

Vowel Processing in Cluttered Auditory Scenes

Beatrijs van Meerveld

Cover:

Printed by: ..

ISBN: .. (printed) / .. (electronic)



**university of
groningen**



This research was made possible thanks to STW grant DTW 7459.



university of
 groningen

Vowel Processing in Cluttered Auditory Scenes

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on
Monday 25 April 2016 at 14.30 hours

by

Beatrijs van Meerveld

born on 24 October 1979
in Ede

Supervisors

Prof. L.R.B. Schomaker

Prof. D. Baskent

Co-supervisor

Dr. T.C. Andringa

Assessment committee

Prof. S.L. Denham

Prof. J. Nerbonne

Prof. M. Cooke

Preface		vii
1	General introduction: The problem of spoken key-word spotting	1
1.1	Key-word spotting	1
1.2	Scope of this work: Vowels and phonetic knowledge	2
1.3	Signal-driven speech processing	3
1.4	Knowledge-guided speech processing	5
2	Machine vowel processing: Signal-driven processing	9
2.1	Speech representations	9
2.1.1	Acoustic-phonetic features	9
2.1.2	Spectro-temporal Gabor features	11
2.1.3	Glimpses of speech in noise	12
2.1.4	Speech representations evaluated	12
2.2	Robust feature extraction: Local features	14
2.2.1	How formants are related to key-word spotting	14
2.2.2	Experiment 1: Formants by interpolation for spoken vowels	15
2.2.3	Experiment 2: Formants by interpolation in continuous speech	25
2.2.4	Experiment 3: Formants on peak locations in continuous speech	34
2.2.5	Local features built on a harmonic complex as alternative representations of speech	40
2.3	Robust key-word spotting: Global speech representation	42
2.3.1	Experiment 4: The effect of selecting voiced-speech-components on the performance of an HMM-based classifier	42
2.4	Robust representations in signal driven speech processing	50
2.4.1	Conclusions	50
3	Human vowel processing: Knowledge-driven & signal-guided processing	51
3.1	Local features in models for human speech processing	51
3.2	Experiment 5: Audiovisual vowel perception	53
3.3	Experiment 6: Auditory vowel perception	71
3.4	Speech processing in noise needs local features and knowledge-driven processing	81

4	General discussion: Key-word spotting by humans and machines	85
4.1	Key-word spotting	85
4.1.1	Signal-guided feature detection	85
4.1.2	Knowledge-driven feature weighting	86
4.1.3	Integrating signal-guided and knowledge-driven processes	88
4.2	The problem of spoken key-word spotting	88
	Samenvatting (summary in Dutch)	99
	Summary (summary in English)	103
	Author publications	107

Preface

This is a dissertation about the recognition of speech components in noisy environment by humans and machines.

Groningen, December 2015
Bea Valkenier

Chapter 1

General introduction: The problem of spoken key-word spotting

1.1 Key-word spotting

Automatic spoken Key-word Spotting (AKS) refers to the detection of predefined individual words by machines, similar to word spotting in handwritten manuscripts (Van der Zant, Schomaker & Haak, 2008; Rath & Manmatha, 2007). AKS differs from Automatic Speech Recognition (ASR) in the focus on the detection of a few target words instead of the recognition of sentences or phrases. Applications of AKS are, for instance, found in mobile telephones or navigation systems. In mobile telephones, a telephone call can be activated by saying the name of the person in question. In navigation systems in cars the navigation goal can be entered by pronouncing it. However, such applications generally limit the input to target word input and only function appropriately in noise-free conditions or when noise is mild and stationary or predictable.

In contrast to the limitations posed on the application area of AKS, human listeners can recognise words in many more listening conditions. For example, interesting words (such as ones name) can be discerned spontaneously, even in acoustic mixtures of multiple speech signals (Cherry, 1953) and without help of sentence context information (Wood & Cowan, 1995). In line with this contrast between AKS and human performance Huang, Baker & Reddy (2014) pose that *"despite the impressive progress over the past decades, today's speech recognition systems still degrade catastrophically even when the deviations are small in the sense the human listener exhibits little or no difficulty."* They assign the problem of dealing with uncertainties (such as resulting from noise, speaking rate or speaker dialect for example) as one of six challenges that still need to be taken before ASR can reach human level performance. Li & Allen (2011) argue that *"perhaps ASR performance will improve if we can answer the fundamental question of HSR: How is the speech coded in the auditory system."*

Other arguments extend this view by stressing the importance of the interplay between different levels of aggregation and bottom-up versus top-down processes. Dusan & Rabiner (2005) focus on the different levels of speech representations by arguing that the complexity of the brain might be able to learn many representations, such as phoneme, phoneme transitions and words, and process them in parallel. In contrast, Ellis (1996, 1998) focuses on the utilisation of knowledge when the bottom-up signal contains a mixture of target speech and non-target sounds: *"The key insight is that it will not always be possible to extract a signal from interference in a unique or optimal way, but rather it is necessary to bring to bear a wide range of contextual constraints and prior biases in*

a heuristic search for an account of the signal that is at least reasonably satisfactory.”

This interplay of bottom-up signal driven processing and top-down knowledge guided processing is demonstrated with an equivalent from the visual domain (Figure 1.1). In this picture, the processing of the figure starts with a signal-driven input, the black and white spots. Recognition of the dalmatian is derived by knowledge of the animal and the edges of the depicted dalmatian can be specified only once the dalmatian is recognised. Signal and knowledge are both needed for the correct recognition of the dalmatian. Similarly, speech processing may profit from the investigation of both signal-driven and knowledge-guided speech processing such that it can be applied flexibly.



Figure 1.1: *In this figure a dalmatian can be discerned despite the absence of edges. It illustrates the interplay of bottom-up signal driven processes and top-down knowledge-guided processing. Only with the help of both signal-driven and knowledge guided processing, the dalmatian can be recognised.*

To improve our understanding of automatic and human speech processing in variable conditions we investigate signal-driven speech representations and knowledge-guided speech processing. In the current work we focus on the processing of vowels and we limit the influence of knowledge to phonetic knowledge. This choice is substantiated below.

1.2 Scope of this work: Vowels and phonetic knowledge

A robust speech processing model depends on components of both Signal-driven Speech Processing and Knowledge-driven Speech Processing. Namely, the expectancy-driven selection of elements must be driven by signal evidence. This is an intricate, dynamic interplay that changes with varying listening conditions and from listener to listener.

The scope of this work is limited to vowels in both signal- and knowledge-based processing and to phonetic knowledge in knowledge-based processing. Vowels exhibit relatively high energy levels and are therefore robust to noise (Andringa, 2002). Also,

vowels contribute significantly to the intelligibility of speech (Cole, Yan, Mak, Fanty & Bailey, 1996). Cole et al. (1996) performed an experiment where either vowels or consonants were replaced by noise in sentences with an equally balanced number of vowels and consonants. They found that twice as many words were recognised when vowels were retained than when the consonants were retained and conclude from their findings that recognition of words depends more upon vowels. Additional to the high information value of vowels for speech processing, they provide information on vocal characteristics such as vocal tract length and pitch of speech, which can be used in a wide range of tasks such as gender/talker identification (Whiteside, 1998; Mury & Sigh, 1980) and emotion/prosody perception (Kienast & Sendlmeier, 2000). Finally, the formant movement of vowels is indicative for preceding and following consonants by upward or downward frequency shifts of formants. For these reasons the vowel can be considered as an anchor in speech perception in the context of this research; the vowel can disambiguate speech sounds by backward- and forward predictions. Disambiguation by predictions is done in the model of Barker, Cooke & Ellis (2005). They used salient time-frequency structures as a bottom-up input to disambiguate between target sound and contaminating sounds by applying an iteration of expectation-driven hypotheses and signal-driven evidence. Signal-driven vowel processing may function as a slightly more precise bootstrapping seed in such an expectation-driven model. High-level knowledge, knowledge of grammar or words for example, is built upon low-level knowledge such as knowledge of speech sounds. By limiting our research to the processing of vowels, we limit the influence of knowledge to low-level knowledge.

1.3 Signal-driven speech processing

A poor representation of the speech elements in machine speech recognition is argued to be one of the factors for levelling performance growth in ASR (Li & Allen, 2011; Dusan & Rabiner, 2005). The first requirement that we concentrate on in this thesis is the signal-driven estimation of a robust representation of vowels. In this thesis we focus on speech related time-frequency components; structures that are connected in time and frequency. This approach for speech representation differs from most existing approaches for ASR (evaluated in an overview article by Li, Deng, Gong & Haeb-Umbach, 2014; Cutajar, Gatt, Grech, Casha & Micallef, 2012) where featural representations are calculated to represent the whole spectrum over pre-set intervals of the time-frequency representation (TF-representation).

The representation of speech with these whole spectrum representations, generally Mel Frequency Cepstral Components (MFCC)s, deteriorates quickly in noise conditions because the overall spectral shape is determined by the target speech as well as non-target sounds (noise). This effect is illustrated in Figure 1.2 where the spectral shape is plotted for both speech in quiet and speech in noise (taken at the same time-frame). One of the reasons that global features are often relied on is that they are developed to be used with Hidden Markov Models (HMMs), a classification method that has been and still is highly valued (but see van Oosten & Schomaker, 2014, for a different viewpoint

on the effectiveness of HMMs). As a result, for current ASR approaches, the reliance on the sentence context to recognise a word becomes especially important when noise levels increase. Currently, to have a properly functioning system in real-life speech conditions, the task settings are often simplified, such as with single speaker input or with a speech upon request procedure where speech is only processed when requested. If reliance on sentence context information in noise can be reduced, by the use of noise-robust representations this would enhance the applicability of automatic speech processing approaches. Also, for AKS it would be an advantage if natural boundaries are available in the representation such that linguistically relevant components (for example phoneme, syllable or word boundaries) can be processed irrespective of the recognition of a sentence.

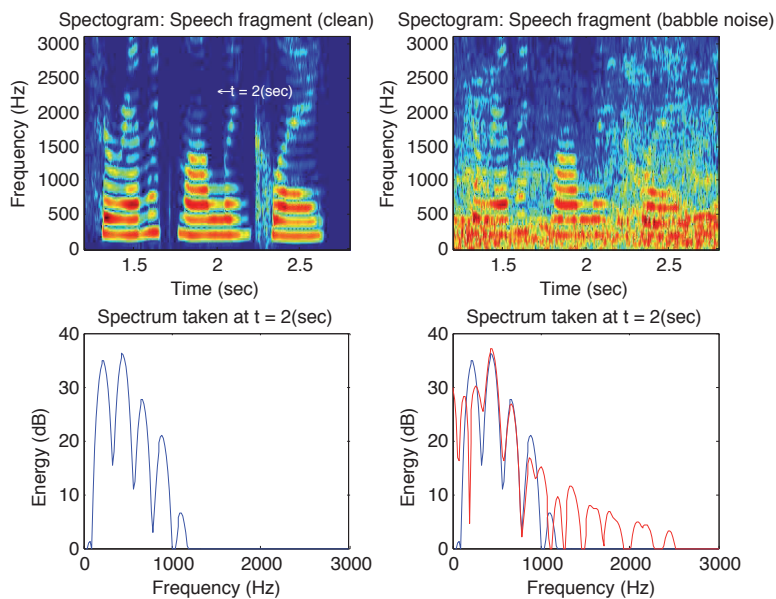


Figure 1.2: *The spectral envelope changes with noise. The upper two panels show speech in clean (left panel) and 0 dB babble noise (right panel). The lower two panels shows the spectrum taken from these spectrograms at $t = 2$ seconds (as indicated in the upper left panel) from clean (left panel) and noisy speech (right panel).*

Despite the vulnerability to noise of the overall spectral shape, human listeners can rely, at least partly (and in clean speech laboratory settings), on the whole spectral shape for the recognition of vowels (Ito, Tsuchida & Yano, 2001; Kiefte & Kluender, 2005). This may be related to some sort of preprocessing, such as investigated by Bregman (1990) on sound processing. Bregman (1990) investigated how sound is organised into perceptually meaningful elements in human sound processing; Auditory Scene Analysis (ASA). One of the effects he demonstrated are primitive perceptual processes that lead to grouping and segregation of elements. With regard to speech processing Green, Cooke & Crawford (1995) argue that *"If ASA depends on these unconditional, primitive processes, they may be viewed as a natural preprocessing stage for ASR."* Thus, the primitive processes may extrapolate to speech conditions and function as a preprocessing mechanism in speech recognition. Bregman (1990) investigated the main properties of this human preprocessing. He summarises the principles that together form the human

capacity to group acoustic elements that originate from a single source (we call these Signal Elements) into a single stream: stream segregation. These grouping principles lead to the segregation of a target stream from non-target noises. Such effects may function as a preprocessing mechanism for human listeners also in speech processing. Cooke (1993) investigated the applicability of the grouping principles on speech processing, he applied Bregman’s grouping principles as a means to segregate speech from noise by a machine approach. If humans apply such grouping principles as a preprocessing method it can improve the stability of the whole spectral shape in conditions with noise or competing speakers. When all target harmonic are correctly determined, the spectral shape is preserved. Additive energy in the harmonics changes the spectral shape only to the degree that non-target energy is added. Because human listeners may rely on both local (time-frequency components) and global (spectral shape) speech representations we investigate both representations of speech in noise in Chapter 2.

In Section 2.1 of this thesis, we describe *local speech representations* that are developed as alternatives to the commonly applied (global) feature representations (such as MFCC). Analysis of the alternative structures provide data for the understanding of strengths and weaknesses of speech representations. We consider it important to understand which characteristics of speech representations are desirable for speech recognition purposes. This is in line with Boulard, Hermansky & Morgan (1995) who argue that improvements in ASR might need to be developed with the help of new approaches that do not instantly lead to improvement of recognition scores. The alternative representations that we describe in this thesis have in common that they focus on the extraction of local time-frequency structures. Local structures can be noise robust because (1) they can be selected such that they have a high local SNR, and (2) noises affecting low-frequency structures of the time-frequency representation do not affect high frequency structures and vice versa which also makes them less vulnerable (Kleinschmidt, 2003). In contrast to these local speech representations we describe *global spectral envelope structures* in Section 2.3. The representation of the spectral shape after application of this preprocessing approach is investigated and described in Section 2.3. We show that spectral shape representations show increased robustness to noise when harmonic grouping is performed. The approach, investigated in this thesis (Section 2.2) to estimate a signal-driven robust representation of speech, is to *combine the robustness in ASA approaches with the robustness of local features*.

1.4 Knowledge-guided speech processing

The second factor associated with the contrast between human and machine speech processing performance, is the *structure that knowledge and expectation can impose on sensory input*. The second task of key-word spotting that we therefore concentrate on in this thesis is to investigate the structuring effect of knowledge in human perception of speech sounds. Davis & Johnsruide (2007) argue that *”interactions between higher-level linguistic knowledge and bottom-up perceptual processes are necessary for successful speech perception.”* Humans exploit knowledge to achieve efficient perception. This was

illustrated for the visual domain with Figure 1.1 where the dog is perceived with the structuring effect of knowledge of the visual characteristics of dalmatians. We consider the effective use of knowledge as one of the factors leading to improved performance levels on key-word spotting.

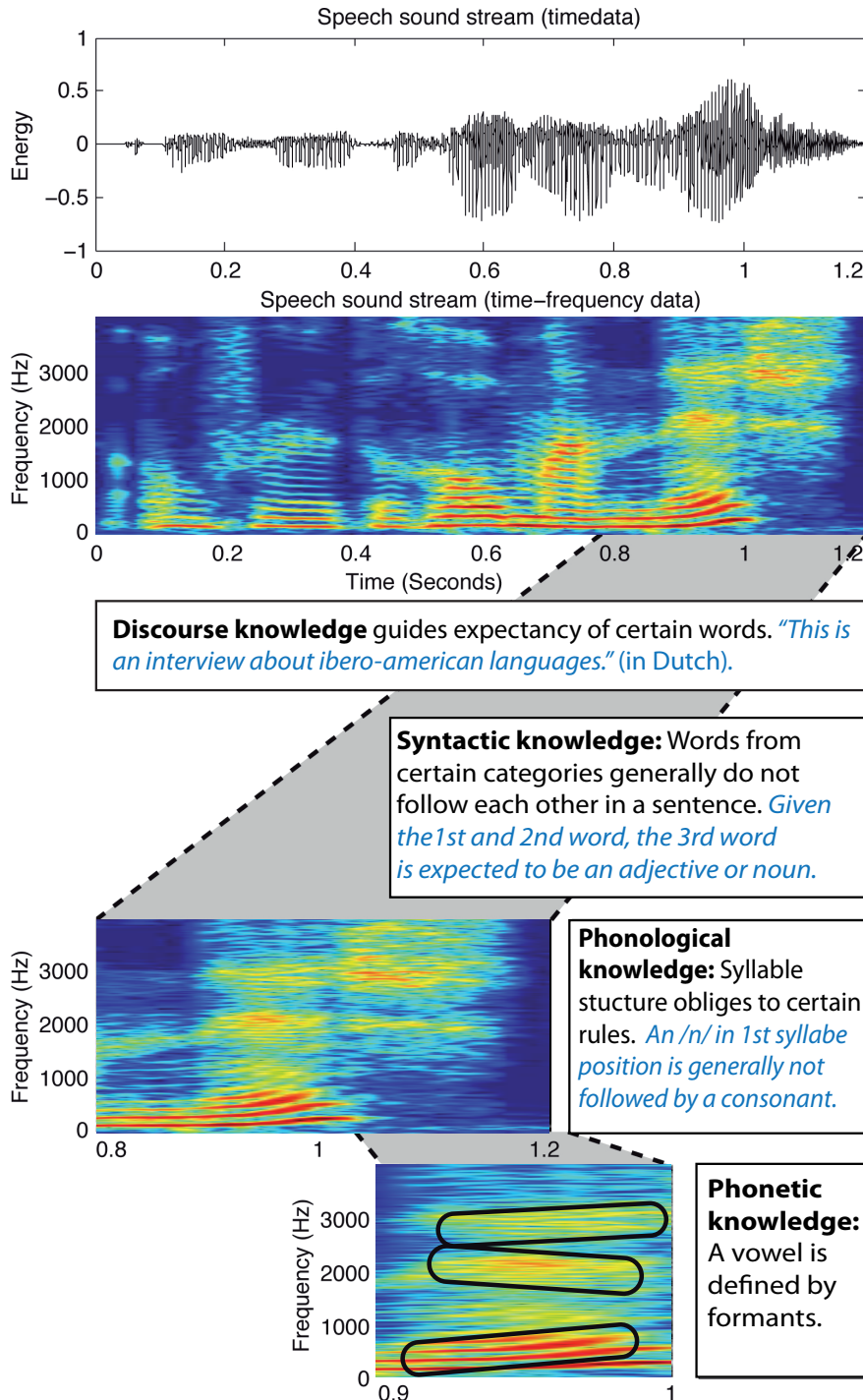


Figure 1.3: Knowledge of different levels of aggregation can all impose structure upon seemingly unstructured sensory input, such as a continuous stream of speech sounds.

Experimentally, the effect of knowledge on human perception of speech is demonstrated, for example, when the same acoustical signal can be perceived as two different words depending on the context (McQueen, Norris & Cutler, 1994). One of the examples that McQueen et al. (1994) give, is the input string /barti/. This input can lead to the perception of /bar/, /art/ and /tea/. McQueen et al. (1994) show that given the context of the whole string (which can be considered knowledge) the words /bar/ and /tea/ are perceived, as they cover all input sounds. The influence of knowledge on perception is generally accepted, but a debate is still ongoing (for a theoretical overview see Simpson, 1984; Tabossi & Zardon, 1993) of whether this knowledge is applied early (McClelland & Elman, 1986) or late (McQueen et al., 1994) in the recognition process. In general, it can be stated that knowledge-driven expectations guide the perception process, either early or late. Expectations may be formed based on linguistic knowledge such as phonetic, phonological, syntactic, semantic, and discourse knowledge. The above example given by McQueen et al. (1994) is based on linguistic knowledge. Figure 1.3 illustrates the structuring effect that expectations, based on different types of linguistic knowledge, can have on speech perception. However, expectations may be generated on the basis of other knowledge as well. For example Barker et al. (2005) use the time-frequency energy-envelope as a non-linguistic cue to bootstrap their model. They show how an iteration of bottom-up signal processing processes and top-down structuring processes can lead to robust speech decoding without the special need of linguistic knowledge. However, they do discuss the potential advantage of pitch cues to decode the target and non-target information in the speech.

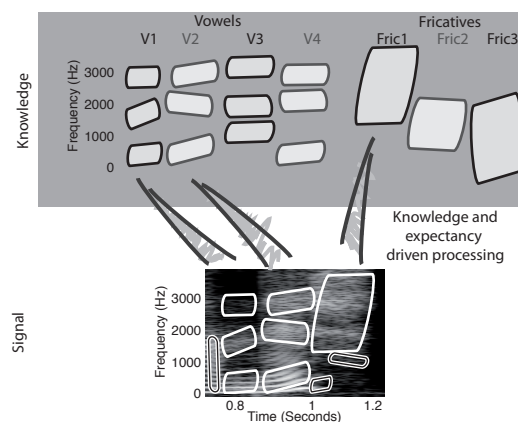


Figure 1.4: *The picture illustrates how knowledge of clusters of acoustic elements that are related to speech can help to select target elements from the input stream. Knowledge can lead to the processing of the known subset of components*

In this thesis we suggest (in line with Mattys, Davis, Bradlow & Scott, 2012) that when task demands increase, for example when speech is embedded in unpredictable / non-stationary noise, processing efficiency can improve by using knowledge (knowledge-driven processing) in the selection stage instead of in the recognition stage. In theory, this approach seems especially effective when integrated with a method to determine local components in a signal-driven approach as illustrated in Figure 1.4. This figure illustrates how knowledge of the relation between acoustic components and speech

sounds can help to detect target components from the input stream while ignoring non-target components. This way knowledge assists in breaking down the sound stream into elements or clusters of elements that comply with predictions based on previous observations (following the approach of Barker et al., 2005). By means of experimental perception research we aim to improve our understanding of the role of knowledge in speech processing. Therefore, we focus on *knowledge-driven speech processing by humans and its implications for automatic approaches*.

Chapter 2

Machine vowel processing: Signal-driven processing

2.1 Speech representations

Features are computable aspects of the signal that help in discriminating between different pattern classes. In this respect there is a distinction between computing in engineering systems and neural computation in the brain. Features that are commonly applied in systems for speech recognition (Mel Frequency Cepstral Components; MFCCs) are not optimal and have been thought to be one of the causes of the levelling performance growth in speech recognition research (Li & Allen, 2011; Dusan & Rabiner, 2005), especially for automatic recognition of speech in noise.

Recent findings show high potential for statistical features that are learned by a Deep Neural Network (DNN) as these systems outperform MFCC based models in clean speech conditions (Hinton, Deng, Dong, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath & Kingsbury, 2012) and equivalent to MFCC based models in noisy speech conditions (Seltzer, Yu & Wang, 2013). However, in this thesis focuses on what we call structural features. They have a direct relation with the TF-representation. The power of structural features is that they have high domain-specificity in contrast to statistical approaches that need high numbers of samples. The goal of the current evaluation is to understand the relevant characteristics of robust speech representations. In addition to robustness of speech representations, the quality of phonetic descriptions that the features capture can be relevant for improved understanding of speech characteristics.

Analysing diverse speech representations may improve our understanding of the relevance of different spectral characteristics for speech representation. To investigate the benefits and limitations of different features, we will discuss three alternative speech representations that are described in the literature: Learned and hand-defined Acoustic Phonetic features (AP-features), spectro-temporal Gabor Features and Glimpses. Not all three approaches are equally concerned with noise-robustness but all approaches provide an alternative viewpoint to the MFCC features that are vulnerable to noise. Understanding the strength of different representations helps to develop features that comply with most of the characteristics.

2.1.1 Acoustic-phonetic features

Although Acoustic-phonetic features (AP-features) are not recently developed (literature ranging from 1983 for Bush (1983) to 2009 for Strik, Truong, de Wet & Cucchiari (2009)), we describe them here because their direct relation to the signal is interesting

to understand the relevance of different characteristics of the signal for speech processing. AP-features are acoustic correlates to articulatory actions (Holmes & Holmes, 2002; Cohen & Mercer, 1975). They are the spectral changes in the time-frequency plane that co-occur with articulatory actions as deduced from phonemic annotations. For example, from the phonemic annotations of [k, p and t] it can be deduced that a plosive is articulated. The articulated plosive co-occurs with a pulse in the time-frequency plane. Table 2.1 shows how different phonemes are described by articulatory actions. For example; the phoneme [k] is formed by an articulatory stop (described by ”+ plosive”) at velar position (described by ”+ velar”). AP-features are the local acoustical characteristics in the time-frequency domain of these actions. The AP-features can be automatically extracted from a learned or hand-defined representation (De Mori & Flammia, 1993; Kirchhoff, Fink & Sagerer, 2002; Wester, 2003) as explained in the next two paragraphs.

Learned acoustic-phonetic features

When AP-features are learned, the phonemes in an annotated database are automatically rewritten into Articulatory Features. This can be done by a reference table as demonstrated in an example in Table 2.1. De Mori & Flammia (1993); Kirchhoff et al. (2002) and Wester (2003) apply an artificial neural network approach to learn acoustic features from such rewritten annotations. With this approach the respective results are 77% feature correct rate (De Mori & Flammia, 1993, clean speech stops and nasals), 91% word correct rate (Kirchhoff et al., 2002, 12 digits clean speech) and 87% phoneme correct rate (Wester, 2003, syllables excised from continuous clean speech). With these results it is demonstrated that acoustical local structures, smaller than a phoneme can function as speech representation. The results should be evaluated while taking into account that all approaches use a simplification of the task settings. Both De Mori & Flammia (1993) and Wester (2003) simplify the task by using phonemes excised from clean speech sentences to train and test the features. Kirchhoff et al. (2002) simplify the task by applying an ASR system and a relatively simple database consisting of spoken digits.

Table 2.1: Example of a reference table to rewrite phonemes into acoustic-phonetic features

phoneme	voicing	manner	place	. . .
[k]	-voice	+plosive	+velar	. . .
[g]	+voice	+fricative	+velar	. . .
[m]	+voice	+nasal	+labial	. . .
[. . .]

Hand-defined acoustic-phonetic features

An alternative approach to obtain acoustical representations of articulations is to define how acoustical structures are related to articulatory actions and extract the structures that comply to this definition. Such definitions can be based on phonetic knowledge (hence: Acoustic Phonetic features) or they can be based on the developers’ intuitions. Defining AP-features by hand is therefore less straightforward than learning them. However, results obtained with hand-defined features can provide new insights into represen-

tations that can be useful to learn.

Bush (1983) was one of the first to investigate AP-features. Features, based on spectral energy, were determined and tested by eye; human inspection of spectrogram representations resulted in feature descriptions for phonemes. Strik et al. (2009) developed AP-features, based on the development of the overall spectral energy. Their purpose was to provide pronunciation feedback to non-native speakers. The goal of this investigation was to distinguish velar plosive /k/ sounds from velar fricative /x/ sounds from a database with the two sounds excised from read sentences over a telephone connection. They reported 93% scoring accuracy as calculated by

$$100 * \frac{\text{accept}(\text{correct}) + \text{reject}(\text{correct})}{\#\text{tokens}} \quad (2.1)$$

The approach is useful to distinguish phonemes for the case of excised target-phonemes, context information is not used to disambiguate between sounds.

Another approach that does not demand pre-segmentation is described by (Liu, 1996; Abdelatty Ali, Van der Spiegel, Mueller, Haentjens & Berman, 1999). They performed both automatic phoneme segmentation and phoneme classification. Liu (1996) developed features based on the energy development in selected frequency-bands. He reported 86% feature correct rate on read sentences in 30dB SNR. Abdelatty Ali et al. (1999) used features related to spectro-temporal energy to segment continuous clean speech into stops, fricatives, sonorants and silences (92% accuracy). Subsequently, the segmented stops and fricative were classified. They reported 86% accuracy for stops (Abdelatty Ali et al., 1999; Abdelatty Ali, van der Spiegel & Mueller, 2001) and 90% accuracy for fricatives (Abdelatty Ali et al., 1999; Abdelatty Ali, Van der Spiegel & Mueller, 2001).

The investigation of AP-features is to provide a phonetic analysis. Two advantages of AP-features are that they 1) can provide insight in the correctness of pronunciations (Strik et al., 2009) and 2) that asynchronous feature changes are captured, which can become important when not all features are extracted equally reliably Wester (2003).

2.1.2 Spectro-temporal Gabor features

Gabor Features are speech representations that are the result of convolving the signal by Gabor filters; linear filters with variable orientation and frequency. The resulting representations are time-frequency components with varying direction and scale. The Gabor Features are based upon research into the human visual system and are applied for edge-detection in image processing. These features were adopted for auditory processing by (Kleinschmidt, 2002b, 2003). In terms of auditory processing these features capture local energy structures in temporal and spectral direction with varying orientations. A speech structure can be a pulse (related to plosives), a tone or a group of similarly spaced tones (related to a harmonic complex). Kleinschmidt (2003) argues that spectro-temporal features have the advantage that they (1) are noise robust as a result of the locality of the features, (2) detect diagonal patterns in the time-frequency representation (tones that change through time), (3) can be adapted to different types

of patterns and (4) incorporate also the characteristics of existing features.

The performance levels obtained from a system using Gabor features in combination with a simple artificial neural network are comparable to those obtained with a standard ASR system (Kleinschmidt, 2002a). Kleinschmidt (2002a) reports 99% word correct rate on a 12 digit clean speech task which is comparable to common ASR approaches. Also, Kleinschmidt (2002a) and Heckmann, Domont, Joublin & Goerick (2008) show that systems based on Gabor Features perform slightly better than systems based on MFCC when trained on clean speech and tested in noise. Kleinschmidt (2002a) obtained word correct rates of 33% and 20% on a 12 digit task in speech shaped noise (0 dB and -5 dB respectively) using the Gabor Features and 14% (both 0 dB and -5 dB) using cepstral features. Heckmann et al. (2008) reported similar results in factory noise on the same database using a hierarchical structure of three layers of spectro-temporal features and an hidden markov model (HMM) approach.

2.1.3 Glimpses of speech in noise

In contrast to the features discussed so far, Glimpses are developed with speech in noise as a starting point. Glimpses are spectro-temporal regions where the signal energy is above the local average energy. Cooke (2006) showed that Glimpses, extracted by taking pixels that (1) exceed the energy levels of the noise by 3 dB and (2) are connected in time or frequency, lead to recognition patterns that are similar both in robustness to noise and in the trends of the identification pattern of phonemes to that of human listeners as tested on a 16 consonant recognition task and averaged over different noise conditions. Glimpses, as structures, is useful for explanatory purposes of human speech perception.

However, because Glimpses are generally calculated using a model of the noise, they seem less directly applicable for speech recognition purposes in ASR. An attempt to effectively use glimpses are the missing data techniques (described in Cooke, Green, Josifovski & Vizinho, 2001). These are based on glimpses and are used to find the best sequence of words in ASR. They are based on noise estimation to decide whether fragments are reliable or not. With this method glimpses can be calculated when noise is relatively predictable. Such an approach (Gemmeke & Cranen, 2009; Gemmeke, Cranen & Remes, 2011) leads to high robustness to noise if the noise can be estimated reliably for relatively long time-periods. They report phoneme correct rates of 98%, 95% and 87% (10 dB, 0 dB and -5 dB noise mixture) on a digit database (Gemmeke & Cranen, 2009) and 93% to 89% (15 dB to 0 dB multi-speaker babble noise) on a continuous speech database (Gemmeke et al., 2011). *If reliably extracted, glimpses can be effectively used for ASR.*

2.1.4 Speech representations evaluated

The representations discussed here have a focus on local, relatively high energetic structures. This is mainly due to the fact that we chose to evaluate features that can be directly related to the TF-representation. Wester (2003) argues that the temporal asyn-

chronicity in such features may become relevant when some local representations can be more reliably extracted than others. Also, local structures are robust to noise when they are chosen such that they exhibit high energy levels (Kleinschmidt, 2003). The combination of high energy and locality of structures can help to disentangle problems with noise in AKS while capturing information to provide phonetic descriptions of the speech sounds. Of the evaluated representations the Glimpses (Cooke, 2006) inhabit both qualities, they rely on high energy levels and can be determined prior to segmentation.

Similarly, harmonic complexes are structures based upon local, high energetic representations; tonal signal elements. In human speech recognition voiced speech is described by both local structures such as harmonic complexes and formant-tracks (Molis, 2005; Ito et al., 2001; Kiefte & Kluender, 2005) and global, spectral shape representations (Molis, 2005; Ito et al., 2001; Kiefte & Kluender, 2005). Therefore, in the next two sections (Section 2.2 and 2.3) we investigate the robustness of both local and global representations based on harmonically related tones.

2.2 Robust feature extraction: Local features

2.2.1 How formants are related to key-word spotting

We discussed three speech representations, different from the commonly used MFCC features, with the goal to map the useful acoustical characteristics of speech representations. We concluded that the locality in the time-frequency domain of these alternative representations have two advantages over more global representations. First, they can be of help to find representations that are robust to noise. Second, they capture information that is suited for phonetic descriptions of the articulations.

Bregman (1990) investigated human perceptual processes. His findings suggest primitive principles that lead to grouping of tones such that simultaneous or successive tones are perceptually grouped (this paradigm is called Auditory Scene Analysis; ASA). This is later used to develop approaches to perform ASA automatically on speech; Computational Auditory Scene Analysis (CASA). The harmonic complex consists of high energy, tonal structures that segment speech into voiced parts and are robust to noise. Therefore, our goal is to define structures that are based on harmonic complexes and describe speech. Because the current work focuses on vowels, formant structures are extracted from harmonic complexes.

Formants are the resonance frequencies of the vocal tract; they change with the shape of the vocal tract. *Formants can be sufficient to understand speech.* Barker (1998) investigated the usefulness of formant tracks for ASR with sine-wave speech. Sine-wave speech consists of three tones that are played at formant track positions and is understandable for trained human listeners (Remez, Rubin, Pisoni & Carrell, 1981). Barker trained and tested existing ASR systems on sine-wave speech and found that 85% of the sine-wave speech could be recognised correctly. From this, we conclude that the reliable extraction of formants in different acoustical conditions can function as a first step for key-word spotting in noise. *The development of a noise robust formant extraction algorithm can bring us one step closer to understanding the problem of noise robust key-word spotting.*

Standard formant tracking approaches are developed to process large corpora with clean speech elements. They generally rely on Linear Predictive Coding (LPC). LPC based methods have some limitations. Jacobi (2009) describes that when using LPC, formants are less well extracted when f_0 and f_1 interact, when formant peaks lie close, or when formants are low (such as in high back vowels). Other approaches focus on formant estimation to analyse speaker differences (and not primarily on the estimation of the formant track). These methods are generally based on Principle Component Analysis (PCA) and the problems that arise from LPC are reduced with PCA based techniques (Plomp, Pols & van der Geer, 1967; Jacobi, 2009). Both LPC and PCA are developed for phonetic analysis and are not optimised for noisy speech conditions. A few attempts are made for formant extraction in noisy speech conditions (Mustafa & Bruce, 2006; Yan, Vesghi, Zavarehei, Milner, Darch, White & Andrianakis, 2007) but these methods still show results that deteriorate quickly in noise. One exception to this can be found by

the system that was developed by Gläser, Heckmann, Joublin & Goerick (2010). They use a Bayesian technique to predict formant tracks with an added preprocessing method to enhance the formant structures in the spectrograms. In this thesis we focus on the extraction of features for speech processing in noise and in the current chapter this focus is on determining the formant-tracks in noisy speech conditions where we will compare our results against the work of Gläser et al. (2010).

In the current chapter (Chapter 2.2), we investigate whether we can use the extractions of a harmonic grouping algorithm to robustly extract local features. Two different grouping algorithms and two algorithms to extract local features were used in this work. The first grouping algorithm (Krijnders, Niessen & Andringa, 2010, in this thesis dubbed COCHL) is based on tonal components extracted from a simulation of the cochlear response (described globally in Section 2.2.2). The second grouping algorithm (in this thesis dubbed SPECT) is based on a spectrogram representation of the speech sounds (van de Vooren, Violanda, van Elburg & Andringa, 2010a,b). The SPECT method is preferred over the COCHL method because it allows a broader range of speech input, such as speech in noise and continuous speech.

Two algorithms to extract local descriptions of voiced speech components are described. We first describe an algorithm to estimate formant positions, based on interpolation over harmonics. Formant detections resulting from this algorithm are found to be stable (Section 2.2.2) over different noise conditions for a spoken vowels database. From a continuous speech database, with more formant movements, it follows that formant tracks are less well estimated at formant movement positions (Section 2.2.3) both in clean and noisy speech conditions. Therefore, a *second* algorithm is developed. This algorithm is based on the relative energy in harmonics of a harmonic complex. This algorithm leads to stable results for regions with and without formant movement (Section 2.2.4) in both clean and noisy speech conditions. However, these extractions are less directly related to formants. *The results indicate that harmonic complex extraction supports robustness of speech representations.*

2.2.2 Experiment 1: Formants by interpolation for spoken vowels

A modified version of this chapter was previously published as:
 Valkenier, Krijnders, van Elburg & Andringa (2011). "Psycho-acoustically motivated formant feature extraction". *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011* 11:218-223

In this experiment we focus on a reliable extraction of formants in noise from spoken vowels with relatively little formant movement. Because formants are relatively energetic parts in the harmonic complex we can extract the same or similar formant values in noisy as well as clean conditions. Human listeners can detect and recognise speech in uncontrolled environments with relatively little interference from background noises (O'Shaughnessy, 2008). Humans seem to apply many top-down mechanisms to enhance

perception of degraded speech (Başkent, 2012; Grossberg & Kazerounian, 2011). One of these mechanisms, as applied to speech in the field of CASA is the grouping of components to recombine components from different sound sources into a single percept. These grouping principles can be applied, provided the individual components are separable from background noise depending on the signal to noise ratio (SNR). In general, systems based on grouping of harmonics are applicable in uncontrolled environments and do not rely on training. However, harmonic complexes are sometimes overlooked and pitch estimations can fail in one or more octaves.

Experiment 1: Algorithm

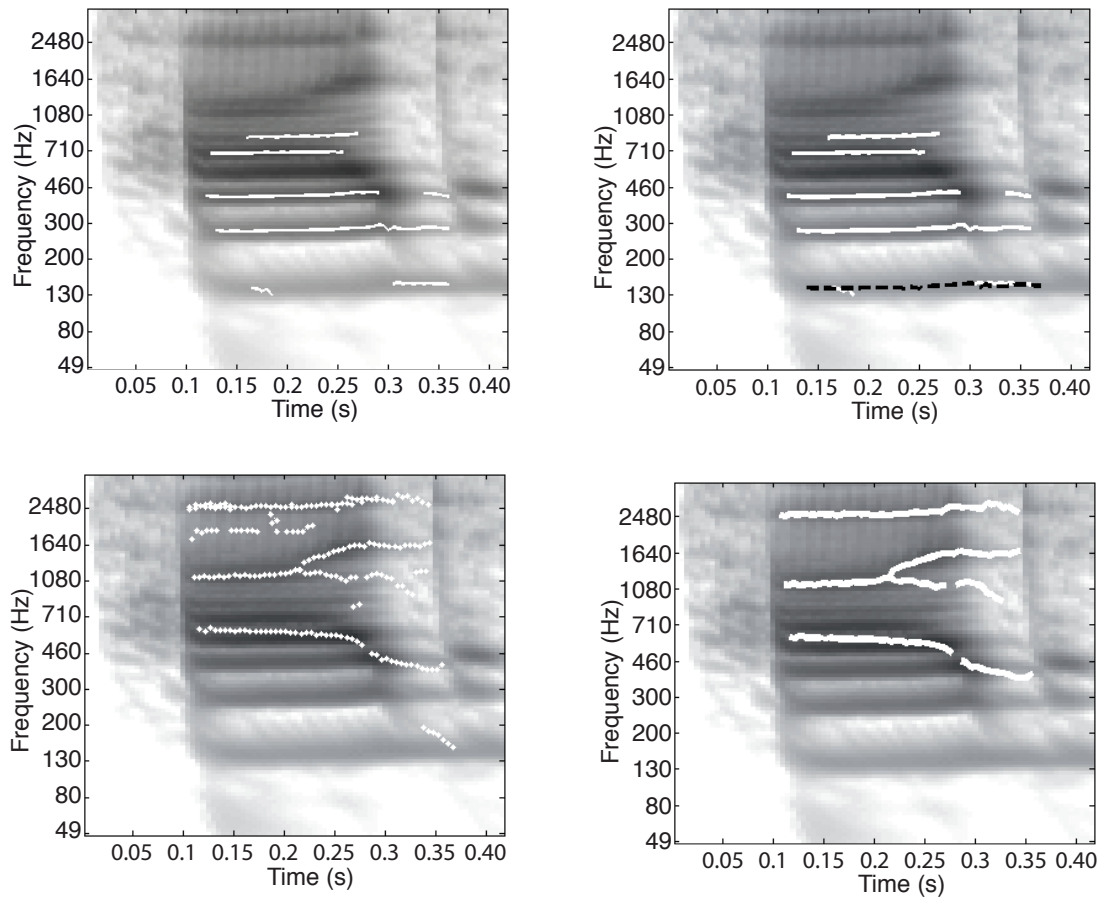


Figure 2.1: Results of the different steps in the algorithm represented on a cochleogram of a male speaker pronouncing [hud]. (top left) energetic signal components; (top right) selected HC, the fundamental frequency is given by the dashed line; (bottom left) formant detections based on this fundamental frequency and its overtones that fall below 4000 Hz; (bottom right) selected formants.

To estimate the resonance frequencies of the vocal tract we perform peak interpolation over harmonics in a harmonic complex (HC) as obtained with the "cochleogram method".

First, the time signal is converted to the time-frequency domain by a gamma-chirp filterbank (Irino & Patterson, 1997). Its filter coefficients (h_{gc}) are defined by,

$$h_{gc} = at^{N-1}e^{-2\pi bB(f_c)t}e^{j(2\pi f_c t + c \log(t))} \quad (2.2)$$

where $N = 4$ is the order of the gamma-chirp. The coefficients ($a = 1$, $b = 0.71$, $c = -3.7$) are based on Irino & Patterson (1997) but were adjusted such that the response is narrower in frequency and the tonal components are emphasized. The frequency range f_c is fully logarithmic from 67 to 4000 Hz over 100 channels. The bandwidth (B) of the filters is given by Moore (1996),

$$B(f_c) = 24.7 + 0.108f_c \quad (2.3)$$

We call the averaged and logarithmically compressed result a cochleogram.

Second, harmonics are extracted from the cochleogram using a measure called tone-fit (Krijnders, 2010; Krijnders et al., 2010). The tone-fit is calculated per segment and per channel. It is a measure of how closely the local shape of the cochleogram matches the ideal shape of a tone. The tone-fit of a segment is the normalised difference between the energy of the top of the tone minus the mean of the energy at one sine-broadness. Sine-broadness is calculated by $sb = sb_1 + sb_2$ with sb_1 the difference between the frequency position of the top and the frequency position of the upward slope (towards higher frequencies) at threshold value and sb_2 the difference between the frequency position of the top and the frequency position of the downward slope at threshold value. The threshold value is given by $th_n\omega_n$ with typically $th_n = 2$ for all segments and ω_n the standard deviation of white noise at cochleogram channel n .

The tone-fit is calculated by

$$TF = \frac{E(n) - \frac{1}{2}((E(n - sb_1) + (E(n + sb_2)))}{th_n\omega_n}, \quad (2.4)$$

with E the energy. The difference of the energy at the top and the mean energy at sine-broadness positions is normalised by the local noise standard deviation ω_n . Subsequently, energy patterns of neighbouring channels that resemble the excitation of a perfect tone are extracted and are described as a line - a temporal sequence - through the best matching location. We call such a description a signal component (Figure 2.1, top left).

The final step in harmonic complex extraction combines signal components into harmonic complexes (Figure 2.1, top right). To that end, HC hypotheses are generated from energetic signal components (Figure 2.1, top right) that partly overlap in time and have an approximate harmonic frequency relation to each other. Initially a hypothesis consists of a fundamental frequency (f_0) estimate and energetic signal components. Additional signal components are added later to each hypothesis if they increase the score of that hypothesis. This score S depends on the number of signal components and the number of which are sequential harmonics, the availability of f_0 and congruence with

the f_0 value based on the harmonics. The score S is defined as (Krijnders, 2010; Niessen, Krijnders & Andringa, 2009):

$$S = n_{sc} + b_{f_0} + n_h - \sum_{sc} rms_{sc} - \sum_{sc} \Delta f_{sc}, \quad (2.5)$$

where n_{sc} is the number of signal components in the group, b_{f_0} is one or zero depending on the existence of a signal component at the f_0 , n_h is the number of sequential harmonics in the group, rms_{sc} are the root mean square values of the differences of the signal component f_0 after the mean frequency difference is removed, and Δf_{sc} is the mean frequency difference divided by harmonic number. To reduce octave errors additional hypotheses at octaves above and below each hypothesis are added and scored. In the formant extraction phase only the hypothesis with the highest score is used.

The resonance frequencies of the vocal tract might be located between two harmonics. Therefore, a three point quadratic interpolation over the harmonics around the harmonic with (local) maximum energy is used to estimate the formant location (Figure 2.1, bottom left). Subsequently, formant estimates with minimal distance in the adjacent frames in the time-frequency plane are connected into formant tracks. Only tracks of sufficient duration (7 frames or more, Figure 2.1, bottom right) are kept. These long formant tracks constitute our final formant estimate.

Experiment 1: Material

The formant extraction algorithm was tested on the American English Vowels dataset (AEV) (Hillenbrand, Getty, Clark & Wheeler, 1995). The dataset consists of 12 vowels pronounced in /h-V-d/ context by 48 female, 45 male and 46 child speakers. The AEV dataset is automatically annotated and subsequently hand-corrected for the first four formants at 8 points in time for each vowel (Hillenbrand et al., 1995), which makes it a suitable ground truth. We added pink noise in decreasing SNR, from 30 dB to -6 dB SNR (30 dB, 20 dB, 10 dB, 5 dB, 0 dB, -2 dB, -4 dB, -6 dB). Pink noise was chosen because it has a spectral shape close to the long-term speech spectrum, and hence masks speech evenly across the speech spectrum range.

Experiment 1: Evaluation

For the goal of evaluating the extracted formant-tracks on usefulness for classification and robustness to noise we need 4 measures. A distance measure alone is not suitable for our goal because we do not extract exactly three formants at each annotated time-location. This is illustrated in Figure 2.2 where extractions are pictured by dotted lines and annotations are pictured by dots. Because not always three or four formant-tracks are determined, due to a limitation of the algorithm, the distance can not always be calculated reliably. In order to assess the *usefulness for classification*, we determine the consistency (the distance and the relative number of extra peaks) between the extracted formant tracks ($f_{extr(clean)}$) and the annotated formant frequencies (f_{ann}) in clean speech conditions. Additionally, in order to evaluate the *robustness to noise* we determine the

similarity between the extracted formant tracks $f_{extr(noise)}$ from noisy speech conditions with the extracted formant tracks $f_{extr(clean)}$ from clean speech conditions. The corresponding measures are explained in more detail below.

General efficiency: We specify two measures to determine the consistency between f_{extr} and f_{ann} and compute these in clean speech conditions. Together this indicates how useful the $f_{extr(clean)}$ are for classification based on a high hit-rate and a low false-positive rate.

(1) We define a hit (f_{hit}) as an extracted formant (f_{extr}) corresponding with an annotated formant (f_{ann}), given by

$$f_{hit} = f_{extr(clean)} \cap f_{ann}, \quad (2.6)$$

where the difference in formant frequency between f_{extr} and f_{ann} may fall within the range of 15% (1st formant), 12% (2nd formant) and 8% (3rd formant) to fulfil the criterion of intersection. This equals a mean accepted error of respectively 95 Hz, 316 Hz and 266 Hz. This range is chosen such that formants that were considered correct (based on visual inspection) are included.

We measure the "ratio correct formants" ($r_{correct}$) by weighting the number of hits (f_{hit}) by the number of annotations (f_{ann}). The $r_{correct(x)}$ gives the fraction of annotated formants (calculated per formant) that is consistent with our detections,

$$r_{correct(x)} = \frac{\#f_{hit(x)}}{\#f_{ann(x)}}, \quad (2.7)$$

where x is the formant number ranging from 1 to 3.

(2) A false positive is an extracted formant (f_{extr}) that can not be related to an annotated formant (f_{ann}). We measure false-positives by "ratio spurious peaks" (r_{sp}) calculated over all f_{ann} . The r_{sp} gives the ratio between the number of $f_{extr(clean)}$ that cannot be related to f_{ann} , and the number of f_{ann} ,

$$r_{sp} = \frac{\#f_{extr(clean)} - \#(f_{hit})}{\#f_{ann}}. \quad (2.8)$$

For these two clean speech consistency measures we use $\#f_{ann}$ as a reference value because $\#f_{ann}$ provides a fixed reference (Figure 2.2).

Robustness to noise: We specify four measures to determine the robustness of the extractions to noise. In order to determine how well the extractions are captured in noise we take the correct extractions from clean speech (f_{hit}), as a ground truth. The robustness of the extracted formant tracks is measured by recall and precision and summarised by the F score. Also, the usefulness of the extractions is demonstrated with a

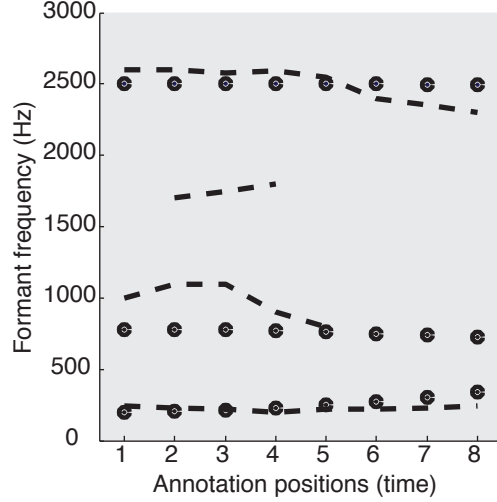


Figure 2.2: Schematic representation of annotations and extractions. Schematic representations of the instantaneous formant frequencies (f_{ann} given by dots) that are given as annotations and of the formant tracks (dotted lines) that are extracted by our algorithm. Limitations with the evaluation of the extracted formants are illustrated by the fourth extracted formant track at annotation positions 2,3 and 4 and the missing formant track at annotation positions 6,7 and 8.

classification experiment.

We define a robust formant f_{robust} , as a formant extracted from the noisy speech condition $f_{extr(noise)}$ corresponding with the ground truth formant f_{hit} , given by

$$f_{robust,x} = f_{extr(noise,x)} \cap f_{hit(x)}, \quad (2.9)$$

where x is the formant number ranging from 1 to 3. The difference in formant frequency between $f_{extr(noise,x)}$ and $f_{hit(x)}$ may fall within the range of 15% (1st formant), 12% (2nd formant) and 8% (3rd formant), based upon visual inspection of the extractions, in order to be included. The calculation of recall and precision are based on f_{robust} with x , the formant number ranging from 1 to 3.

(1) The recall reflects the fraction of the ground truth, $f_{hit(x)}$, that fulfils the criteria of robustness as determined by $f_{robust(x)}$,

$$recall(noise, x) = \frac{\#f_{robust(x)}}{\#f_{hit(x)}}. \quad (2.10)$$

(2) The precision gives the fraction of noisy speech extractions $f_{extr(noise,x)}$, that meets the criteria of robustness,

$$precision(noise, x) = \frac{\#f_{robust(x)}}{\#f_{extr(noise,x)}}. \quad (2.11)$$

(3) We calculated the weighted average of precision and recall per noise conditions and per formant number x , the f-score:

$$Fscore_{(noise,x)} = 2 * \frac{precision_{(noise,x)} * recall_{(noise,x)}}{precision_{(noise,x)} + recall_{(noise,x)}}. \quad (2.12)$$

(4) Finally, the detected formants that are analogous to the ground truth formants are further investigated in how well they are able to classify the vowels in the test material. To that end, a feature vector is constructed, consisting of the frequency values of only the subset of detected formants that are analogous to the reference formants. Due to missing values, i.e. formants that were not detected, we were limited to a small number of classification algorithms to choose from. The Best First Tree (BFT) search algorithm from the WEKA toolbox (Witten & Frank, 2005) allows a weighting of different features. This is a relevant characteristic because different formants represent a different informational value and should be weighted accordingly. We used the BFT search algorithm using a tenfold cross validation method on the detected formants.

Experiment 1: Results

General efficiency: The correct rate ($r_{correct}$) and fraction of spurious peaks (r_{sp}) were calculated for clean speech conditions.

(1) We found correct rates of: $r_{correct(1)} \sim 90\%$, $r_{correct(2)} \sim 75\%$ and $r_{correct(3)} \sim 75\%$. Most of the annotated formants were determined by our algorithm correctly.

(2) We determined an r_{sp} of $\sim 11\%$ over all annotated formants; the percentage of extractions that could not be related to the annotations.

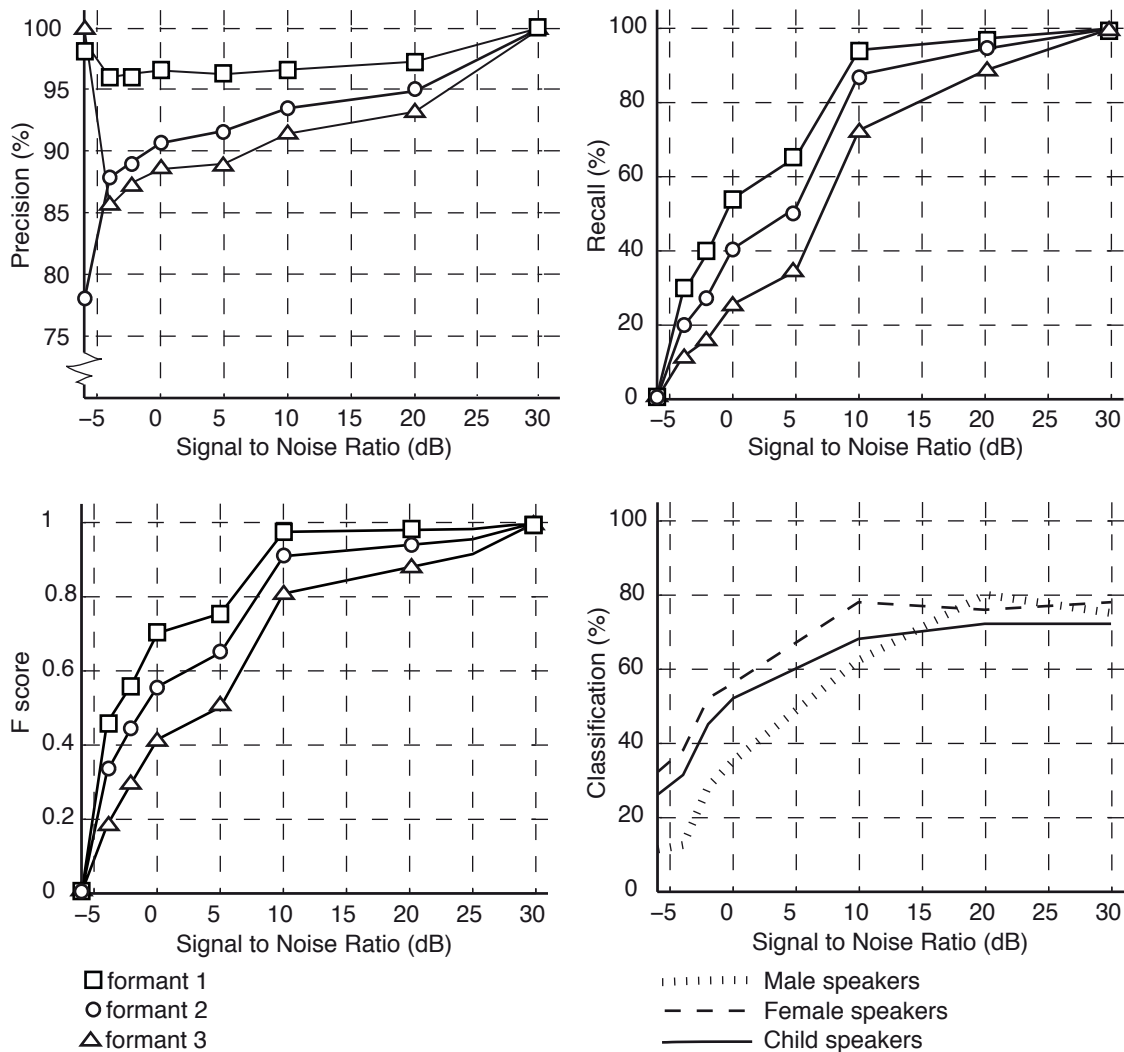


Figure 2.3: Formant extraction in noise tested on a vowels database. Upper Left panel: Precision, the percentage of detected formants that coheres with the annotated formants (i.e. relative error falls within the range of 15% - 1st formant -, 12% - 2nd formant - and 8% - 3rd formant -) in increasing SNR levels in pink noise. Upper Right panel: Recall, the percentage of target formants that is detected in increasing SNR levels in pink noise. Bottom Left panel: F scores, the combined representation of precision and recall. Bottom Right panel: Classification scores obtained with breadth first tree search algorithm on the formant extractions. Lines are added to guide the eye.

Robustness to noise: Figure 2.3 shows the *precision*, the *recall*, the *Fscore* and the *classification* scores as a function of increasing signal-to-noise ratio in pink noise.

(1) The upper right panel in Figure 2.3 shows the recall-values per formant. The recall-values remain high for all three formants for SNRs above ~ 10 dB but decrease rapidly when SNRs decrease. These results were evaluated further by evaluating the results of the extraction of harmonic complexes. We found that the harmonic complexes were often not, or not correctly extracted as demonstrated in Table 2.2. The table shows the occurrences of harmonic complexes that are not detected and the occurrences of harmonic complexes where the detected pitch makes an octave error as compared to the f0 annotations in Hillenbrand et al. (1995).

	SNR(EdB)	30	10	0	-4	-6
female	not extracted	0 %	1 %	18 %	35 %	51 %
	octave error	1 %	3 %	10 %	13 %	11 %
male	not extracted	2 %	8 %	41 %	74 %	81 %
	octave error	8 %	10 %	7 %	3 %	3 %
child	not extracted	0 %	1 %	17 %	39 %	51 %
	octave error	1 %	2 %	4 %	9 %	8 %

Table 2.2: Type of mismatch (octave error or not extracted) for detection of the harmonic complex for male, female and child speakers in pink noise. For male speakers more harmonic complexes are missed and more octave errors are made than for female or child speakers.

(2) The upper left panel in Figure 2.3 shows the precision-values per formant. Precision remains relatively high at all SNR values and for all three formants. These results imply that the formants that were extracted (the recall scores) were similar to the ground truth values f_{hit} .

(3) The F-score, depicted in the panel at the left bottom panel of Figure 2.3 shows the combined score for precision and recall.

(4) In the right bottom panel of Figure 2.3 the classification scores obtained with the BFT search algorithm are shown. Recognition in clean speech is 75% for all three speaker classes. In 0dB SNR, recognition for female speakers is 58% and recognition for male speakers 35%. Table 2.4 shows the confusion matrix of the classifications of the vowels from all speakers pooled together. Relatively much confusions can be found between the vowel sounds ae, eh and ah, aw. Those four vowels are confused with one of the other sounds for 25% percent of the vowels. The same vowels are reported to be confused most often by human listeners (Hillenbrand et al., 1995).

	ae	ah	aw	eh	ei	er	ih	iy	oa	oo	uh	uw	
ae	92	3	2	35	1	0	2	0	0	0	4	0	ae
ah	34	66	16	11	0	1	1	0	0	1	8	1	ah
aw	0	21	96	0	0	1	1	0	0	5	13	2	aw
eh	37	2	3	86	0	0	3	0	0	5	3	0	eh
ei	1	0	0	0	115	1	12	9	0	0	0	1	ei
er	0	0	2	0	3	129	1	0	1	3	0	0	er
ih	2	0	0	4	9	3	114	3	0	2	0	2	ih
iy	0	1	0	0	11	2	5	114	1	3	0	2	iy
oa	0	0	3	0	0	1	0	1	106	14	3	11	oa
oo	0	2	0	1	0	5	0	0	17	96	4	14	oo
uh	0	10	19	0	1	1	1	0	4	13	89	1	uh
uw	0	0	0	0	0	2	3	1	4	14	0	115	uw

Figure 2.4: Confusion matrix of classification task in clean speech pooled over all speaker classes.

Experiment 1: Discussion

We described and tested a method to automatically extract formants based on robust parts of the acoustic signal, namely the harmonics at formant positions. We showed that it is possible to extract formant values over SNRs from 30 dB to -6 dB in pink noise using the harmonics of a harmonic complex. Harmonics provide a solid basis for the extraction of formants if a harmonic complex can be extracted. The formants are important acoustical cues for the identification of phonemes. These initial results support the hypothesis that harmonic grouping can be used as a basis for speech processing. However, harmonic grouping based on the COCHL method as implemented in this study did not meet our needs as clearly visible Harmonic Complex (HC)s were too often not extracted.

One of the characteristics of vowels pronounced in isolation is the relative stability of the formant frequencies. The clearly pronounced vowels in H-vowel-D context, such as used in this study, shows relatively little formant movement. Therefore, *we consider the present results viable for situations with little formant movement*. However, such speech conditions are encountered relatively seldomly. Conditions with abundant formant movements are much more common because they are characteristic for the continuous speech encountered in everyday live. Formant movements not only characterise voiced parts of speech, but also provide context for the correct perception of other phonemes. As a result, the formant movement is important for the correct perception of continuous speech. Therefore, we also tested the algorithm on a database consisting of continuous speech (Section 2.2.3).

2.2.3 Experiment 2: Formants by interpolation in continuous speech

In Experiment 1 we tested a formant extraction algorithm on a database consisting of vowels pronounced in isolation (the AEV-database). We showed that formants were reliably extracted in different noise conditions, provided the harmonic complex was extracted. Also, we concluded that the results obtained from spoken vowels do not extrapolate to voiced components of natural speech, especially because formant movement is more common in natural speech. Therefore, we make three adjustments to the experimental set-up from Experiment 1. First, (1) we replace the COCHL method by the SPECT method (described in 2.2.3) as a means to extract harmonic complexes. This adjustment is made because the COCHL method did not meet our needs. Informal tests showed a high flexibility of the SPECT method to different recording and noise conditions. Second, (2) we test the formant extraction algorithm on a database where vowels are pronounced as part of continuous speech, because the estimation of formant tracks may be harder with increased formant movement. Third, (3) in order to interpret the results within the contexts of results from other approaches, we follow the experimental set-up of Gläser et al. (2010) as closely as possible. Similar to Gläser et al. (2010) we evaluate the performance of the formant extraction algorithm in babble noise.

Experiment 2: Algorithm

Peak interpolation over the harmonics in the harmonic complex is performed (Section 2.2.2) where the harmonic complex is determined with the SPECT method as described here:

Segment selection. The first step, to select segments, is illustrated in the upper left panel of Figure 2.5. Segments are determined by a method described by Violanda, van de Vooren, van Elburg & Andringa (2009). Two Fourier transforms (STFT) are calculated, one with high and one with low frequency resolution to reduce the effect of window parameters on the spectral and temporal resolution (see also Nakatani & Irino, 2004). From this we determine dominant points, the frequencies that dominate a few frequency bins in the TF plane using Otsu’s threshold selection method (Otsu, 1979). The second step, to apply an energy filter, is illustrated in the upper right panel of Figure 2.5. To extract segments, we apply a threshold of one standard deviation above the mean of the dominant point. This leads to dominant time-frequency tracks when the dominant points that pass the threshold are connected.

Preselection of segment-pairs. For all N segments, i.e. time-frequency tracks, the frame-numbers, corresponding frequencies and energy-values were given as input for the next step. Tonal segments with a minimal frequency of 50Hz were analysed for grouping. For every pair of segments the temporal overlap length was calculated. All pairs with a minimal overlap length of 10 frames were stored in an N -by- N matrix. Component pairs were captured when the frequency modulation error (ε_{fm}) did not exceed the threshold-value of 5% with ε_{fm} the frame-wise difference of the normalised frequencies

of the segments as calculated by:

$$\varepsilon_{fm}(\#fr, fqNorm_{seg}) = 100 \times \sqrt{\frac{\sum_{Nfr=1}^{\#fr} (fqNorm_1(Nfr) - fqNorm_2(Nfr))^2}{\#fr}}. \quad (2.13)$$

Nfr is the frame number and $\#fr_{seg}$ the number of frames of the segments (seg). $fqNorm_{seg}$ is the frame-wise normalised version of the segment frequency $freq_{seg}$ as calculated by:

$$fqNorm_{seg}(Nfr) = \frac{freq_{seg}(Nfr)}{AF_{seg}}, \quad (2.14)$$

and

$$AF_{seg} = \frac{\sum_{Nfr=1}^{\#fr} freq_{seg}(Nfr)}{\#fr_{seg}}. \quad (2.15)$$

Conflict based grouping. The grouping algorithm is based on the calculation of conflicts. Every criterion leads to a conflict matrix with conflicts documented for all segment-pairs that should not be clustered. In the current experiment we used harmonicity as a criterion. Signal-components are considered not belonging to the same group when harmonic error-values (ε_{harm}) exceed a threshold-value of 5%, where ε_{harm} is calculated as the difference of the average frequency AF_{seg} of two segments after dividing by their optimal fractions, as calculated by:

$$\varepsilon_{harm}(AF_{seg}, N, D) = 100 \times \sqrt{\frac{\sum_{Nfr=1}^{\#fr} (AF_1/N - AF_2/D)^2}{\#fr}}, \quad (2.16)$$

with N and D the reduced numerator and denominator of the fraction $\frac{AF_1}{AF_2}$. The resulting conflict matrix is the basis for clustering to groups.

Segment-pairs are processed one by one, starting with the pair with the lowest ε_{harm} . By doing so groups are gradually filled with harmonically related segments (step 3, bottom-left in Figure 2.5). The groups are tested on octave-errors; the more harmonics a group consists of, the more reliable the fundamental frequency can be determined. Based on the mean fundamental frequency of the groups remaining, isolated segments are grouped and their energy at harmonic positions is inserted (step 4, bottom-right in Figure 2.5).

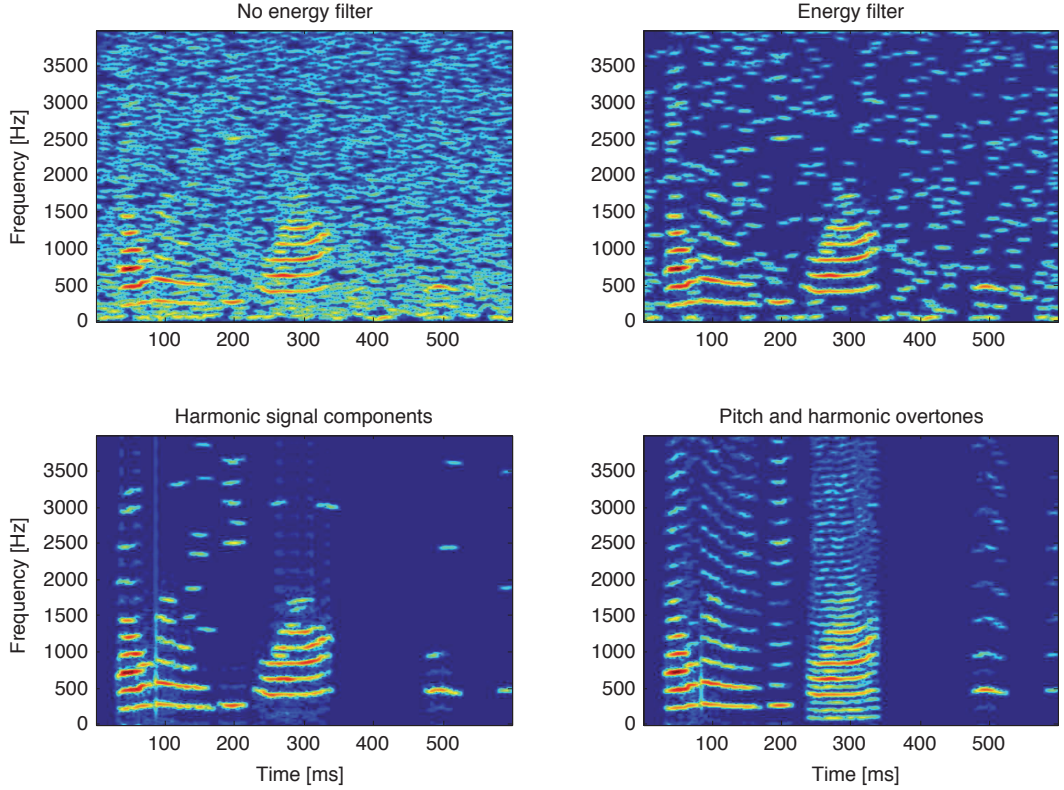


Figure 2.5: Time-frequency representation of the stimuli. The example shows the resynthesised stimuli, obtained from the 0dB pink noise conditions, of the sentence "Elderly people are often excluded" (Timit faks0 sx43). Upper left panel: Segments / time frequency tracks. Upper right panel: Time frequency tracks after removing the less energetic segments. Lower left panel: Filtered segments that are harmonically related. The regions without harmonically related segment were filled with silence. Lower right panel: Energy at harmonic positions of the pitch resolved from the harmonically related segments. The harmonic complex that starts at 250 ms shows the effect of an octave error. The octave error is probably due to the tiny noise-related extractions (visible in the lower left panel) that are extracted at positions in between two harmonics.

Experiment 2: Material

The formant extraction algorithm was evaluated on the test set of the VTR-Formant database (Deng, Cui, Pruvencok, Chen, Momen & Alwan, 2006). This database consists of recordings of 34 utterances spoken by male speakers and 56 utterances spoken by female speakers. These sentences are automatically annotated and subsequently hand-corrected for the frequency values of the first three formants (Deng et al., 2006) that serve as a ground truth to evaluate our algorithm.

Following Gläser et al. (2010) we added multi-speaker babble noise to the clean speech signal, in decreasing signal-to-noise ratios (SNRs), from 30 dB to -6 dB SNR (30 dB, 15 dB, 12 dB, 9 dB, 6 dB, 4 dB, 0 dB, -3 dB, -6 dB). We estimated the signal-to-noise ratio while excluding the non-speech samples from the energy calculation. Following the procedure taken by Gläser et al. (2010) we used the exact start and end of the

speech sample from the phonetic annotations that are given with the TIMIT (The Texas Instruments and Massachusetts Institute of Technology) database, the database that provided the sentences for the VTR-Formant database.

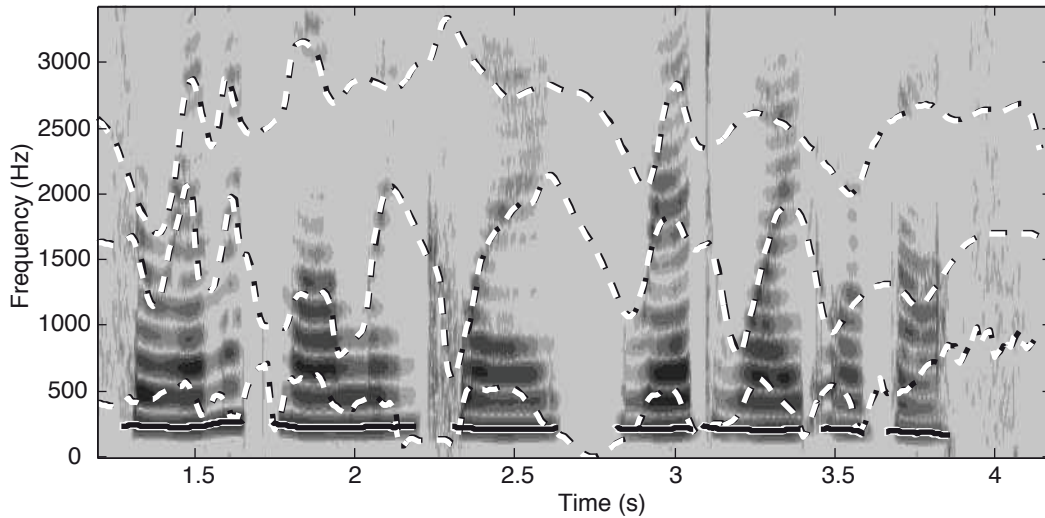


Figure 2.6: Illustration of the annotations and extractions in continuous speech. The *white dotted lines* show the annotations of the formant tracks that served as a ground truth to test our algorithm. The *black solid lines* show harmonics extracted from continuous speech by the SPECT method. Only these time-regions are taken into account in the evaluation of the formant extraction algorithm. The *sentence* that was used for this figure is labeled "si756" where a female speaker reads "Materials: Modelling clay, red, white or buff."

Experiment 2: Evaluation

Because the formant annotations in the VTR-database do not take into account voicing, the annotations are given throughout the whole sound file, including positions where no speech or no voiced speech is recorded (this is shown in Figure 2.6 by the white dotted lines). To obtain a clear understanding of the suitability of the formant algorithm, we evaluated the formant tracks at the voiced speech positions only. We defined "voiced speech positions" as the samples where a harmonic complex was extracted from clean speech (as illustrated in Figure 2.6 by the black solid lines). This way we obtain an understanding of the formant extraction algorithm with least confoundment from possible faults in the harmonic complex extraction stage.

For the goal to evaluate both usefulness for classification and robustness to noise we define three measures. Usefulness for classification is evaluated by the consistency (the distance and the relative number of extra tracks) of the extracted and annotated formant tracks, ($f.track_{extr}$ and $f.track_{ann}$, respectively). Additionally, in order to evaluate the robustness to noise we determine the similarity between the formant tracks extracted from noise $f.track_{extr(noise)}$ and the formant tracks extracted from clean speech conditions $f.track_{extr(clean)}$. The corresponding measures are explained in more detail below.

General efficiency: We specify two measures to determine the consistency of $f.track_{extr}$ and $f.track_{ann}$ in the different noise conditions. Together these measures indicate how useful the $f.track_{extr}$ are for classification based on a low error-score and a low false-positive rate.

(1) In accordance with the error measure reported by Gläser et al. (2010) we calculated the absolute errors normalised by the annotated formant frequencies. The error was calculated over the mean formant frequency of both $f.track_{extr}$ and $f.track_{ann}$;

$$error(noise, x) = \frac{f.track_{ann}(x) - f.track_{extr}(noise)}{f.track_{ann}(x)}, \quad (2.17)$$

and its value was computed for clean speech and the different noise conditions. The $error$ is calculated per formant-track (x , ranging from 1 to 3) and for the best matching extracted formant track, $f.track_{extr}$. The latter choice, to calculate the $error$ for the best matching $f.track_{extr}$ can, in case of a missed extraction, lead to an increased error-score when a spurious peak is erroneously handled as formant (illustrated in figure 2.8).

(2) Additional to the $error$ we calculated the number of false positive extractions (spurious peaks). A high number of spurious peaks is less successful for classification purposes. We determined the number of spurious peaks per harmonic complex;

$$\#peaks_{spurious}(noise, HC) = \#f.track_{extr}(noise, HC) - \#f.track_{ann}(HC). \quad (2.18)$$

where the number of annotations $\#f.track_{ann}(HC)$ is always three and HC ranges from one to the maximum number of extracted harmonic complexes in that particular speech file. The number of spurious peaks is calculated for clean speech and the different noise conditions.

Robustness to noise: Calculation of the $recall$, the fraction of correctly detected formant tracks in clean speech conditions, that remains available in noisy speech conditions, is given by

$$recall(noise) = \frac{\#(f.track_{extr}(noise) \cap f.track_{hit})}{\#ft_{hit}}, \quad (2.19)$$

where

$$f.track_{hit} = f.track_{extr}(clean) \cap f.track_{ann}. \quad (2.20)$$

Here, an $f.track_{hit}$ is an extracted formant track $f.track_{extr}$ whose value is similar to the value of an annotated formant track $f.track_{ann}$. The accepted value difference between $f.track_{ann}$ and $f.track_{extr}$ is 15% for the 1st formant, 12% for the second formant and 8% for the third formant (congruent to Section 2.2.2).

Experiment 2: Results

General efficiency: Formant-wise error-scores (evaluation measure 1) are given in the left panel in Figure 2.7. The error scores are relatively independent of SNR. Error scores are approximately $\sim 35\%$ $error(f_1)$, $\sim 10\%$ $error(f_2)$ and $\sim 16\%$ $error(f_3)$. These percentages equal a maximum error of 280 Hz, 140 Hz and 460 Hz respectively as calculated for the vowel with the highest formant value. Visual inspection of the results reveals that the relatively high error scores can be attributed to (a) undetected formant movements and (b) characteristics of the annotations.

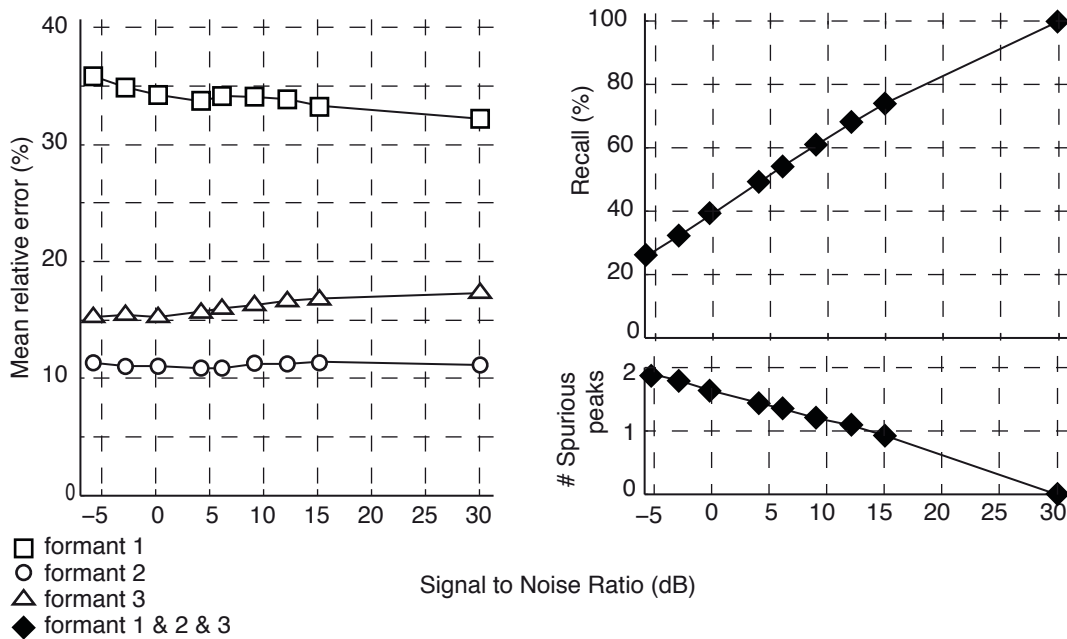


Figure 2.7: Results of formant extraction in noise tested on a continuous speech database. Left panel: Mean relative error as distance measure for extracted and annotated formants. Upper right panel: Percentage of extracted formants (relative error falls within range; 15%, 12% and 8% for f_1 , f_2 and f_3 respectively). Lower right panel: Mean number of spurious peaks. Lines are added to guide the eye.

- (a) An example of undetected formant movements is presented in Figure 2.8 (inlay "Undetected Formant Movement"). In the inlay the word "clay" is shown. The voiced diphthong at the end of the word $[\epsilon i]$ is characterised by a transition towards the vowel $[i]$. The typical formant movements for this transition are a downward slope for the first formant from $f_1[\epsilon i]$ is 659 Hz towards $f_1[i]$ is 399 Hz, (Adank, van Hout & Smits, 2004) and an upward slope for the second formant from $f_2[\epsilon i]$ is 2097 Hz towards $f_2[i]$ is 2276 Hz, (Adank et al., 2004). The start-point and end-point of the first formant and the middle part of the second formant are extracted congruent to this analysis. However, the formant movement itself is not extracted in this example.

- (b) An example of a characteristic of the annotations that confounds the error scores is presented in Figure 2.8 (inlay "unvoiced speech"). The inlay shows the transition from the word "modelling" to the word "clay". Here, formants are annotated despite the fact that part of the transition is unvoiced (the phoneme [k] is unvoiced). This is an artefact of the annotation algorithm because formants only describe the vocal tract resonance at voiced speech positions. In the current example, this artefact results in a steep drop in frequency for the annotated formants, also at voiced speech positions. Because the formants of our algorithm are not similarly influenced, such incongruencies in the annotations result in increased error rates.

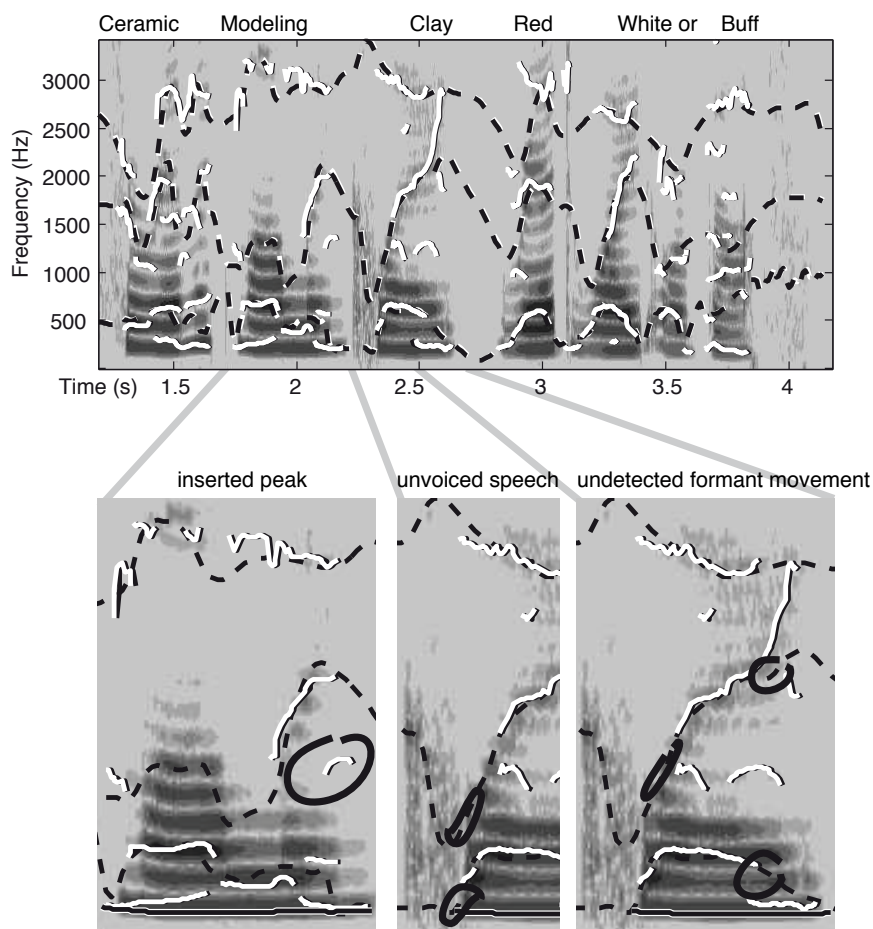


Figure 2.8: Three anomalies in formant extraction by interpolation over harmonics. The white solid lines show formants extracted from continuous speech through interpolation over harmonics. The black dotted lines show the formants that were used as ground truth. Sentence: "Ceramic modelling clay, red, white or buff". Inlays illustrate anomalies:

Inlay 1 (inserted peak): The circle indicates an additional peak extracted from a nasal speech sound.

Inlay 2 (unvoiced speech): The circles indicate annotated formant movements close to unvoiced speech.

Inlay 3 (undetected formant movement): The circles indicate formant movements that seem to be correctly annotated but is not extracted.

The mean number of spurious peaks per HC (evaluation measure 2) is given in the lower right panel of Figure 2.7. Mean $peaks_{spurious(noise,HC)}$ increases from 0 towards 2 when noise increases from SNR of 30 dB towards SNR of -5 dB. Visual inspection of the results reveals that the spurious peaks can be attributed to (A) the target speech and (B) the noise.

- (A) Spurious peaks related to the target-speech can be due to low energetic frequency regions of voiced speech incorporating local non-formant peaks. This is the case, for example, in the spectral zero regions of nasals such as illustrated in Figure 2.8 (inlay "Inserted Formant"). Such peaks can be assessed by criteria related to the relative energy in the harmonic complex.
- (B) Spurious peaks related to the noise signal result mainly from harmonic complexes in the babble noise. Especially in lower SNRs where the noise level can exceed the target speech energy, the probability becomes high to extract formants originating from the babble noise instead of the target speech and these will be classified as spurious peaks. In this work we only focused on grouping by the principle of harmonicity. Sequential grouping principles might help to select HCs that are likely to belong to the same speaker and as such lead to a reduction of spurious peaks. However, at this moment the sequential integration of harmonics does not have our primary focus, as we first want to understand the relevant characteristics of speech representations.

Robustness to noise: Recall calculated over all formants is given in the upper right panel in Figure 2.7 as a function of noise. The *recall* decreases from 100% in an SNR of 30 dB to 25% in an SNR of -6 dB.

Experiment 2: Discussion

We described a method to automatically extract formants based on interpolation over harmonics. Initial results on a database with vowels (Experiment 1, Section 2.2.2) were promising. Because harmonic grouping based on the COCHL method did not meet our needs we replaced it by the SPECT method while retaining the rest of the algorithm, and performed a second experiment. The SPECT method has the additional advantage that it also permits us to evaluate the results of the algorithm on continuous speech that is more commonly used in everyday life than vowels. In contrast to formants in vowels, continuous speech is characterised by formant movement.

Formant extraction from vowels versus continuous speech

Two differences are found when comparing the results obtained from continuous speech with results obtained from vowels:

- (1) Formant extraction as measured by the *error*, especially for the first formant, is less precise in continuous speech. The mean error of f_1 was $\sim 35\%$ (280 Hz) for continuous speech, where the accepted error was set to 15% (95 Hz) for f_1 in the vowels experiment ($\sim 95\%$ consistent extractions). Similarly, the mean error of f_2

was $\sim 10\%$ (140 Hz) for continuous speech, where the accepted error was set to 12% (316 Hz) for f_2 in the vowels experiment ($\sim 90\%$ consistent extractions). The higher error for continuous speech is explained by the increase of formant movements in continuous speech. Visual inspection of the results showed that (a) formant movements remained undetected more often than steady formants and (b) the annotations algorithm had led to exaggerated formant movements near voiceless positions.

- (2) The second difference between the results of the continuous-speech experiment and the vowels experiment concerns the *recall* that remains relatively high in the continuous-speech experiment as compared to the recall in the vowels-experiment, even in heavy noise conditions such as SNR of -6 dB. We attribute this performance gain to the change to the SPECT method for harmonic grouping.

Formant extraction methods compared

In order to be able to compare our results to the findings of Gläser et al. (2010) we tested our algorithm on the same database in similar conditions. The error values in clean speech conditions are higher than the ones reported by Gläser et al. (2010) which implies that our extractions are less well related to the reference annotations. The error scores of both methods converge in SNR of 0 dB and lower. The current extractions are more stable than the ones obtained by Gläser et al. (2010). It can be concluded that local features based on harmonics that belong to a harmonic complex are highly stable in noise. However, the current extractions are not clearly related to the reference annotations. An important reason for this seems to be that the current algorithm is rather vulnerable to formant movements, which increases the error scores. Additional criteria may tweak the results to better fit the annotations. However, it does not bring us closer to the goal to better understand robust speech representations and robust speech processing.

Usefulness of formant features for speech recognition

In contrast to other approaches for formant extraction, the current algorithm extracts formants only when harmonic complexes are extracted. An extracted formant indicates that the input has a positive local signal to noise ratio and can be voiced speech. However, the algorithm does not in all cases extract exactly three formants; formants can be missed and extra peaks are inserted. For current approaches in speech recognition such input features can not be successfully processed. In order to investigate the usefulness of such input features, we investigate and discuss in Chapter 3 how human listeners reason with missing formants or added peaks.

Conclusions: *The current approach to extract features based on a harmonic complex leads to robust representations that are related to formants.* However, the current estimation method to determine speech representations does not satisfyingly capture the formant movements, we developed (Section 2.2.4) an additional algorithm to also capture these regions. Another disadvantage of the current method is that formants are not always extracted by the algorithm and sometimes additional peaks are extracted.

Because current technology is not developed for such relatively unstructured input we investigate (Chapter 3) how human listeners reason with this type of information.

2.2.4 Experiment 3: Formants on peak locations in continuous speech

In Experiment 1 and 2 we showed that formant-related speech components were robustly extracted from the harmonically related tonal components in the speech signal with added pink and babble noise in different SNR levels. However, formant movements, the most important characteristic for correct identification of speech sounds, were not satisfyingly estimated. Therefore, we propose an algorithm to extract different characteristics from the harmonic complex.

In Experiment 1 and 2 the vocal tract resonances are estimated by interpolation because they are not directly available in the tf-plane as formant tracks. These formant tracks are indicated by lines in Figure 2.9 (left panel). The vocal tract resonances sampled by harmonics are directly available in the TF-representation. Namely, the energy in the vocal tract resonances protrudes the TF-representation only at frequency locations where harmonics function as carrier for the resonances. In Figure 2.9 (right panel) these locations are indicated by circles.

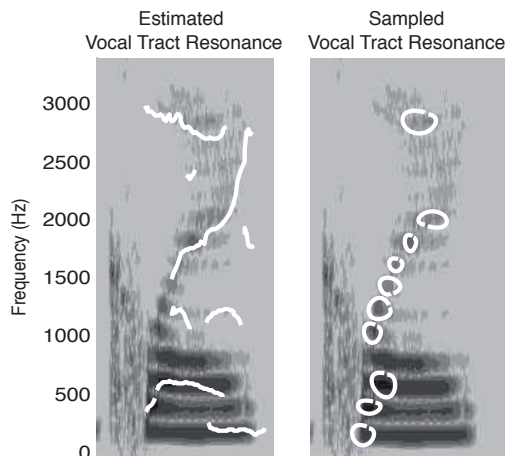


Figure 2.9: Estimated and sampled vocal tract resonance. Visual example of the difference between an estimated VTR (formants) such as used in phonetics research and sampled VTR such as available in the signal.

The sampling of the vocal tract resonances changes with the natural fluctuations in the pitch of speech, resulting in variable representations of further similarly pronounced phonemes. The interpolation approach is valuable for the goal to describe speech because formants derived from the same phoneme are only minimally affected by these pitch changes. This is schematically illustrated in Figure 2.10 showing how interpolation can lead to similar extractions when the pitch of the speaker is doubled. In theory, this is an advantage of the interpolation approach. However, the findings of Experiment 2 show that formant movements are not optimally estimated. Therefore, we adopt a different, intuitive approach to extract robust representations of speech. We extract the

energetic structures. Intuitive because the high energetic components (ECs) are available in many types of noise and in SNR levels of different severity.

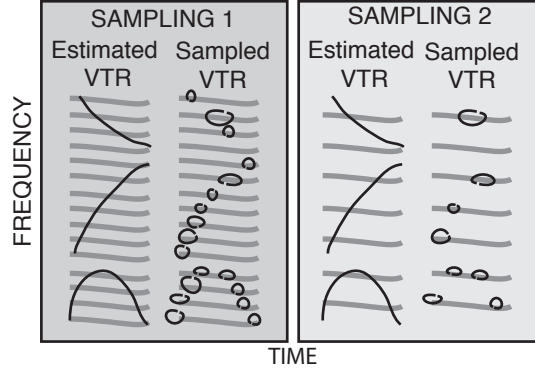


Figure 2.10: The effect of sampling at different pitches is limited when formants are estimated. This schematic representation of the sound (given as t-f representation in Figure 2.9) gives a visual example of effect of sampling of the VTR by harmonics when belonging to speech with different pitch. The effect is dissimilar for formants estimated by interpolation and for high-energy structures.

In voiced speech the ECs are the components where the vocal tract resonances co-occur with harmonicity. These components represent the vocal tract resonances sampled by intervals in the frequency domain. In contrast to the findings for interpolated formants, we expect this representation to profit from formant movements. In Figure 2.10 we illustrate how formant movements cross the harmonics of the HC, which eventually leads to high energy locations within the harmonics at movement positions.

Experiment 3: Algorithm

In this Experiment we used the output of SPECT (as described in Section 2.2.3) to estimate the Vocal Tract Resonance (VTR) at locations where they co-occur with harmonic strands (the term strand is defined by Cooke (1993): "Each strand aims to define the time-frequency behaviour of a single spectral component."). The resulting Energetic Components (EC)s are related to formants, where formants estimate the VTR and the ECs represent the VTR as sampled by harmonics.

An energetic component, EC , is defined as the segment of a harmonic strand where the frame-wise difference between $HE(harm, fr)$; the frame energy of the harmonic and $HT(harm)$; a threshold calculated by the harmonic energy, is higher than 0:

$$EC(harm, fr) = (HE(harm, fr) - HT(harm)) > 0, \quad (2.21)$$

with the threshold $HT(harm)$ taken as two standard deviations above the mean energy of the harmonic strand $HM(harm)$. This threshold was chosen upon visual inspection of the extractions, such that the co-occurrence of VTR and harmonics were captured in clean and noisy speech.

$$HT(harm) = HM(harm) + 2HS(harm), \quad (2.22)$$

with mean harmonic energy $HM(harm)$ calculated by

$$HM(harm) = \frac{\sum_{fr=1}^{\#fr} HE(harm, fr)}{\#fr_{harm}}, \quad (2.23)$$

and standard deviation of the harmonic energy $HS(harm)$ calculated by

$$HS(harm) = \sqrt{\frac{1}{\#fr} \sum_{fr=1}^{\#fr} (HE(harm, fr) - HM(harm))^2}. \quad (2.24)$$

Figure 2.11 shows an example of the final step in the adjusted algorithm. The figure depicts the energy development for each extracted harmonic of a HC. For each of the harmonics the mean and spread of the energy are calculated (depicted in the inlay in Figure 2.11 for the 14th strand).

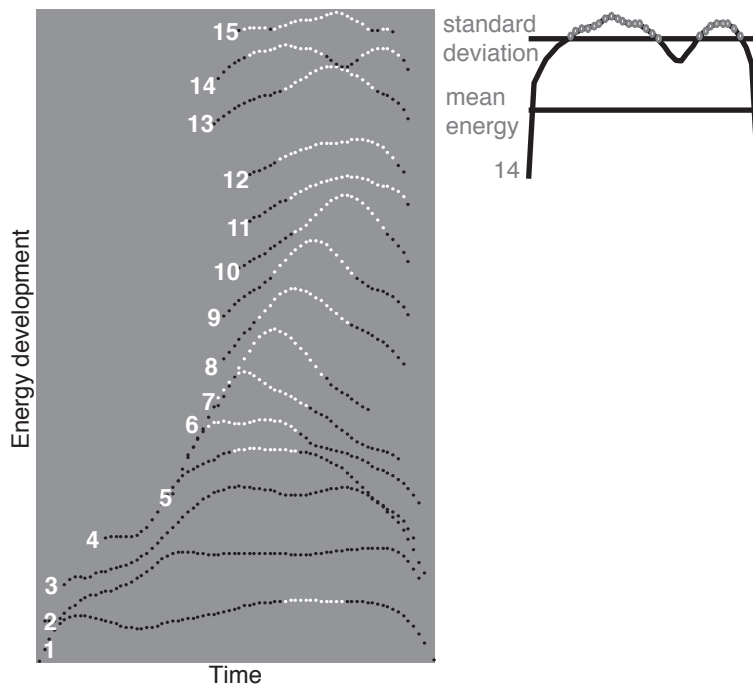


Figure 2.11: Visual example of the algorithm to calculate energetic components. The energy in the subsequent harmonic strands is depicted by the relative height of the dotted lines in the Figure. All dots with an energy exceeding one standard deviation of the mean energy of that strand are white, the other are black. An example of this calculation is given for the 14th harmonic strand at the upper right part of the figure.

Experiment 3: Material

The VTR-Formant database and the same noise type and noise levels, described in Section 2.2.3, are used for this experiment.

Experiment 3: Evaluation

The algorithm selects the highest energy regions in an HC as calculated per harmonic strand. These are the locations where vocal tract resonance and harmonic strand co-occur. The extractions can not be numerically compared to the formant annotations because they are sampled versions of the vocal tract resonance. Therefore, we focus on the stability of the extractions in noise and assume the extractions to be informative (we will elaborate further on this in Section 2.2.5).

Robustness to noise, harmonic complex: Recall, the fraction of clean speech target elements that remains accessible in noisy speech conditions, is calculated for the second last step (HCs 2.25) and the last step (ECs 2.26) in the algorithm. The recall of the HCs is calculated for different noise conditions by

$$\text{recall.HC}(\text{noise}) = \frac{\#(\text{HC}_{(\text{noise})} \cap \text{HC}_{(\text{clean})})}{\#\text{HC}_{\text{clean}}}. \quad (2.25)$$

Robustness to noise, energetic components: The recall of the ECs is calculated similarly by

$$\text{recall.component}(\text{noise}) = \frac{\#(e.\text{component}_{(\text{noise})} \cap e.\text{component}_{(\text{clean})})}{\#e.\text{component}_{\text{clean}}}. \quad (2.26)$$

Recall of the ECs is calculated for all intervals for which we could estimate an HC in the noise conditions. As such we determine the robustness of both the HCs and the ECs.

Experiment 3: Results

Figure 2.13 shows the recall of both the HCs and ECs. The measurements at an SNR of 30 dB were taken as a reference and as a result they are set to 100% for both the recall of HCs and the recall of ECs. The combined recall at -6 dB SNR is ~25%; ~50% of the HCs is recollected and ~50% of the ECs is recollected with the recollected HCs as an input.

Robustness to noise, harmonic complex: Recall of the HCs is given in the left panel of Figure 2.13. The recall declines from 100% in 30 dB SNR to ~50% in -6 dB SNR with the slope of the curve decreasing with increasing noise. In order to improve our understanding of the results an example of an input sound (first panel) and the extractions (second and third panel) is given in Figure 2.12. In the second panel it can be seen that all HCs are extracted for this sentence in SNR of 30 dB, these extractions function as a reference or base-line extractions. In SNR of 4 dB incorrectly extracted HCs are indicated by opaque white regions. The HC is marked as not being recalled at these locations. In this example half of the HCs is correctly extracted, which is congruent with the reported recall of HCs at approximately 55%.

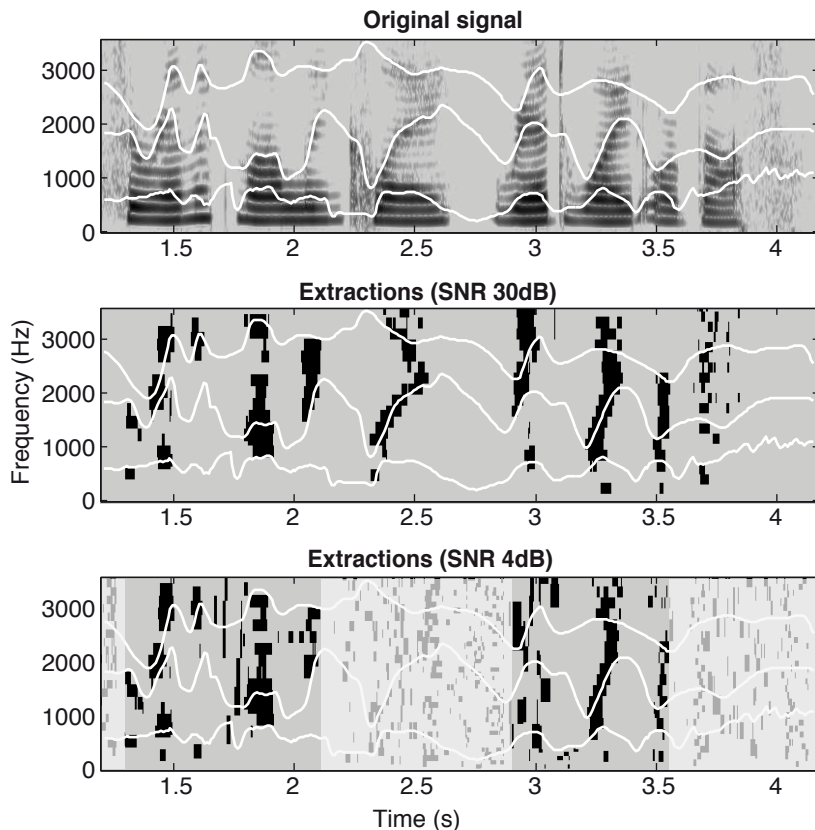


Figure 2.12: Energetic components in clean and babble noise continuous speech. White lines: annotated formants. Black spots: Extracted energetic components from SNR of 30 dB (middle panel) and 4 dB (lowest panel). The white regions in the lowest panel indicate locations where the harmonic complex was not correctly (e.g. not with the original fundamental frequency) extracted.

Robustness to noise, energetic components: Recall of the EC is given in the right panel of Figure 2.13. The recall declines from 100% in 30 dB to $\sim 50\%$ in -6 dB with an increasing decline rate when noise increases. In Figure 2.12 it can be seen that the shape of the extractions remain similar when the HC is correctly extracted. An important effect of the noise is the narrowing of extractions. For SNR of 4 dB we found a recall of the EC (given a correctly extracted HC) of approximately 75%.

Experiment 3: Discussion

We described and tested a method to automatically extract local, speech-related elements based on HCs. Because formants estimated by interpolation do not accurately represent formant movements (Section 2.2.3) the interpolation step in the algorithm was replaced by a step to extract local, energetic components. We extracted the time-frequency locations where vocal tract resonance and harmonic strand co-occur.

Evaluation of energetic components

Noise leads to a decrease in recall scores for the ECs. This can be explained by two

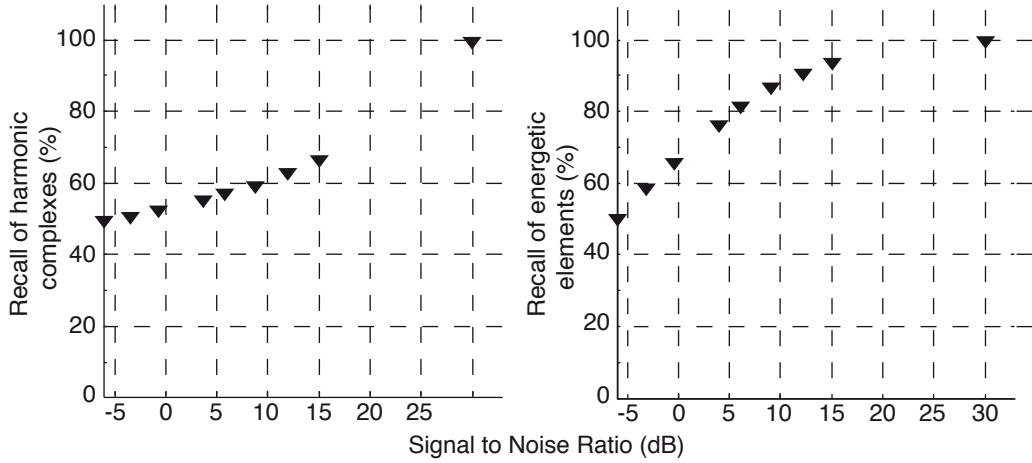


Figure 2.13: Extraction of energetic components from babble noise tested on a continuous speech database. Left panel: Exact recall of the energetic components in babble noise given the availability of harmonic complexes. Right panel: Recall of the harmonic complexes in noise. The 30 dB SNR condition was added as a near-clean condition.

factors. *First*, some HCs are extracted at a deviant pitch in noisy speech conditions, resulting in changed extractions of the ECs. *Second*, noise leads to a narrowing of the extractions when noise levels increase as a result of a higher mean energy and lower deviation in energy fluctuations. As a result, the extractions become narrower in the frequency direction; smaller parts of the harmonic chain are extracted. Because the narrowing extractions retain the original spatial relations to each other, this decrease in recall scores does not necessarily go together with a decrease in usefulness of the representation.

Formant features and energetic components compared

With this experiment we showed that the robustness to noise is similar for the ECs and the interpolated formants. For both the HCs and the ECs the recall declines from 100% in 30 dB SNR to 50% in -6 dB SNR. This is similar to the recall of formants by interpolation, where 25% of the formants remained accessible. However, the accepted error was much higher for the interpolated formants (15%, 12% and 8%) than for the ECs (exact recall); the ECs were more precisely recalled. Also, the findings suggest that formant movement is better captured by the shape of the ECs than by the interpolated formants.

Usefulness of energetic components for speech recognition

We could not relate the extracted ECs to a ground truth to evaluate the usefulness for classification. However, we expect them to be useful for classification because of three characteristics of human speech perception that are related to the extractions. (1) The energetic part of a harmonic originating from speech is generally the interception of a vocal tract resonance and a harmonic strand. The VTRs are known to be useful for classification (Barker, 1998) and (2) the selections of the VTR that are carried by harmonics are directly accessible to human listeners in the cochlear representation. Analogous to this sampling of the VTR by harmonics in auditory perception, visual

perception successfully deals with sampled objects; target objects that are partly occluded by non-target objects. Furthermore (3), the VTR and, in effect, the ECs, are highly correlated to phonemes which is a necessary prerequisite for the learnability of a representation. Therefore it is viable that a VTR-like representation is learned from the sampled VTR input.

Similar to the findings for interpolated formants the algorithm does not extract the same type of features at all spatio-temporal locations, as is needed for existing speech recognition systems. Extractions can be missing or noise induces extra peaks. For current approaches in speech recognition such input features can not be successfully processed. All commonly used statistical methods based on linear algebra use a fixed number of feature-vectors. In order to investigate the usefulness of such input features, we investigate and discuss in Chapter 3 how human listeners reason with missing and added peaks.

Conclusions: *The energetic components are robust to noise and capture characteristics that can be related to formant movements in both clean and noise conditions.* The usefulness of ECs for speech recognition purposes can not be explicitly tested. Qualitative evidence suggests that the ECs capture the relevant aspects of the time-frequency plane. However, because current technology is not suitable for this type of input features we will investigate (Chapter 3) how human listeners reason with this type of information.

2.2.5 Local features built on a harmonic complex as alternative representations of speech

In this chapter (Chapter 2.2) we introduced and evaluated two algorithms to extract features for a noise-robust representation of speech. Both algorithms utilise the robustness of harmonics in a harmonic complex. Harmonics are noise robust because they are characterised by relatively high energy levels. Grouping, the combination of harmonically related tonal components, leads to a robust representation of the original speech sound. Human listeners apply primitive principles, such as grouping of tonal components, as a preprocessing method in sound perception (Bregman, 1990) and grouping can function as one of the factors that leads similarly to robustness in speech processing. Besides using the robustness of harmonic complexes, both algorithms utilise the robustness of local structures. It was concluded in Section 2.1 that existing robust speech representations focus on the extraction of local structures. Local structures are considered noise robust because (1) they can be chosen such that they exhibit relatively high energy levels and (2) they are relatively independent of effects at frequency locations other than the frequencies associated with the feature. By selecting local structures from harmonic complexes we combined two qualities that are related to noise robust representations.

The first algorithm, based on interpolation over harmonics, was tested on a vowels database with added pink noise (Section 2.2.2) and on a continuous speech database with added babble noise (Section 2.2.3). We obtained results similar to those obtained by Gläser et al. (2010), using the same database and similar noise. The second algorithm, based on energy in individual harmonics, was tested on the same continuous speech database and babble noise (Section 2.2.4). With this second algorithm we obtained

better results in terms of robustness to noise, but the results could not be evaluated in terms of usefulness for classification. It can be concluded that it is possible to develop an automatic method to extract features that are stable over SNRs from 30 dB to -6 dB by using the robustness of harmonics and the concept of locality.

The interpolation approach often fails at formant movement regions which is the most important characteristic for phoneme identification. As a contrast to this the ECs profit from formant movements. Also ECs are more robust to noise than interpolated formants; recall is more precise. However, a difficulty with the ECs is that the quantitative evaluation on usefulness for classification is not possible. Therefore, we need to rely on a qualitative evaluation. The ECs represent vocal tract resonances that are sampled by a carrier; the harmonics in a harmonic complex whose frequencies change with changing pitch. Vocal tract resonances are considered sufficient to support perception of the linguistic message (Barker, 1998; Remez et al., 1981) and shown appropriate for a phonetic description of speech (Bladon & Lindblom, 1981; Diehl & Lindblom, 2004). This characteristic is captured in the ECs and represented as a sampled version of the VTR.

The well-known Glimpses (Cooke, 2006) are similar to ECs, in terms of theoretical foundation and in terms of robustness. In terms of theoretical foundation both representations are defined as structures that exhibit relatively high energy levels and as a result, both representations are robust to noise. An advantage of the ECs over the Glimpses is that the ECs are extracted without prior knowledge of the noise. However, ECs only incorporate structures that are part of the harmonic complex. Opposed to this glimpses incorporate structures of the complete TF-representation such as pulses, that are not included in the harmonic complex. It is shown that glimpse-like extractions can be effectively used for automatic speech recognition (Gemmeke & Cranen, 2009; Gemmeke et al., 2011) and it is suggested that features similar to glimpses may be a basis for human speech perception (Cooke, 2006). An advantage of ECs is that they are extracted without prior knowledge of the noise. If ECs represent the same elements of speech as Glimpses do, they can be potentially powerful representations for speech processing. They may provide the bootstrapping seed in an interactive bottom-up, top-down approach such as proposed by Barker et al. (2005). This way robust representations of vowels may even function as an anchor for the extraction of other speech characteristics that are less robust to noise.

2.3 Robust key-word spotting: Global speech representation

From experiments 1 to 3 it was concluded that local features can be chosen such that they exhibit high energy levels. We presented an algorithm based on this analysis but the resulting detections could only be qualitatively evaluated. Therefore, it is not clear whether the extractions are useful for the recognition of speech. It was shown that both the recall of harmonic complexes and the recall of ECs decreases. Therefore, because we can evaluate the relation of the recall of HCs to recognition. In this chapter (Chapter 2.3) we investigate, by resynthesis of the sound in concatenation with a standard recogniser, the usefulness of HCs for recognition of speech in noise.

2.3.1 Experiment 4: The effect of selecting voiced-speech-components on the performance of an HMM-based classifier

Experiment 4: Introduction

Dusan & Rabiner (2005) assign the shortcomings in ASR technology to three causes: the utilised method of speech representation, the incomplete language and context models, and the still limited processing capacity of modern computer systems. Here, we address the problem of the representation of speech. The speech representations used in ASR are mainly focused on describing the spectral envelope. Effective for clean speech, but in a more natural setting interfering sounds (noise) will affect the spectral envelope of the target speech. As a consequence, the derived speech representation is disrupted when noise affects the speech signal. Noise subtraction techniques (see Gong, 1995, for a review) reduce this problem, but their effectiveness is limited to predictable (primarily steady-state) noise conditions (Cooke et al., 2001) and even than the estimated noise disrupts the spectral envelope (Gong, 1995).

Human listeners, in contrast, focus on the spectro-temporal structures of the target speech. When it comes to speech recognition, human performance is superior to that of many ASR systems in both quiet and degraded environments (Lippmann, 1997). This is even more so when the noise is non-stationary (Miller & Nicely, 1955). From the field of auditory scene analysis (ASA), humans seem to perceive segments of sounds as belonging together when they comply to Bregmans grouping principles such as continuity, common onset, common fate or harmonicity (Bregman, 1990). In a computational system these principles can be used as well to mimic human ASA (this is called "computational ASA" or CASA; Wang & Brown, 2006) and as such serve as front end for an ASR system to improve the quality of input speech (in Cooke et al., 2001, it is argued that methods based on ASA can function as an alternative to noise subtraction techniques).

CASA comprises at least a segmentation stage and a grouping stage. In the segmentation stage, segments are usually defined as the areas with a positive local SNR in a time-frequency (TF) representation of the acoustic signal. In the grouping stage, the segments are grouped together, mainly by applying Bregman's grouping principles. Ideally, each group corresponds to one sound source and, as a consequence, the spectral enve-

lope derived from the target group is considered to contain target energy predominantly. However, this CASA-derived spectral envelope still deviates from the ideal spectral envelope for three reasons: (a) only energy caught into segments have been conserved; (b) as the grouping process is far from trivial, grouping errors might be numerous; and (c) target energy masked by interfering sounds, most likely under severe noise conditions, can not be recovered. Where the last issue is an inevitable consequence of the sound-mixing process, appropriate algorithms for segment extraction and segment grouping can minimise the effects of the respective former two issues.

In this study we investigated the effects of segmentation and grouping on the discriminability of clean speech and speech in noise. The effects of segmentation (a) and grouping (b) are quantified by comparing phoneme classification results on segmented-variant samples, and segmented-and-grouped variant samples obtained from clean speech, with the results obtained from the original clean speech samples as a control condition. If the extraction of segments or groups leads to improved quality of input speech, the corresponding recognition results are expected to exceed those of the control condition. The effect of masked target energy (c) is objectified by comparing the phoneme classification results on segmented-variant and segmented-and-grouped variant samples obtained from speech in noise, with the classification results obtained from the original speech in noise samples as a control condition. The addition of noise changes the spectral envelope, resulting in decreased recognition scores for the original speech samples. Segmentation, by selecting areas with a positive local SNR, is expected to delay this process by delaying the change in spectral envelope.

Experiment 4: Method

Resynthesis techniques are available to transform a set of segments in the time-frequency plane to an acoustic signal (Cook, 2002). Resynthesised segments were used as input to a state-of-the-art recogniser (HTK) to investigate the effect of segmentation and grouping on the representation of clean speech and speech in noise. The segments were the results of four different stages in a CASA process: (1) extraction of segments, (2) application of an energy filter on the segments, (3) selection of harmonically related segments, and (4) addition of missing harmonics based on the pitch that was determined from (3). These basic steps are illustrated in Figure 2.5 from left up to bottom right.

Experiment 4: Algorithm

The algorithm is described in Section 2.2.3.

Experiment 4: Database

The Texas Instruments and Massachusetts Institute of Technology (TIMIT) database was used for training and testing. The database consists of 6300 sentences, recorded from 630 speakers from 8 major dialect regions of the United States. The training and test-set of the database do not overlap. The SA* recordings were excluded as they are pronounced by all speakers and the phonetically identical sentences can bias the results

because co-articulation is same or similar for those recordings (Lee & Hon, 1989; Young, 1992). This resulted in a dataset of 2340 sentences. Also, the 64 phonetic labels provided by TIMIT were folded into 39 phoneme categories, following the approach of Lee & Hon (1989) and Young (1992).

Experiment 4: Stimuli

Pink noise was added to the speech files at signal-to-noise ratios (SNRs) ranging from 30 dB to -15 dB (SNRs of 30 dB, 15 dB, 12 dB, 9 dB, 6 dB, 3 dB, 0 dB, -3 dB, -6 dB, -9, -12, -15). Pink noise was chosen because other types of noise pose an extra challenge on extracting harmonically related segments due to voiced structures in other types of noise. Pink noise serves our goal to determine how well the speech information is retained in the different stages of a CASA process. The control stimuli consisted of the speech files with added pink noise. Experimental stimuli were derived from these control stimuli by extracting segments using the four consecutive steps of the algorithm to extract a harmonic complex. Subsequently, each segmented file was resynthesised to sound. Resynthesis could be performed straightforwardly for the first two steps in the algorithm. The segments derived from step 3 and step 4 in the algorithm were mainly located at the voiced positions of speech, resulting in empty positions at non-voiced positions. Therefore, for these conditions we used the original (noisy) signal at the positions where no harmonic complex was extracted. This resulted in stimuli that were partly resynthesised and partly identical to the control stimuli. This way, we captured possible coarticulation effects consistent in the experimental and control conditions.

Experiment 4: Recogniser

The HMM-based classifier in this study adopts a standard monophone-based single mixture HMM. In total 39 MFCC parameters were extracted; 12 mel-based cepstral coefficients, e-normalised energy, delta and delta-delta features. Because our interest focuses on the difference between conditions and not on the overall performance, we did not make an effort to find the best performing context-independent recogniser. Context was not modelled, as this would make the results less interpretable; the context model could influence classification results unevenly. For all four conditions the classifier was trained on the stimuli obtained from the clean speech training set and tested on the stimuli obtained from the clean or noisy speech test set.

Experiment 4: Evaluation

We calculated recognition results for vowels. The reason for this is that tonal segments and harmonic complexes can characterise complete vowels, but describe voiced consonants (plosives such as "b" and "d", voiced fricatives such as "v" and "z" and nasals) only partly. We calculated the phoneme correct rate (PCR) for vowels in percentages (Following McCowan et al., 2004)) given by $PCR = H / Nr$, where H is the number of correctly recognised phonemes and Nr is the number of reference transcriptions. Because

insertions from the non-target phonemes could confound the results, we did not take into account insertion errors. We used per-sentence bootstrapping to compute 95% confidence intervals (Bisani and Ney, 2004).

Experiment 4: Results

Segmented-variant samples. The effectiveness of segments for application in ASR was determined for clean as well as noisy speech with segments before (step 1) and after (step 2) filtering. The left panel of Figure 2.14 shows the PCR for the vowels for both filtered and unfiltered segmented-variant samples. The figure shows a flat slope for the two experimental conditions in decreasing SNR in contrast to a steep downward slope for the control condition. In the clean speech condition, the PCR is higher (55%) for the control condition than for both experimental conditions (45%). In noisy speech conditions the experimental conditions outperform the control condition.

Differences in PCR between the two experimental conditions show the effect of selecting the more energetic segments. The figure shows that in clean speech and mild noise levels the filtering of the more energetic elements leads to improved performance. However, in more severe noise ($\text{SNR} < 0\text{dB}$) performance deteriorates for the filtered segments relative to the unfiltered segments.

Segmented-and-grouped variant samples. The effectiveness for application in ASR of harmonic complexes was determined for clean and noisy speech conditions with harmonic complexes before (step 3) and after (step 4) supplementing secondary harmonics. The right panel of Figure 2.14 shows the recognition results for the vowels for both conditions of the segmented-and-grouped-variant samples. A flat slope is obtained for the two experimental conditions in decreasing SNR as a contrast to the steep downward slope for the control condition and similar to the results in the segmented-variant samples. In the clean speech condition, the PCR is higher (55%) for the control condition than for the two experimental conditions (35% and 44% for step 3 and step 4 respectively). In all other noise conditions the two experimental conditions outperform the control condition.

Comparison of the two experimental conditions shows the effect of supplementing secondary harmonics to the extracted harmonic complex. In clean speech and mild noise levels recognition scores were improved when secondary harmonics were added. In more severe noise conditions ($\text{SNR} < -12\text{dB}$) no effect was found of adding secondary harmonics.

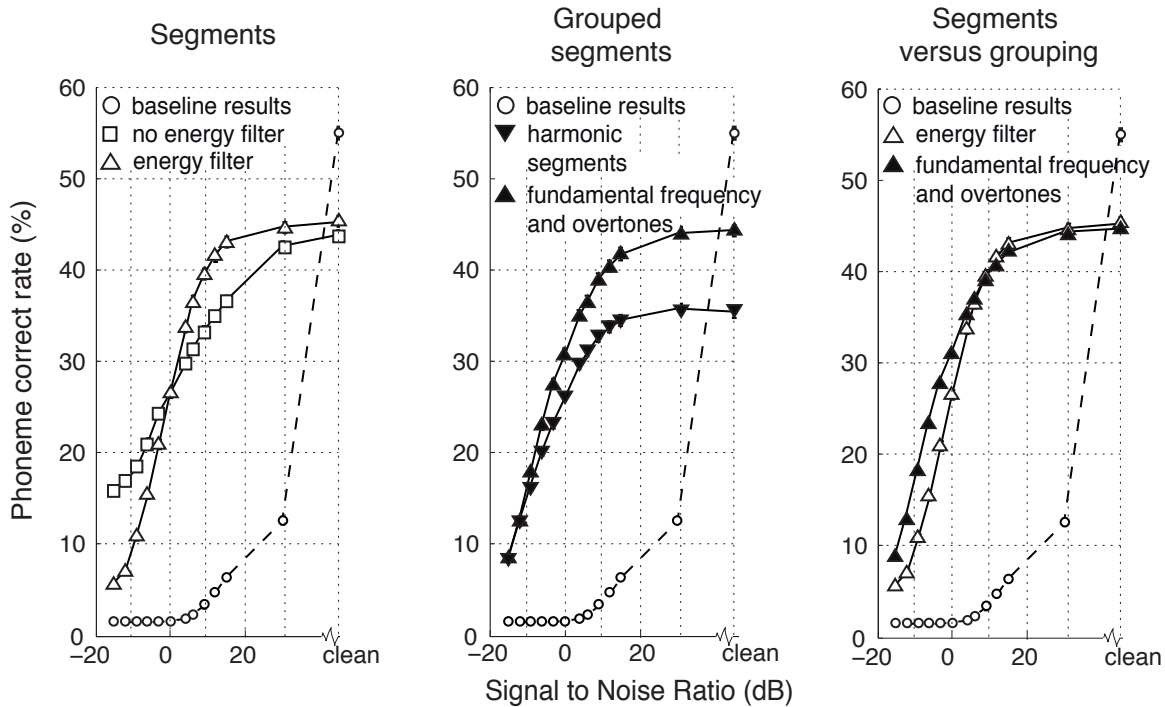


Figure 2.14: The figure shows Phoneme Correct Rate (PCR) in percentages for the vowels as obtained from different noise levels and different types of stimuli. Baseline results are obtained from original speech files in clean conditions and with added noise ranging from 30dB to -15dB and given by a dotted line and open circles in all three panels. Confidence intervals were calculated for all conditions by per-sentence bootstrapping, but too small to stick out from the symbols.

The left panel illustrates the recognition scores for segments without grouping, recognition scores are given for resynthesised segments before (open squares) and after (open triangles) applying an energy filter. The middle panel illustrates the recognition scores after grouping with results before (downside filled triangles) and after (upside and filled triangles) adding harmonics that are congruent with the calculated fundamental frequency. The right panel compares the results for the filtered segments and the completed harmonic complexes.

Experiment 4: Discussion

The objective of this study was to investigate how well the information that is relevant for speech recognition was retained in the segmentation stage and grouping stage of a CASA approach to speech processing. It was argued that methods based on CASA improve the quality of input speech and can function as an alternative to noise subtraction techniques. However, three aspects of the CASA process were identified as potential negative influence on the representation: (a) the effect of energy not caught into segments, (b) potential grouping errors and (c) masking of the target energy as a result of noise. Four subsequent steps in a CASA approach were evaluated on their effectiveness as representation of speech, in both clean and noisy (pink noise) speech conditions.

The experimental set-up consisted of a state-of-the-art ASR system with resynthesised extractions from a CASA process as sound input to the system. In modern ASR-technology, the sound input is transformed into Mel Frequency Cepstral Components

(MFCCs) prior to training. Recognition is based on pattern matching of the MFCC features that represent the whole spectral shape. We did not incorporate a context model in the HMM system and training was performed in clean speech. Therefore, if the spectral shape is changed in one of the experimental conditions, this leads to a decrease in recognition scores.

Performance in clean speech conditions. The results obtained from clean speech conditions (SNR 30dB) show the effects of segmentation and grouping. Potential effects of the CASA processes in clean speech are (a) energy not caught into segments and (b) potential grouping errors. The performance level obtained from clean speech with a standard HMM recogniser without context models, is set as the baseline or control condition and determined at 55% PCR. This baseline level is relatively low, which can be explained by the fact that HMM systems for speech recognition rely heavily on context models (Huang et al., 1991) which we did not incorporate in the HMM recogniser.

In clean speech conditions, the performance levels obtained for the experimental conditions (PCR 45%, 45%, 33% and 44% respectively for step 1 to 4 in the algorithm) are below the baseline level. This loss in performance, as opposed to the control condition, is first observed in the segment selection stage conducted in step 1 and does not deteriorate further with energy filtering performed in step 2. From this we conclude that the segments, regardless of their energy level, do not optimally capture the information needed for the current task, where the spectral shape must be captured optimally to provide useful input for the speech recogniser. This can be accounted for as an effect of (a) energy not caught into segments. Subsequent grouping based on the harmonic relations between segments, performed in step 3, leads to additional loss of information. A possible explanation for the drop in performance level is that the spectral shape becomes sparsely represented because the grouping process omits segments that are not harmonically related to other segments, being a further effect of energy not caught into segments. For the current task a sparse representation of the spectral shape is suboptimal because it can alter the spectral shape easily. An alternative explanation is (b) the possibility of grouping errors. However, we consider this explanation less likely for the current, clean speech conditions, because grouping errors are expected to influence the results of both grouping conditions, while the performance level is recovered with the addition of secondary harmonics, as conducted in step 4. Therefore we conclude that the current CASA process has the negative effect of not capturing all energy into segments. This results in a direct (sparsening) effect on the representation of the spectral shape, leading to a decrease in performance levels for state of the art speech recognisers that rely on a representation of the spectral shape.

Performance in noisy speech conditions. The results obtained from noisy speech show the effects of segmentation and grouping on the representation of speech. Segmentation is expected to restore the spectral shape, segments are considered the noise robust part of the speech stream. However, target energy can be masked as a result of noise. The effect of noise is different for different SNR levels. In order to discuss the results straightforwardly, we divided the results into three clusters (mild, intermediate and heavy noise).

Mild noise: $signal\text{-}to\text{-}noise\text{-}ratio > 10dB$. The addition of noise (SNR 15dB) leads to a drop in performance level in the control (or baseline) condition from 55% PCR to 10% PCR. Because training was performed in clean speech conditions with no context model incorporated, the effect can be attributed to a changed spectral shape as a result of added noise. As a contrast, for all four CASA steps a relatively stable performance level is found for signal to noise ratios of 30dB to approximately 10dB. The results suggest that in mild noise conditions the CASA process recovers the original segment-based spectral shape such that the level of recognition obtained in clean speech conditions can be retained.

Intermediate noise: $10dB < Signal\text{-}to\text{-}noise\text{-}ratio > -3dB$. In further deteriorating conditions (SNR $< 10dB$) the performance levels also decrease for the CASA derived representations, but they remain substantially higher for all experimental conditions than for the control condition. These results suggest that the target energy in the original segment-based representations is affected by noise for SNR $< 10dB$.

Filtering of the segments in step 2 leads to an improvement of the spectral representation, and hence a higher performance level, as opposed to the unfiltered segments in step 1. This can be explained by the removal of noise induced non-target segments. In intermediate noise levels the non-target segments exhibit low energy levels and are therefore removed by filtering in step 2, which leads to an improved representation. Subsequent segment-selection based on harmonic relations in step 3 leads to a degraded representation in intermediate noise levels as compared to both unfiltered segments derived in step 1 and filtered segments derived in step 2. Noise might have lead to detuning of some of the extracted segments resulting in exclusion as tonal component in a harmonic complex. This explanation is especially likely because addition of secondary harmonics in step 4 improves the representation leading to performance levels that are similar to the performance levels obtained for the filtered segments derived in step 2. From this we deduce that, in intermediate noise levels, the differences in performance levels are not due to changes in the energy of the segments but related to exclusion or inclusion of segments.

It was found that the representation obtained in the segmentation stage is retained in the grouping stage. In steady noise conditions, such as investigated here, the advantage of grouping is not explicit. However, in conditions where the noise incorporates tonal components, such as babble noise, the grouping stage can eliminate tonal components that are not harmonically related to the target speech and hence grouping can lead to improved recognition results.

Heavy noise: $signal\text{-}to\text{-}noise\text{-}ratio < -3dB$. In heavy noise the performance level for the unfiltered segments (step 1) remains higher (PCR in heaviest noise level 16%) than the performance level in the control conditions after addition of only mild noise (SNR of 30dB, 10% PCR). Heavy noise is likely to mask target energy. However, the tonal components, characteristic of voiced speech and crucial for the recognition of vowels, exhibit high energy levels. Therefore, they are masked latest of all tonal components in speech which presumably explains the advantage of unfiltered segments over

control speech samples, even in heavy noise conditions. Filtering in step 2, selection of harmonically related segments in step 3 and addition of the secondary harmonics in step 4 all lead to a deteriorated representation as opposed to the unfiltered segments in step 1. Supposedly, the filtering, either based on relative energy or on harmonicity, excludes the few segments that remain accessible in step 1. The slightly better results obtained with the unfiltered segments in step 1 indicate that some information for the identification of voiced phonemes might be left. However, the currently and generally applied ASR approach cannot satisfyingly process the available information. These approaches for ASR work fine as long as the spectral shape is intact.

The results show that in low to intermediate noise conditions both the "energy filtered segments" and the "completed harmonic complexes" give highest recognition scores whereas in heavy noise conditions the "non-filtered segments" seem to capture the spectral shape best. The combination of these results suggest that recognition scores can be optimised by determining an optimal cut-off point when one representation outperforms another. However, this would not be straightforward as it ignores the fact that we applied the algorithm in pink noise while other noise conditions presumably lead to different cut-off points. In competing speaker conditions the segments will represent speech component from both the target and the distractor speech and as such they might not lead to the best results in noise. In conditions with high level competing speaker noises, the algorithm for grouping can become vital as a method to distinguish between target and distractor. The current results show that grouping and completion after grouping does not affect the results of segmented speech recognition negatively illustrating the potential advantage of grouping when groups are captured correctly.

Experiment 4: Conclusions

In summary, we have demonstrated that in clean speech conditions segment selection leads to a significant reduction of the representation of the spectral shape as compared to unsegmented speech. In all noise conditions however, the different steps of CASA processing all recover the spectral shape such that the recognition scores are significantly higher than recognition scores for unsegmented speech. Also, we showed that energy-filtered segments and grouped segments lead to similar recognition scores which suggests that in more demanding conditions, where the noise leads to additional tonal segments, correct grouping is expected to capture the energy of the target speech.

2.4 Robust representations in signal driven speech processing

In Section 2.1 we argued that harmonics in a Harmonic Complex (HC), being local structures that are relatively energetic, serve noise robustness and segmentation of the input stream. To investigate the effectiveness of such representations we developed a method to automatically extract speech representations from HCs. We investigated how the extractions of a harmonic grouping algorithm serve the robustness of formants as dynamic local representations (Section 2.2) and spectral shape as global representation (Section 2.3). Local, dynamic representations are less easily disturbed by noise than global features because noise at a certain frequency does not disturb local representations at a different frequency level while it disturbs representations of the whole spectral shape. Also, local features preserve temporal information and capture information to provide phonetic descriptions of the speech sounds. Our studies showed that both the local features and the global features profit from the robustness of the extracted HC. However, a back-end system needs high flexibility to take optimal profit from the local features because the number and type of extractions are not fixed. Therefore, to understand processing of local features better, we investigate local representations by focussing on human speech processing.

2.4.1 Conclusions

- Representations based on both local and global features profit from the robustness of tonal components in an HC. This rationale to focus on local structures in the TF-representation is not yet applied to other non-tone like phonemes such as plosives and fricatives.
- Local features have the advantage that some segmentation information remains accessible. For AKS this would provide the advantage that potential words can be processed irrespective of the recognition of a sentence.
- Local features can still provide information when only part of them are extracted. Partial extraction can be due to noise or, as discussed by Wester (2003) may be part of the signal when it is due to coarticulation effects.

Chapter 3

Human vowel processing: Knowledge-driven & signal-guided processing

3.1 Local features in models for human speech processing

The high flexibility of Human Speech Recognition (HSR) is attributed to a highly suitable low-level representation of speech (i.e. features) by human listeners (Lippmann, 1997) as a contrast to problems in modern approaches for Automatic Speech Recognition (ASR) that are associated with a poor representation of the speech elements (Li & Allen, 2011; Dusan & Rabiner, 2005).

In Chapter 2 we investigated robust speech representations and concluded that local representations based on the harmonic complex (energetic components; ECs) serve both segmentation and robustness to noise in speech processing but are not suited for most current ASR-systems, generally statistical methods based on linear algebra. Especially, the fact that the number of local features varies, for example as a result of changing acoustical conditions, does not fit to the demands of the current speech recognition systems. We assume that human listeners apply local representations, such as ECs, for speech processing (Section 2.2). To improve our understanding of the processing of local, dynamic representations we investigate human speech processing.

We showed for ECs that noise leads to missing features and superfluous features (illustrated in Figure 3.1); additional extractions that are not related to the target-speech. Missing and superfluous features lead to ambiguity of the feature input. However, most experiments and models on Human Speech Recognition (HSR) focus on ambiguity at a linguistic speech processing level, instead of ambiguity at the feature input level that we suggest to be an effect of noise.

Models for HSR explain experimental findings such as sentence context effects on tasks with lexically ambiguous input (Field, 2003; Simpson, 1984) or lexical effects on phonemically ambiguous input (Cutler, Mehler, Norris & Seguí, 1987; Rubin, Turvey & van Gelder, 1976). These and other findings have been important in modelling the effects of lexical context in HSR, generally performed at the level of words (Cutler & Norris, 1979; McClelland & Elman, 1986; Norris, McQueen & Cutler, 2000). Lexical and phonemic ambiguity is explained by both context-dependent (interactive) models and context-independent models. In the interactive model described by (McClelland & Elman, 1986) context influences bottom-up processing. In contrast, in the context-independent model described by (Cutler & Norris, 1979) frequency and recency of encountering words influence perception and context does solve ambiguity after bottom-up

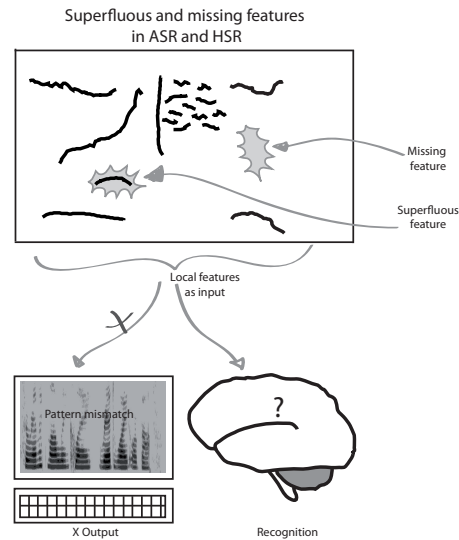


Figure 3.1: We assume that humans apply local representation of speech sounds. Local descriptions have the side effect that features can be missing or superfluous. Automatic approaches for speech recognition can not deal with such input. Therefore we investigate human processing of speech input when features are missing or superfluous.

processing, but context does not influence bottom-up processes. This debate on early or late integration of knowledge on word and phoneme perception is still ongoing. The focus on linguistic phenomena does not succeed to distinguish between bottom-up and top-down HSR models (for a more thorough explanation see Norris et al., 2000).

However, most models on HSR do not focus on ambiguity at the feature input level; HSR models are generally designed to process all input features. An additional research paradigm where the effects of missing features and superfluous features on speech processing is modelled, may lead to new perspectives on speech processing and models for HSR. To improve our understanding of the effects of missing and superfluous target-segments, we investigated human vowel processing in such conditions. Vowels were presented in isolation such that the results were not confounded by knowledge of a higher level than phoneme-level. In one experiment the auditory signal is degraded such that features are missing. In another experiment we induced additional features by adding visual information incongruently to auditory input. Although there is not necessarily a one-to-one relation between audio-visual integration and auditory confusions we chose to use the audio-visual paradigm because it provided us with a method to obtain structured confusions of perceptions when features were added.

3.2 Experiment 5: Audiovisual vowel perception

A modified version of this chapter was previously published as:
Valkenier, Duyne, Andringa & Baskent (2012). "Audiovisual perception of congruent and incongruent Dutch front vowels". *Journal of speech, language, and hearing research* 55(6):1788-801

Experiment 5: Introduction

Perception of spoken language is not an auditory phenomenon only; it is also heavily influenced by visually perceived pronunciation information. The influence of visual cues on speech perception has been shown for a variety of speech tokens, such as consonants (see Massaro, 1987, for an overview) (Massaro, 1989; Massaro & Cohen, 1990) and vowels (Robert-Ribes, Schwartz, Lallouache & Escudier, 1998; Traunmüller & Öhrström, 2007), and conditions, such as hearing-impairment (Başkent & Bazo, 2011; Grant, Walden & Seitz, 1998; Miller & D'Esposito, 2005). This interaction is so strong that, when the auditory and visual components are incongruent, they may fuse into a single percept different than both the original auditory and the original visual stimuli, also known as the McGurk effect (McGurk & MacDonald, 1976). For spoken man-machine interaction devices and video applications such knowledge of audiovisual integration is crucial. For example, the precision with which the auditory and visual information are aligned in video-conferencing tools follows directly from research on audiovisual integration of temporally mismatching stimuli (McGrath & Summerfield, 1985; Miller & D'Esposito, 2005). Also, appropriate audiovisual alignment is especially important for users of rehabilitative communication devices, such as cochlear implants and hearing aids. Since the auditory signals are less well transmitted, hearing impaired listeners rely heavily on the visual cues (Champoux, Lepore, Gagnéú & Théoret, 2009; Rouger, Fraysse, Deguine & Barone, 2008; Başkent & Bazo, 2011). When auditory information is correctly aligned with visual information, listeners, especially hearing-impaired listeners, profit significantly from the visual information for understanding speech (Başkent & Bazo, 2011). However, when audiovisual information is not correctly aligned, disruptive interactions may be observed in addition to the loss of positive interaction. Disruptive interactions of audiovisual information have been shown with the McGurk effect for consonants but are not as extensively investigated for the case of vowels. However, it was recently shown that the contribution of vowels to the auditory intelligibility of speech is significant, and could even be more than the contribution of consonants in some listening situations (Cole et al., 1996; Kewley-Port, Burkle & Lee, 2007; Kewley-Port et al., 2007) Argued that hearing-impaired listeners are even more dependent on the correct perception of vowels because in most cases of hearing impairment high frequencies (associated with consonants) are lost more readily than low frequencies (associated with vowels). Thus, correct alignment is shown to be important for audiovisual interaction devices and although vowels are shown to be important for speech intelligibility, research has focused on audiovisual incongruence with consonants. As vowels are of higher intensity and have longer duration than consonants, the effect of visually incongruent information, as for example in cochlear-implant or hearing-aid users, might be different for vowels than for

consonants. In the present study, therefore, we investigated the perceptual processes that play a role in the audiovisual perception of vowels, more specifically the Dutch high and mid-high front vowels ([i, y, e, ɤ], as in the Dutch words "biet", "fuut", "beet", "hut" respectively), with congruent and incongruent audiovisual features.

Based on acoustic information, the first and second formants of a particular vowel are most crucial for its recognition (for an overview see Rosner & Pickering, 1994). Regarding the vowels of interest of the present study, the first formant (F1) is generally associated with the height feature and the second formant (F2) with the backness feature (Ladefoged, 1982; Rosner & Pickering, 1994). Furthermore, the literature suggests that F2 is also related to the lip-rounding feature for some vowels (Lisker & Rossi, 1992). Masking one of the formants by noise leads to perceptual confusions. By establishing confusion matrices for different levels of white noise, Pickett (1957) observed relatively structured confusions. These relatively systematic perceptual changes can be explained by the fact that different vowels have shared or similar formants. In short, height, by virtue of the perception of F1, is the most robust acoustic feature, followed by backness (F2).

In addition to the acoustic cues, visual cues also influence the perception of high front vowels. Robert-Ribes et al. (1998) have quantified the facilitatory influence of visual cues on the French high and mid-high front vowels [i, y] and [e, ø] by using congruent audiovisual stimuli presented with white noise at different levels. In most cases, the visual and auditory cues are complementary (Massaro & Stork, 1995); for instance, lip-rounding is a strong visual cue, whereas height is a strong auditory cue. Similarly, Miller & Nicely (1955) showed that most features of consonants that were easy to identify from a talker's face were hard to identify from hearing them and vice versa. Summerfield (1987) labeled and described those findings as complementarity in audiovisual processing. Complementarity of the two modalities improves the perception of congruent audiovisual stimuli, especially when the auditory input is deteriorated (such as in background noise). However, if the audiovisual stimuli are incongruent, fusions may occur, such as in the McGurk effect (McGurk & MacDonald, 1976). In short, when a visual [ga] stimulus was concurrently presented with an auditory [ba] stimulus, the resulting perception was that of /da/. The McGurk effect is extensively investigated on different pairs of consonants. However, research has not yet established the limits and the magnitude of the fusion effect in the acoustically more stable vowels. One reason for this could be that such investigation is relatively difficult to do in English where the visually most distinctive feature, lip-rounding, is not an independent distinctive feature of vowels. In other languages with an independent lip-rounding feature, however, an experiment can be conceived that uses vowels that share all perceptual features but rounding. In Swedish, for example, Traunmüller & Öhrström (2007) have found a shift in the auditory response from the Swedish high unrounded front vowel /e/ to the high rounded front vowel /ø/, when an auditory [e] stimulus was shown concurrently with a visual [y] stimulus. This effect was, however, not generalizable, as it was only observed with a subgroup of participants who were more prone towards using visual speech cues.

The aim of the present study was to establish the extent to which the acoustic and visual domains influence audiovisual vowel perception, both in quiet and in background noise. In addition to congruent audiovisual vowel perception, taking advantage of the

lip-rounding feature of Dutch vowels, the perceptual fusion was investigated using incongruent audiovisual stimuli. If the visual and acoustic features are complementary, as argued by Robert-Ribes et al. (1998), the visually more salient (i.e. prominent) feature (lip-rounding) leads to a stronger McGurk effect than the visually less salient one (height). For this purpose, we measured confusions (similar to Traunmüller & Öhrström, 2007; Robert-Ribes et al., 1998) with the Dutch high and mid-high front vowels of [i, y, e, ɤ]. These vowels allowed vowel pairs that would only differ in height or lip-rounding. Hence, in the incongruent stimuli, conditions of audio and video input that differed in height only, rounding only, or both, could be tested. Traunmüller & Öhrström (2007) analysed the data for the subset of participants that were more prone towards using visual cues. In contrast to this we included all participants without a pre-selection. A visual bias, i.e. an increased reliance on visual information in audiovisual perception, was induced for all participants by systematically adding noise to the auditory channel. The advantage by doing so is that the results can now be generalised to not only the sub-group of perceivers that are more prone towards visual cues, but to the entire group of normal-hearing listeners and their audiovisual speech perception in sub-optimal listening conditions.

Experiment 5: Method

Subjects

Sixteen native speakers of Standard Dutch participated in the experiment. The data of one of the participants were not reliable because some data points were missing; therefore all data of this participant were excluded from analysis. The data of 15 participants (11 men, mean age: 24.8 years, SD: 1.9; 4 women, mean age: 23.8 years, SD: 0.5) was analysed. All participants reported normal hearing and normal-to-corrected vision. Participation was voluntary, with the possibility of withdrawing at any time during the study. Participants were fully informed about the study and their written consent was obtained prior to data collection.

Stimuli

- **Selection of speech material and speech context**

In order to give an impression of the Dutch vowel system, Figure 3.2 shows the two-dimensional representation by the first and second formant of vowels (vowel diagram) of the Dutch vowels. The vowel diagram was created with the formants as determined with PRAAT Boersma (2001) from two pronunciations of each vowel, produced in isolation by a 31-year-old female speaker of Standard Dutch. The figure is meant to provide insight in the Dutch vowel system. Not all known characteristics are given in this diagram, formant movement and third formant value are not given. In the present study, we investigated the audiovisual perception of the Dutch high and mid-high front vowels [i, y, e, ɤ]. These vowels were selected because lip-rounding and height features of these vowels cross in the acoustic, as well as the visual domain and there are no other confounding features (For a more extensive analysis and justification of the selected vowels see Valkenier et al. (2012)).

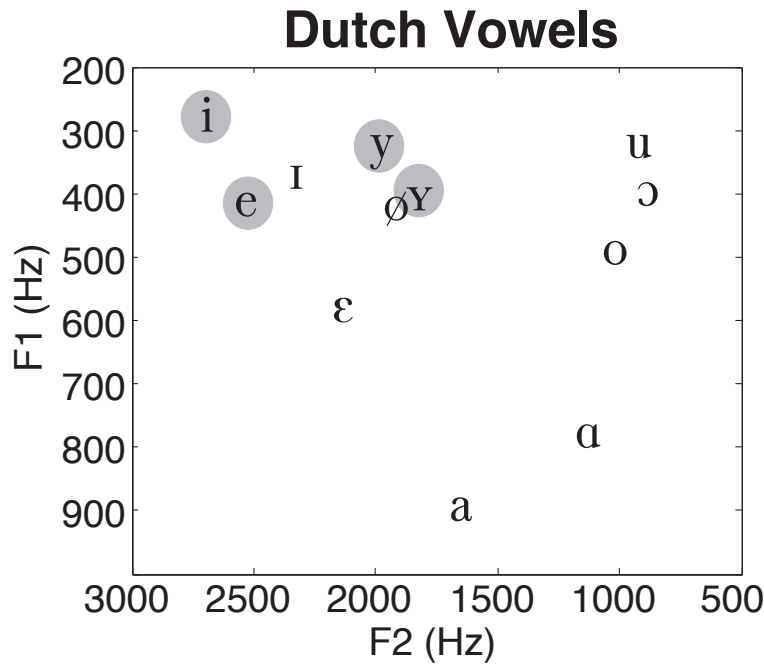


Figure 3.2: Vowel diagram of the mean of the first and second formant of Dutch vowels produced twice in isolation by one female speaker. Shaded vowels are the vowels that were used in the current experiment.

The vowels [i, y, e, ɣ] were recorded in the context of $[\chi V \chi]$, where $[\chi]$ represents a voiceless velar fricative (such as in the Dutch word "acht"). This choice was based on the argument by Traunmüller & Öhrström (2007) that velar consonants hardly affect the visibility of vowel features since the lips and the jaw do not need to be in a particular position. The Dutch language does not have a voiced velar plosive [g] as was used in the Traunmüller & Öhrström (2007) study and the voiceless velar plosive [k] lead to (semantically) meaningful Dutch words. As the context of $[\chi V \chi]$ produces phonologically allowed nonsense words for all Dutch vowels while using a velar consonant, this seemed to be the most appropriate context structure.

- **Recording and editing of speech material**

The stimuli were recorded in a quiet room with bright natural daylight against a white background. The speaker was a 22-year-old female native speaker of the standard variety of Dutch. The stimuli were recorded with a Samsung HMX-H106-SP video recorder placed approximately 3 m from the speaker standing against the white background with audio sampled at a sampling rate of 48000 Hz. Recordings were made from the front of the speaker's face, including the entire face and neck and with the mouth at $1/3^{\text{rd}}$ from the bottom of the screen. The total frame size on the computer monitor was 513 cm^2 and the size of the mouth was approximately 3 cm^2 . The front portion of the tongue was visible for the high vowels (see Figure 3.3).





	Lip-rounding	Height	Articulation example
/i/	unrounded	high	
/e/	unrounded	mid-high	
/y/	rounded	high	
/Y/	rounded	mid-high	

Figure 3.3: Summary of the features of the Dutch front vowels used as stimuli with articulation example taken from the experimental stimuli (the section from around the mouth was cut from a still frame from the corresponding stimuli).

For each vowel two utterances were selected where the head movement was minimal and the experimenters agreed on successful pronunciation of the target-vowel. The duration of the video-files of the selected stimuli were cut to equal duration of one second with approximately 0.3 seconds neutral face at the start and end of the video. The long-term root-mean-square (RMS) levels of the audio-recordings were normalised with PRAAT Boersma (2001). Stationary low-pass filtered noise (SLN) was produced by low-pass filtering white noise (filter order: 1, resulting in a slope of -6 dB/octave in filter response). SLN was added to the stimuli at signal-to-noise ratios (SNRs; calculated on RMS levels) of 30 dB (almost quiet), 0 dB, -6 dB, -12 dB, and -18 dB. Audio presentation level of processed stimuli was calibrated to a comfortable level of approximately 70 dBA.

Those prepared recordings were used as control conditions and served as a starting point for the creation of the stimuli of the experimental conditions. In the experimental conditions, the audio tracks with added noise were recombined with differing video tracks to create incongruent audiovisual stimuli in three conditions; fully crossed, incongruent lip-rounding and incongruent height. In the "fully crossed" condition vowel pairs differed in both height and rounding. In the "incongruent lip-rounding" and "incongruent height" conditions, vowel pairs differed in rounding only or height only, respectively (see the specific vowel pairs used on top

portions of the confusion matrices in Figure 3.9). This resulted in 328 stimuli of one second each (8 video vowel tokens + 8 audio vowel tokens * 5 noise levels + 16 incongruent audiovisual vowel stimuli * 5 noise levels * 3 conditions + 8 congruent audiovisual vowel stimuli * 5 noise levels).

- **Experimental procedure**

An identification task was carried out where each participant was tested on the full set of control and experimental stimuli. Stimuli were presented and responses were collected using E-Prime 2.0 software (Psychology Software Tools) via a MacBook (aluminium unibody, spring 2008 edition) running Windows XP SP2 via boot-camp. The participants were seated in a sound attenuated booth at about 70 cm distance facing a 13 inch flat panel led display (resolution 1280 * 800, angular size 32 degrees) and wore Sennheiser HD 600 headphones, directly connected to the MacBook sound-card output.

The actual data collection was preceded by a short introduction with task instruction and symbol explanation (to familiarise the participants with the possible responses and accompanying keys). The participants were informed that auditory, visual and audiovisual stimuli were to be presented. The test instruction was to continuously look at the screen and to indicate by key-press what was perceived.

The test consisted of two blocks of approximately 15 minutes, with a short break in-between. The stimuli were presented with all conditions and all stimuli randomised over both blocks. For each trial the participant could start the presentation of the target stimulus by key-press. A fixation-cross appeared in the middle of the screen for one second, after which the stimulus was presented. In the audio-only condition, the screen was black. After presentation of the stimulus, the response alternatives were shown on the monitor. The possible answers consisted of all rounded and unrounded Dutch high and mid-high front vowels: /y, ʏ, ø, i, ɪ, e/ plus the vowels /u, o, a/. These were indicated on the screen with the grapheme that is normally written in Dutch with a common Dutch word to clarify the intended vowel sound. No limitation was imposed on response time.

- **Methodology of analysis**

Perceptual confusions were measured and confusion matrices were formed to depict patterns of perceptual change. However, in order to determine the significance of perceptual change the data must be quantified differently, which we did by using error rates (ϵ_c) as described below. Error rates (ϵ_c) were calculated for each experimental condition (c) by subtracting the accuracy (acc; the mean correct responses) from the highest possible error score of 1 (multiplied by 100 to obtain percentages), where acc was calculated as

$$acc(c) = \sum_{pp=1}^{N_{pp}} \frac{N_{CORRECT(pp,c)}}{N_{TRIALS(c)}}, \quad (3.1)$$

where $N_{TRIALS}(c)$ was the number of trials for condition c and $N_{CORRECT}(pp,c)$ was the number of correct responses for participant pp in condition c. Either the visual

or the auditory stimulus was used as a truth reference in order to determine $N_{CORRECT}$. As a means to determine the interaction effects in the audiovisually congruent conditions error rates for multisensory responses were predicted (ϵ_p) from the accuracy scores for the "auditory only" and "visual only" conditions as

$$\epsilon_p = 100 * (1 - (acc(A) + acc(V) - acc(A)acc(V))), \quad (3.2)$$

where $acc(A)$ is the accuracy score for the "audio only" condition and $acc(V)$ is the accuracy score for the "visual only" condition, and $acc(A)*acc(V)$ the probability that both are correct. This way we omitted effects of statistical facilitation and only determined the possible effects of multi-sensory interaction.

A second measure, relative transmitted information score (T_{REL}), was used to analyse the availability of speech features in different noise conditions (for an overview and explanation see van Son, 1994). T_{REL} was the ratio between the transmitted information, T , and the maximum rate of transmission, T_{MAX} , in percentages, such as

$$T_{REL} = 100 * \frac{T}{T_{MAX}}, \quad (3.3)$$

where

$$T_{MAX} = H_{STIM} + H_{RESP} \quad (3.4)$$

and

$$T = T_{MAX} - H_{CM}. \quad (3.5)$$

H_{STIM} and H_{RESP} were mean logarithmic products (entropies) for stimulus and response, respectively, and H_{CM} the entropy of the confusion matrix, calculated by

$$H_{CM} = - \sum_{i,j} p(i,j) * \log_2 p(i,j), \quad (3.6)$$

where $p(i,j)$ was the probability of observing response j for stimulus i in a two-dimensional vector or confusion matrix (H_{CM}), and was replaced by either $p(i)$ (H_{STIM}) or $p(j)$ (H_{RESP}) for a one dimensional vector. T_{REL} was calculated per feature; the analysis was performed on matrices representing either rounded and unrounded stimuli and responses, or high and mid-high stimuli and responses. The relative rate of transmission represented the ratio of the responses that can be predicted from the stimuli (Miller & Nicely, 1955).

Experiment 5: Results

Complementarity in congruent audiovisual vowels

Figure 3.4 shows the confusion matrices aggregated over all noise levels for the congruent conditions. Chance level performance equals 11% correct recognition. Note that the visual-only [y] was more likely to be perceived as /y/ than as /ɣ/. All other single-channel stimuli were perceived mostly correct. Error rates (Figure 3.5) and transmitted

information scores (Figure 3.6) were calculated for every noise condition separately. Also, the multi-sensory error rates as predicted from the auditory and visual error rates are presented. Figure 3.5 shows the error rates, for the "audio only" (filled triangles), "video only" (filled squares), "audiovisual congruent" conditions (filled circles) and "audiovisual as predicted" (ϵ_p , open circles) as a function of noise level. Vowel discrimination benefited from combined audiovisual input, which is reflected in slightly lower error rates in the "audiovisual congruent" condition than ϵ_c , the multi-sensory error rates as predicted from the auditory and visual error rates (Friedman's test, $\chi^2 = 2.67$, p one-sided = 0.051). Here and throughout this experiment we report p-values. The comparison of error-scores reduced the sample size and therefore the p-values are considered not too much distorted by large sample size. Post-hoc comparison shows that the difference is significant for the SNR levels -6 dB, -12 dB and -18 dB (pairwise Wilcoxon, p one-sided < 0.05 adjusted for Bonferroni correction).

Visual influence in incongruent audiovisual vowels

Because the responses to incongruent stimuli can be evaluated with respect to the audio as well as the video input, we calculated two error rates for each incongruent condition. The left and right panels of Figure 3.7 show the error rates with regard to the auditory and visual parts of the input, respectively. The error rates for the "audiovisual congruent" condition are the same in both panels because the visual and auditory stimuli were the same in this condition. The figure shows that both the auditory and the visual error rates are higher in the three incongruent conditions (open symbols) than in the congruent condition (filled symbols). In all conditions, the auditory perception deteriorates with increasing noise level, which is reflected by upward slopes. In contrast, the visual perception improves with increasing noise, reflected by a similar, but inverse and less profound, pattern with regard to the visual error rates.

Figure 3.8 shows the results of Friedman's test when the visual error rate of an incongruent condition was compared with the congruent condition or the "visual only" condition, and when the auditory error rate of an incongruent condition was compared with the congruent condition or the "audio only" condition. Figure 3.8 also shows the levels for which the post-hoc Wilcoxon test is significant (after correction for Bonferroni).

For all incongruent conditions, the overall auditory and the overall visual error rates are significantly different from the four reference levels (Friedman, $p < 0.001$, pos-hoc Wilcoxon's test, adjusted $p < 0.05$ for all comparisons except for "visual error rate for incongruent lip-rounding" compared with "visual only" for -6 dB, -12 dB and -18 dB). For all but one of the conditions, the error rates in the experimental condition are significantly higher than the reference levels; namely, the overall auditory error rate in the "incongruent height" (and thus congruent lip-rounding) condition is significantly lower than the "audio only" error rate.

The auditory error rates in the "incongruent lip-rounding" and the "incongruent lip-rounding and height conditions are significantly different from the auditory error rates in both the "audiovisual congruent" and the "audio only" conditions for the 0 dB, -6 dB, -12 dB and -18 dB SNR levels ($p < 0.01$). The auditory error rates in the "incongruent height" condition are significantly higher than the "audiovisual congruent" error rates for the SNR of -18 dB ($p < 0.05$) and significantly lower than the "audio only" error

a) stimulus audio only					b) stimulus video only				
	[i]	[e]	[y]	[Y]		[i]	[e]	[y]	[Y]
response (%)					response (%)				
/i/	79.4	1.3	10		/i/	59.4	12.5	3.2	6.3
/I/	5.6	0.6	3.1	10.6	/I/	25	15.6		
/e/	1.3	77.5	0.6	13.1	/e/	3.1	56.3	3.2	3.1
/y/	10		76.9		/y/	3.1	6.3	67.7	40.6
/ø/		15.6		5	/ø/		3.1	12.9	12.5
/Y/	3.1	4.4	8.1	70	/Y/	6.3		9.7	25
/a/					/a/				
/o/		0.6		0.6	/o/			3.2	
/u/	0.6		1.3	0.6	/u/	3.1	6.3		12.5

c) stimulus audiovisual congruent				
	[i]	[e]	[y]	[Y]
response (%)				
/i/	91.3	0.6	1.3	
/I/	6.9	2.5		0.6
/e/		96.9		
/y/	1.3		90	0.6
/ø/			1.3	11.3
/Y/	0.6		5	87.5
/a/				
/o/				
/u/			2.5	

Figure 3.4: Confusion matrices of the results of the experimental control conditions. Summary of the features of the Dutch front vowels used as stimuli with articulation example taken from the experimental stimuli (the section from around the mouth was cut from a still frame from the corresponding stimuli).

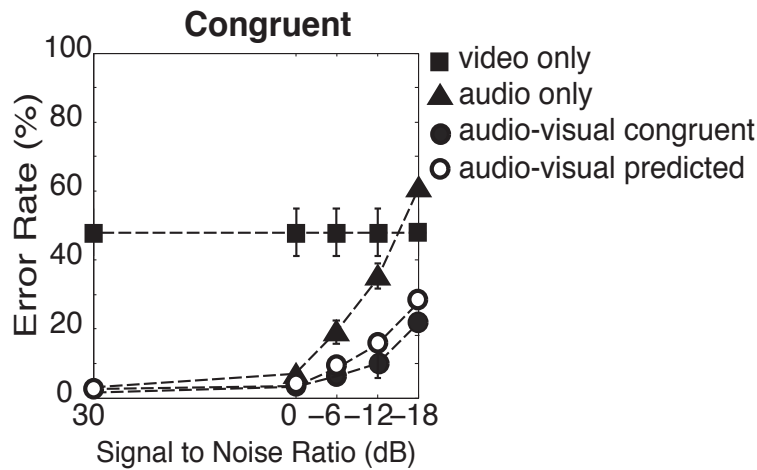


Figure 3.5: Error rates for single channel, audiovisual congruent and audiovisual predicted vowel stimuli. The depicted error rates are averaged across all listeners and shown in percentages as a function of decreasing SNR (i.e. increasing level of the steady low-pass filtered noise, SLN). The error bars show standard errors.

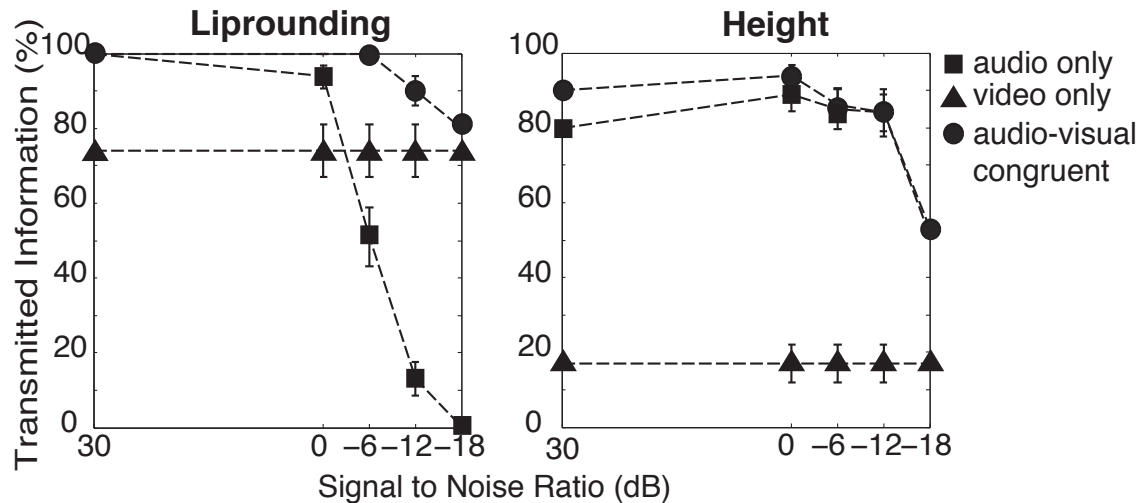


Figure 3.6: Transmitted information in percentages shown as a function of decreasing SNR (i.e. increasing noise level) for the three control conditions. The left panel denotes the transmitted information for lip-rounding and the right panel for height.

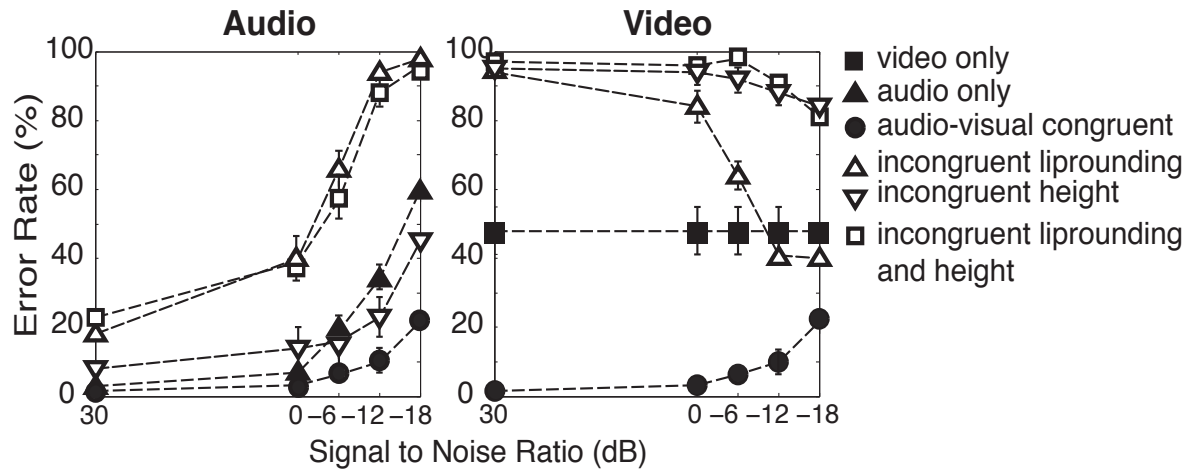


Figure 3.7: Error rates for audiovisually incongruent presented vowel stimuli (open symbols) as well as the reference conditions (audiovisual congruent and single channel; filled symbols). The depicted error rates are averaged across all listeners and shown in percentages as a function of decreasing SNR (i.e. increasing level of the steady low-pass filtered noise, SLN). The left panel shows the error rates with regard to the auditory stimulus and the right panel the error rates with regard to the visual stimulus. The error bars show standard errors.

		Auditory error rates		Visual error rates		
		AV-congruent	audio	AV-congruent	video	
incongruent with regard to...	height	Friedman	$X^2 = 13^{***}$	$X^2 = 13^{***}$	$X^2 = 80^{***}$	$X^2 = 57^{***}$
		significant for SNR levels (dB)	-18*	-18*	all**	all**
	lip-rounding	Friedman	$X^2 = 63^{***}$	$X^2 = 58^{***}$	$X^2 = 70^{***}$	$X^2 = 12^{***}$
		significant for SNR levels (dB)	0, -6, -12, -18**	0, -6, -12, -18**	all**	30, 0*
	lip-rounding and height	Friedman	$X^2 = 65^{***}$	$X^2 = 56^{***}$	$X^2 = 59^{***}$	$X^2 = 80^{***}$
		significant for SNR levels (dB)	0, -6, -12, -18**	0, -6, -12, -18**	all***	all***

Figure 3.8: Significance tests on error rates of incongruent conditions compared to control conditions.

rates for the SNR of -18 dB ($p < 0.05$).

Transmitted information scores

The transmitted information scores provide more detailed insight into the error rates as they show what part of the information was or was not available, when analysed for different features. Figure 3.6 shows the transmitted information scores for lip-rounding (left panel) and height (right panel) for the "audio only", "video only" and "audiovisual congruent" conditions. Highest profit from visual input was in noise; the audiovisually transmitted information for lip-rounding is significantly higher than the auditorily or visually transmitted lip-rounding information for SNRs of -6 dB, -12 dB and -18 dB (Friedman $\chi^2 = 42$ and 23 , respectively, $p < 0.001$; Wilcoxon, adjusted $p < 0.05$). Furthermore the lip-rounding is better transmitted visually than auditorily at SNRs of -12 dB and -18 dB (Friedman $\chi^2 = 7$, $p < 0.01$; Wilcoxon, adjusted $p < 0.001$). The height information is better transmitted auditorily and audiovisually than visually for all SNR levels (Friedman $\chi^2 = 59$ and 66 , respectively, $p < 0.001$; Wilcoxon, adjusted $p < 0.05$).

McGurk effect in incongruent audiovisual vowels

In incongruent conditions, fusions of features were expected to occur, namely features from the visual and auditory input are recombined into a perceived vowel that was not presented in either one of the channels. We originally expected fused percepts that combine the auditorily salient height feature and the visually salient rounding feature. In the present study, apart from these expected fusions also unexpected ones were found, as seen in the confusion matrices aggregated over all noise levels (Figure 3.9) where the bold and underlined numbers represent the expected fusions. All fusions that seem to be a trend in the data are reported in this section and the unexpected findings (that are a trend) are explained in the discussion. Furthermore, the major fusions (the fusions that occur most often per category, predicted or not) are plotted in Figure 3.7 as a function of noise. It will be reported whether the number of fused responses increases (or decreases) significantly with increased noise.

Figure 3.9-a shows the responses in the fully crossed condition. In this condition fusions occurred when the vowels [i, e, y, ʏ] were presented through the auditory channel with the vowels [ʏ, y, e, i] through the visual channel, respectively. While we expected to find the fused responses [y, ʏ, i, e], respectively, the observed fusions lead predominantly to perceived /y, ø, i, ɪ/ instead. The peak of the observed fusions was found at SNR of -18 dB for [i_a] with [Y_v] and at SNR of -12 dB for the other three stimulus-pairs.

Next, Figure 3.9-b shows the responses for the "lip-rounding incongruent" condition. We expected increased visual responses because the auditory and visual height information are combined with the visually salient lip-rounding feature. Next to this expected result we found that auditory [y] presented with visual [e] was sometimes perceived as /ɪ/ (ranging from 9% at 30 dB SNR to 40% at -6 dB SNR). Also, the auditory [e] presented with the visual [y] was sometimes perceived as /ø/ (ranging from 34% at 30 dB SNR to 70% at -12 dB SNR).

Finally, Figure 3.9-c shows the responses for the "incongruent height" condition. We expected increased auditory responses because the auditory and visual rounding infor-

Confusion Matrices of Incongruent Conditions

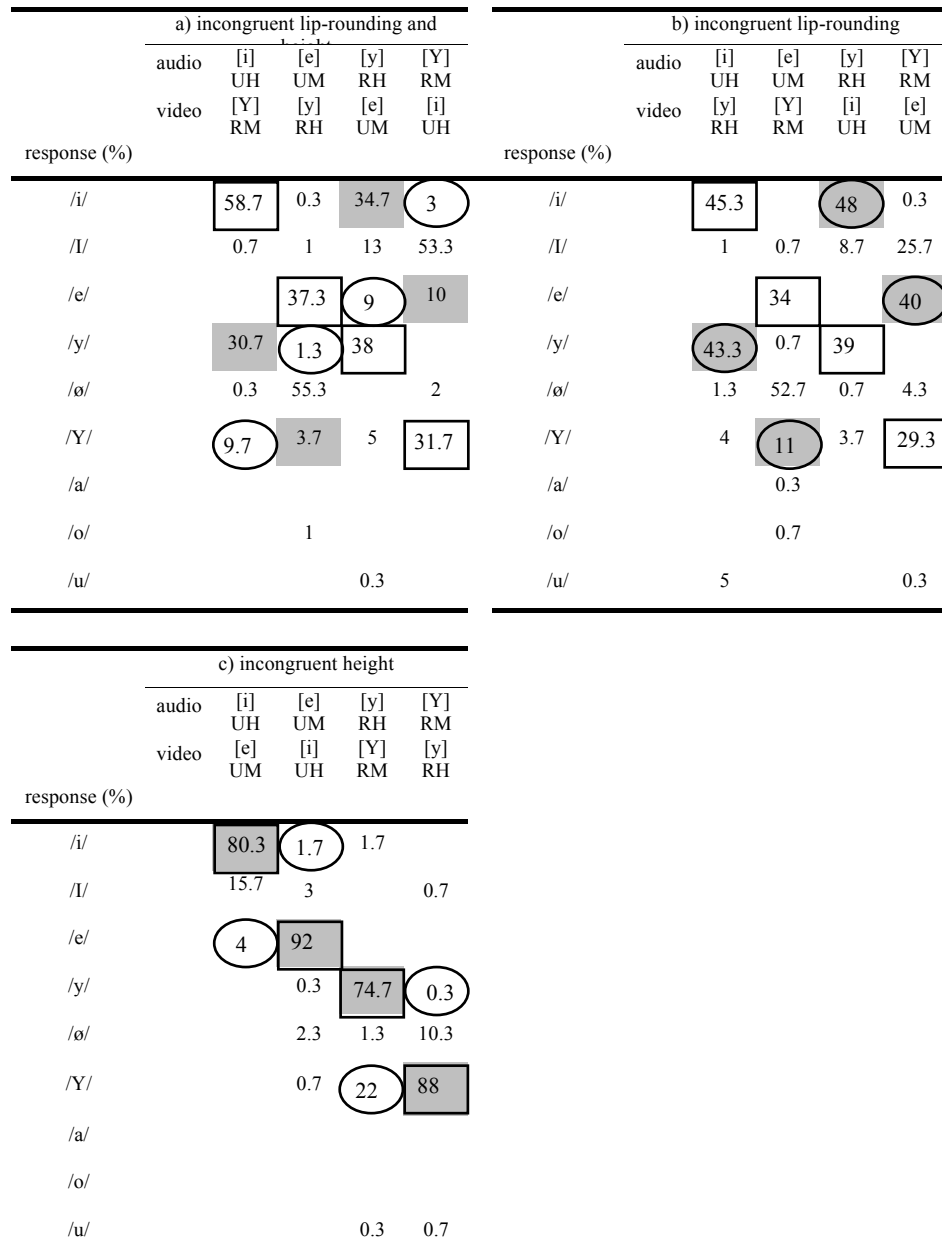


Figure 3.9: Confusion matrices in percentages (%) rounded to one decimal digit. The percentages are calculated from the aggregate of responses to all presentations and noise levels for the incongruent conditions a) "lip-rounding and height" (fully crossed), b) "lip-rounding", and c) "height". The columns and the rows represent the audiovisually presented stimuli and the responses, respectively. A coded description of the vowel features of the presented vowels are given in the top row: U=unrounded, R=rounded, H=high, M=mid-high. Each cell shows the percentage of the aggregated number of times that a response was given at a specific audiovisual stimulus presentation. The outlined cells are the responses that are congruent with either the auditory (rectangle) or the visual (oval) stimulus input. The shaded cells are the expected fusion responses.

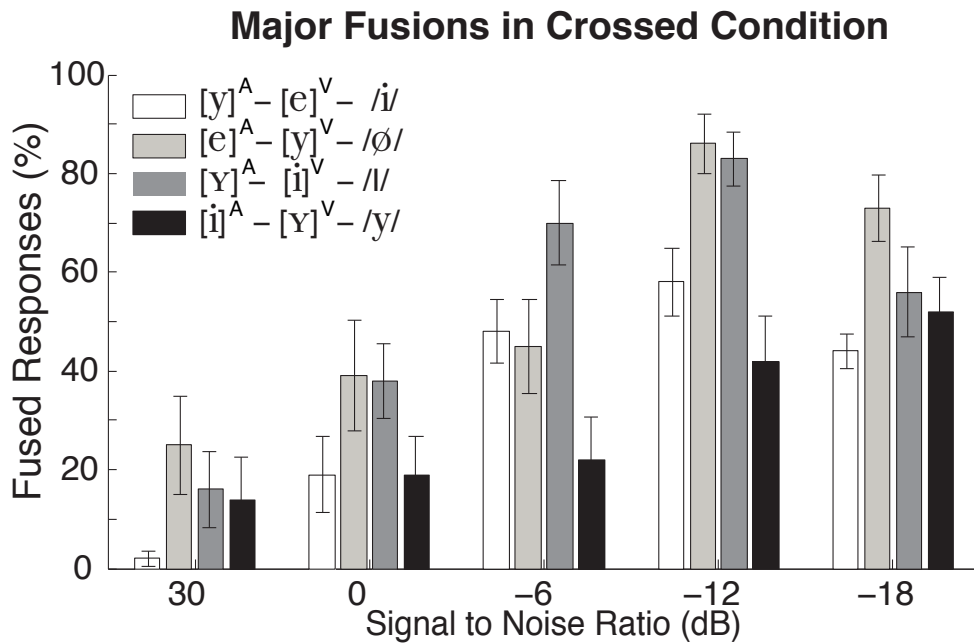


Figure 3.10: Percentage of fused responses for the four different audiovisual vowel pairs in the crossed where lip-rounding and height were both presented incongruently. Fusion targets are the major fusions and not the predicted fusions. The error bars show standard errors.

mation are combined with the auditorily salient height feature. Next to this expected result we found an increase in /i/ responses when [i] was presented auditorily with [e] visually (ranging from 0% at 30 dB SNR to 56% at -18 dB SNR).

The major fusions are shown in Figure 3.10 as a function of noise. A Friedman test with noise-level as factor revealed that the number of fused responses is significantly different for different noise levels ($df = 4$, Friedman $\chi^2 = 202$, $p < 0.001$). A post-hoc Wilcoxon analysis revealed that the number of fusions significantly changed for each increase in noise. The number of fusions increased for SNRs of 0 dB, -6 dB, -12 dB and decreased for SNR of -18 dB (adjusted $p < 0.01$ for all comparisons).

Experiment 5: Discussion

The present study used congruent and incongruent Dutch front vowels as audiovisual stimuli, presented in steady low-pass filtered noise, to investigate to what extent visual cues influence the perception of vowels. The noise, by degrading the auditory input and forcing the participants to rely more on the visual input, served the purpose of producing robust perceptual interactions between the audio and visual cues.

Robert-Ribes et al. (1998) have shown, for the case of vowels, that visual and auditory features are complementary (see also Summerfield (1987)); namely, the feature whose auditory discrimination is hardest can be perceived better through vision and vice versa. When the information is incongruent, the auditory and visual features were expected to interact in a way that can be explained by the ease of perception in either of the two channels. For incongruent stimuli this would yield perceived vowels that combined the most salient auditory cue with the most salient visual cue. This would in turn

lead to fusions of vowel features, similar to the McGurk effect previously observed with consonants.

Complementarity in congruent audiovisual vowels

We reproduced the findings of Robert-Ribes et al. (1998) for Dutch vowels. Our results showed complementarity of the features in the auditory and visual channels; the transmitted information for lip-rounding, for example, was higher in the congruent audiovisual condition than the transmitted information for lip-rounding in the audio-only or video-only condition (see Figure 3.6). Also, for low SNR, the perception of congruently presented audiovisual vowels was better than the score that was predicted based on vowels perceived through either of the single channels (see Figure 3.5).

Visual influence in incongruent audiovisual vowels

The main interest of the present study was the perception of audiovisually incongruent vowels. As vowels are shown to contribute significantly to intelligibility of speech (Kewley-Port et al. (2007) correct perception of vowels can be decisive for speech understanding. Yet, this can be disrupted as a result of misalignment of the auditory and visual signals, for example, in modern audiovisual communication devices. Until now research on audiovisual incongruency has focused on consonants, which needs to now extend to vowels.

In this study we showed that the auditory processing of vowels was influenced by incongruent visual information which was reflected by an increase in auditory error rates in comparison to the "audiovisual congruent" condition for all incongruent conditions (see Figure 3.5). The increased auditory error rate was highest for both conditions when the auditory stimulus was presented with incongruent lip-rounding, but incongruently presented height also lead to a change in the response distributions. Apart from the beneficial influence that congruent visual information has on the perception of speech (Başkent & Bazo (2011)) and more specifically vowels (Robert-Ribes et al. (1998); this study), incongruent visual vowel information is disadvantageous for the correct perception of vowels. Even when the visual input is not very salient (i.e. height), incongruent presentation can disrupt the perceptual process, especially when the auditory signal is less well represented. If processing speed in audiovisual devices can be improved by passing half the auditory information, one can think of special conditions where ignoring the visually salient lip-rounding information in the audio channel of technical devices would improve the alignment through improving the processing speed. This could aid the correct perception of vowels and hence speech, as the information is transmitted through the channel through which it is saliently perceived.

McGurk effect in vowels with incongruent lip-rounding

For the incongruent conditions where both visual and auditory error rates were higher than the "audiovisual congruent" error rates, the perceived vowel was neither the auditorily nor the visually presented one. This was the case in both conditions with incongruent lip-rounding. The confusion matrices for those conditions showed fusions of vowel features (McGurk effect). As was hypothesised, the fusions consisted mainly

of vowels in which the height of the auditory vowel was combined with the rounding of the visual vowel (see Figure 3.9; shaded cells). Exceptions to this were the following: In the "incongruent lip-rounding and height" condition auditory [ɣ] presented with visual [i] was perceived as /ɪ/, and auditory [e] presented with visual [y] was perceived as /ø/. Similarly, the "incongruent lip-rounding" condition showed recombination of auditory [ɣ] presented with visual [e] into /ɪ/ percepts and auditory [e] presented with visual [ɣ] into /ø/ percepts. Although we present them as exceptions, the responses can be interpreted as natural fusions. The vowel [ɣ] was used instead of [ø] because [ɣ] belongs to the same viseme category as [y] (explained in more detail in the Appendix in Valkenier et al. (2012)). Although both [ø] and [ɣ] are called mid-high vowels, their first formant frequencies (F1) are not identical Adank et al. (2004). F1 of [ɣ] is more similar to F1 of [i] than to F1 of [i] or [e]. Also, the first formant frequency of [e] is more similar to F1 of [ø] than to F1 of [ɣ] or [y]. Therefore, the results are not intrinsically different from McGurk-like fusions; an audiovisual vowel is perceived with the rounding of the visually presented vowel and with the F1 closest to the auditorily presented vowel, especially since height is most salient in the auditory channel.

McGurk effect in vowels with incongruent height

Contrary to our expectations, we also found significant visual influence when height was presented incongruently in the auditory and visual channels. Height is not a very visible feature because tongue placement is hidden behind lip articulation. Therefore, we expected results similar to those of the congruent stimuli; i.e., auditorily presented [ɣ] with visually presented [ɣ] would then lead to the auditory height perception of [ɣ] and the visual rounding perception of [ɣ], resulting in a perceived [ɣ]. Indeed, the visual influence of congruent lip-rounding was additive or complementary; auditory identification improved with regard to the "audio only" condition. However, next to this positive influence, we also found a detrimental influence; both auditory and visual identification degraded (i.e. resulted in higher error rates) with regard to the "audiovisual congruent" condition which implies that neither the visual nor the auditory input was effectively perceived.

The confusion matrices show that the detrimental effect in both modalities was due to two effects of non-normal perception / fusions (see Figure 3.9-c). First, auditory [ɣ] presented with visual [ɣ] lead to the perception of either [ɣ] or [ɣ] where we expected a congruency effect leading to predominantly [ɣ] responses. The perception of [ɣ] combined the auditory- and visual perception of lip-rounding with the visual height despite the fact that the visual height was less well transmitted visually than auditorily at all SNR levels (see Figure 3.6). Second, an increase in the number of [i] perceptions was found when [i] was presented auditorily with [e] visually. This was not a purely auditory effect as auditory [i] presented on its own did not often result in [i] percepts (Figure 3.4-a). It must be noted, however, that also in the three control conditions [i] responses were given to both [i] and [e] stimuli. The effect can partly be explained as follows; [ɪ, i, e] belong to the same viseme category of short unrounded vowels (van Son (1994)). Adank et al. (2004) showed that the mean first formant values for those vowels (pronounced by 10 female speakers) are 442 Hz, 399 Hz and 294 Hz for [e, ɪ, i], respectively. Therefore, the perceived [i] combines the audiovisual lip-rounding with a vowel having the height (F1)

in between the height of the presented vowels [e] and [i] despite the fact that height is best transmitted auditorily at all SNR levels. It turns out that the incongruent visual input was sometimes preferred over the more reliably transmitted auditory information (see confusions in Figure 3.9-c).

It can be concluded that in special cases, where perceptual features are crossed, fusions occur in incongruently presented vowels, similar to the McGurk effect commonly observed in consonants. Vowels are longer in duration and higher in energy than consonants and the results show evidence that these intrinsic differences do not prevent the cognitive system from binding information from the different modalities, especially when the auditory signal is less reliable. Further research could reveal audiovisual interactions between vowels and consonants. Audiovisual interactions of long vowels and short consonants could lead to partial incongruence of which the effect is unknown. Also, the audiovisual interaction of people that are hard of hearing might differ from the results found in this study and needs further investigation. Namely, long-standing hearing loss might lead to a different phonological system (for example, a few of the cochlear-implanted participants in the study of Schorr, Fox, van Wassenhove & Knudsen (2005)) gave [ta] responses to the three different stimuli /ka, pa, ta/, indicating that the phonological system is broadened for these participants with regard to these phonemes) which could result in interactions different from the ones found here. Insight in audiovisual interaction in different conditions may help to better understand the problems people experience with misalignment of the auditory and visual channels and where the focus should be with regard to alignment.

The influence of saliency on the McGurk effect

The influence of saliency on the number of fused responses can be related to the transmitted information scores. It was shown that the number of fused responses increases significantly for increasing noise levels up to SNR of -12 dB. The auditory transmitted information scores for height decrease gradually with noise increasing to SNR of -12 dB and hence the reliance on visual information increases; transmitted information for lip-rounding is better through the visual than through the auditory channel for SNR of -6 dB and below. Furthermore it was shown that the number of fused responses significantly decreases for -18 dB SNR with respect to -12 dB SNR. This can similarly be related to the steep drop in transmitted information for height and hence the identifiability of the height feature. Thus, when noise increases, the reliance on visual information increases accordingly, which leads to fused responses as long as the auditory height is perceived correctly.

Experiment 5: Conclusions

In summary, we have demonstrated that the audiovisual information leads to complementarity in congruent vowels. Furthermore, we have shown that incongruent visual input influences the perception of stimuli even if the visual information does not very well distinguish between vowels. Finally, we have shown that this knowledge is not always used optimally, as listeners sometimes used less salient information from one modality even when more salient information was available from the other modality.

The finding that even the visually less salient height feature influences auditory identification stresses the importance of appropriate audiovisual alignment in communication devices, such as cochlear implants and/or video-conferencing tools, especially when the auditory signals are degraded and listeners rely heavily on the visual cues (Champoux et al. (2009); Rouger et al. (2008)). For those types of applications the addition of visual information is of great help, but if not done right, it can also distort the perception of speech.

Experiment 5: Appendix

The high and mid-high front vowels [i, y, e, Y] were selected because lip-rounding and height features of these vowels cross in the acoustic as well as the visual domain with no other confounding features, as explained below in detail:

- With regard to the acoustic features, height and diphthongization were aimed to be matched in pairs of vowels. The Dutch vowels [i, y] are high vowels and [e, I, Y, L] are mid-high vowels (Adank et al., 2004; Pols, Tromp & Plomp, 1973; van Hout, Adank & van Heuven, 2000). van Hout et al. (2000) found that expert listeners judged the vowels [e] and [L] in standard Dutch as relatively monophthongal, although they are conventionally described as diphthongs (Gussenhoven, 1999) or near-diphthongs (Rietveld & van Heuven, 2009). Therefore the vowels [i, y] and [e, I, Y, L] make appropriate candidates for the forming of vowel pairs that are either different from or equal to one another in height.
- With regard to the visual features, the rounded vowels [y] and [Y] belong to the viseme category of short rounded front vowels, whereas [L] belongs to long rounded front vowels (Van Son et al., 1994). The vowels [e, I, i] belong to the viseme category of unrounded front vowels. Therefore, the vowels [y, Y] and [I, i, e] make appropriate candidates for the forming of vowel pairs that are either different from or equal to one another in rounding.
- The crossing of features in the acoustic and visual domains was necessary for analyzing the responses to the incongruent vowel stimuli that is, where a feature can conflict in the auditory and visual domains without other conflicting features. Using crossing of features as a criterion, it was most appropriate to use [e] and [i] as monophthongal and unrounded vowels (mid-high and high, respectively) and [Y] and [y] as monophthongal and rounded vowels (mid-high and high, respectively; see Figure 3.3. As an example, complete crossing can now be achieved by combining the auditory vowel [e] with the visual vowel [y] (crossed on both the rounding and height features, whereas all other features are kept equal).

3.3 Experiment 6: Auditory vowel perception

A modified version of this chapter was previously published as:
Valkenier & Gilbers. "The effect of formant manipulation on the perception of Dutch front vowels".

In Chapter 2 we argued that noise can lead to superfluous and missing features when local speech representations are used. The investigation of speech processing in noise, with noise characterised by missing and superfluous features, can lead to new perspectives on noise robust speech processing and models for HSR. In the current experiment we investigated human vowel processing when target segments were missing. This approach is similar to experiments with spectral restoration that is investigated in the field of auditory scene analysis (Warren, Hainworth, Brubaker, Bashford & Healy, 1997). In spectral restoration experiments, speech is presented with spectral silences that are replaced by noise of different levels. An effect of noise level and noise bandwidth on the degree of perceptual restoration of the input sentence is demonstrated with this type of experiments. In the current experiment we focus on the condition where the silence is not replaced with noise. An in depth analysis of perceptual shifts is performed. Vowels were presented as both complete acoustic signals and as incomplete signals in which the second formant was suppressed. In an open form response task, participants were presented with high and mid Dutch front vowels: particularly, the primary cardinal (unrounded) vowels /i, e/ and the secondary cardinal (rounded) vowels /y, ø/.

Experiment 6: Introduction

The acoustic characteristics of a sound form the basis of speech perception. Traditionally, formants have been considered to be essential cues for the perception of vowel quality (see Rosner & Pickering (1994) for a review). In line with this, Diehl & Lindblom (2004) showed that perceived phoneme identity is affected by the first two or three formant frequencies, whereas other information (such as the spectral shape and the higher formants) mainly influences the judged similarity of segments and not the perceived identity. Nevertheless, Bladon & Lindblom (1981) and Bladon (1982) believe the whole spectrum to be important for the perception of speech sounds. They argue that if only formants are considered, information that may prove to be auditorily relevant, for example spectral zeros (time-frequency regions of reduced energy levels), that are argued to be important for the recognition of nasals, would be neglected. Also, Molis (2005) points out that whole-spectrum representations (incorporating peak location, peak energy and energy in between peaks) always give a richer description of the frequency spectrum than representations that are limited to formant frequency although the latter might be enough for vowel identification. Molis (2005) evaluated a variety of models for vowel representation and concluded that both a whole spectral shape model and a model based on formant frequency and formant energy better predicted perception data than a model based on formant frequency only.

Perceptual studies support the finding that perception data is best explained by

formant frequency and energy or by the whole spectral shape. Evidence is found for both a representation based on the whole spectral shape and a representation based on formant frequency and amplitude. Ito et al. (2001) have shown that if the first or second formant was suppressed (with the rest of the spectrum and the 3rd to 5th formant frequencies and energies equal over all vowel stimuli), synthetically produced /i, e, a, o, u/ were interpreted correctly. However, perceived vowel identity changed when the relative amplitude of the first formant was altered with respect to the formant amplitudes of the 2nd to 5th formants while keeping the formant frequency locations equal. Because the influence of the formant amplitudes could alter the perceived identity of the input vowels, Ito et al. (2001) argue that the spectral shape information can be crucial for the perception of vowel quality. Additional to this, Kiefte & Kluender (2005) showed that although the entire spectral shape indeed influences vowel perception, this effect is reduced when the formants in the vowels are kinematic such as in naturally produced vowels. Aaltonen (1985), Jacewicz (2005) and Kiefte, Enright & Marshall (2010) investigated the effect of formant amplitude on perceived vowel identity. They reported a change in perceived vowel identity when formant amplitude was manipulated in synthetically produced vowels. Aaltonen (1985) showed that a decrease in formant amplitude of the second formant of [y] resulted in increased /i/ responses. Similarly Jacewicz (2005) showed that a decrease in formant amplitude of the second formant of [i] resulted in increased /i/ responses. Kiefte et al. (2010) showed that synthetically-generated [u] is perceived as /i/ when the energy in the second formant was decreased or when it was increased. These results indicate that relative peak energy affects vowel perception.

Most experimental research on vowel perception is performed on synthetically-generated vowels and it is questioned whether the results hold for natural vowels of different speakers (Jacewicz, 2005; Kiefte & Kluender, 2005). In the present study we investigate the effect of formant suppression on the perception of naturally-produced vowels instead of synthetically-produced vowels. If the influence of spectral shape is reduced for the perception of naturally-produced vowels (as suggested by the results of Kiefte & Kluender, 2005) and thus mainly based on formant frequency-location, systematic identification changes are expected when a formant is suppressed. However, if the influence of spectral shape is not reduced and perception is based on the amplitude information of the higher formants when one of the low formants are suppressed, this is expected to lead to unaltered vowel identifications. A second adjustment to the experiment of Ito et al. (2001) is that we used the rounded front vowels /y, ø/ that are part of the Dutch vowel system, additional to the unrounded front vowels /i, e/. Ito et al. (2001) used synthetically-generated unrounded front vowels /i, e/ and rounded back vowels /o, u/ as well as /a/ in their experiments. The unrounded front vowels /i, e/ and rounded back vowels /o, u/ are considered primary cardinal vowels; typologically preferred vowels of language systems (Jones, 1963; Maddieson, 1984). The rounded front vowels /y/ and /ø/ are considered secondary cardinal vowels. As the Dutch front vowels /i, y, e, ø/ are close together in vowel space, rounded and unrounded counterparts can be easily confused. In this way, Dutch differs from languages with a smaller vowel inventory in which the vowels are more dispersed in the vowel space. Therefore, changed formant frequency and changed formant amplitude could have a different effect on the perceptual confusions in

our study than in the one by Ito et al. (2001).

Experiment 6: Methods

Subjects

Thirteen native speakers of Dutch (mean age: 20.6, SD: 1.7) participated in the listening experiment. The participants were male students of the Artificial Intelligence Department (most students of the AI Department were male at that time), who took course credit for participation. All participants reported good hearing and normal to corrected vision. All participants were accustomed to touch typing which we considered important as 16 different keys could be used. The data of one participant who reported dyslexia was removed before analysis.

Stimuli

Recordings were made of the unrounded high and mid-high Dutch vowels /i, e/, the rounded high and mid-high /y, ø/ and five distractor vowels /a, u, ε, ɔ, o/. Two male and two female native speakers of Standard Dutch (colleagues from the auditory cognition group of Artificial Intelligence Department) produced the vowels in varying order (mean pitch: 197 Hz and 119 Hz, SD: 2 Hz and 3 Hz, for female and male speakers respectively, mean duration: 460 msec, SD: 80 msec). In order to avoid interference from intrinsic vowel duration differences, we decided to present all vowels as long vowels.

According to Adank et al. (2004), lax vowels in Dutch have a mean duration of 107 msec. and tense vowels have a mean duration of 191 msec. As the vowels /e/ and /ø/ are tense and long vowels (mean duration 177 msec. and 184 msec. respectively Adank et al. (2004)) and the vowels /i/ and /y/ are tense, but short vowels (duration 93 msec. and 95 msec. respectively Adank et al. (2004)) duration could give a cue as to what vowel was perceived.

The speakers were instructed to produce the sounds as elongated vowels with deliberate concern to reach a constant pitch. The experimenter gave an example of the expected result via a recording. The recordings were made with an Eminent microphone connected to a Mac-Book computer, using PRAAT-software (Boersma (2001)). For each vowel two recordings per speaker were selected that met our criteria for constant pitch and elongated duration. For all target tokens the first three formants of every vowel were determined using PRAAT software (plotted in Figure 3.11). In the figure we added the mean formant values obtained by Adank et al. (2004) for 10 male and 10 female speakers for comparison. Formant frequencies might differ in sustained isolated vowels. A complete (unmanipulated) version of every vowel was stored, as well as an incomplete (manipulated) version in which the area of the second formant with a bandwidth of 1250 Hz was suppressed from the acoustic signal, using the FFT filter in Adobe Audition 1.5. The upper cut-off frequency of the suppressed area was determined by adding half the band-width of the second formant to the second formant value. After F2 suppression, peak normalisation to a normalisation level of 40 dB was performed using Adobe Audition 1.5. in order to present stimuli of similar intensity. Figure 3.11 shows the mean

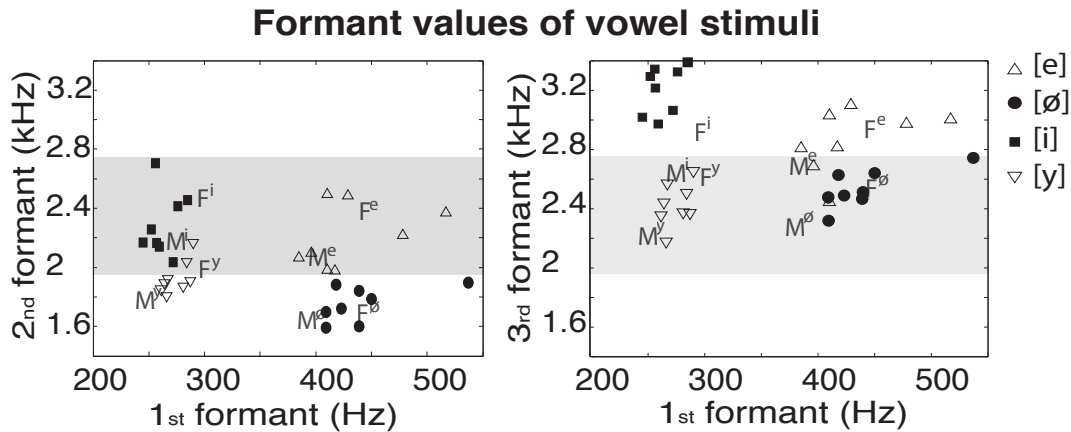


Figure 3.11: Formant values of the first three formants of the recorded vowels as determined with PRAAT (Boersma, 2001). Reference formant values as determined by Adank et al. (2004) are denoted by M(male) and F(female). The grey area is added to make the overlay in third formant values of the rounded vowels (grey area in the right panel) with the second formant values of the unrounded vowels (grey area in the left panel) visible.

spectra (averaged over 400 msec.) of the original as well as the manipulated signals for one of the male speakers.

Experimental procedure

Both variants of all recorded vowels were presented once in random order (9 vowels x 4 speakers x 2 recordings x 2 variants (complete, incomplete) = 144 trials) using E-Prime 2.0 software (Psychology Software Tools). The stimuli were presented diotically at a comfortable level in a quiet experimentation room over EM3561 R1 headphones directly connected to the computer soundcard. The participants were sitting in front of the computer, facing the computer screen. Before the actual task started, the participants were presented with both oral and written instructions. The task was to indicate after each stimulus presentation which vowel they perceived and how certain they were about this percept on a 5-point Likert scale. Subsequently, two examples from the distractor vowels list were presented, following the procedure of the actual task. The participants could ask for clarification of the task instructions after presentation of those two example vowels.

Each presentation of a vowel was initiated by a key-press of the participant. The screen remained black during the stimulus presentation. After the stimulus presentation a list of possible responses appeared on the screen, consisting of all monophthongal Dutch vowels /i, ɪ, ʏ, y, e, ε, ø, o, ə, a, ɑ, u/. Those response options were presented on the screen with the grapheme that is normally used in Dutch ("i.e.", "i", "u", "uu", "ee", "e", "eu", "oo", "o", "aa", "a" and "oe", respectively). Additionally, an example word was presented on the screen for each of the response options, consisting of a word where the vowel is used in its typical form. The participants could indicate what vowel

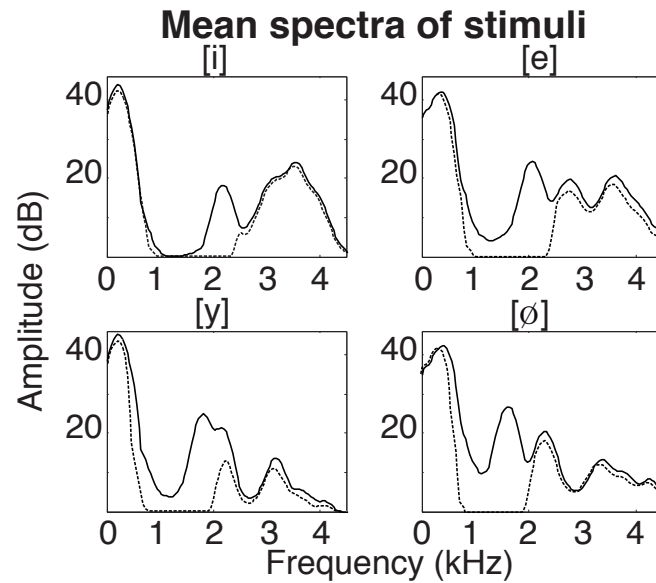


Figure 3.12: Sample spectra of the vowels produced by one of the male speakers (averaged over 400 msec. of both utterances of every target vowel) before (solid line) and after (dotted line) manipulation

they perceived by key press on a QWERTY key-board and subsequently indicate how certain they were about this percept. They had to type the Dutch grapheme belonging to the vowel they perceived, e.g. <uu> for /y/, <ie> for /i/, etc. which is how the writing of vowels is taught at primary school in the Netherlands. The data collection lasted for 20 minutes, and the entire session, including the short preview, was completed in less than 30 minutes.

Response analysis

The answers for the four vowels of interest to this study /i, e, y, ø/ were analysed. Primary inspection of the data revealed a shift in reported identities within this front vowel category in case of incomplete vowels. However, lax vowels such as /ɪ/ or /ʏ/ were rarely chosen, probably due to their short duration in natural speech compared to the long duration of the presented vowels. In order to keep the visualization of classification simple, we chose to label the answers into three categories: "target vowel" (the original vowel), "(un)rounded counterpart" and "other". Thus, /y/ and manipulated /y/ were classified as "target vowel" when they were perceived as /y/ and classified as "counterpart" (unrounded in this example) when they were perceived as /i/. Any other reported percept, including the lax vowel /ɪ/ and /ʏ/, was classified as "other".

Experiment 6: Results

Consistency data

Figure 3.13 shows the confusion matrices of the responses. Figure 3.14 shows the summarised data as explained in Section 3.3. It shows the average consistencies on the perception of the target vowels over all participants. The light grey bars in Figure 3.14

response (proportions)	Presented Stimuli							
	Unrounded Front Vowels				Rounded Front Vowels			
	[i]	[e]	[i]-F2	[e]-F2	[y]	[ø]	[y]-F2	[ø]-F2
/i/	0.98		0.93				0.59	
/e/	0.02	0.94	0.04	0.96			0.01	0.68
/y/					0.93		0.28	
/ø/					0.03	0.99	0.01	0.27
/ɨ/					0.04	0.01	0.05	
/i/			0.01					
/ɛ/		0.06	0.02	0.04				0.03
/o/								0.02
/u/							0.06	
/a/								
/ɔ/								
/ɑ/								

Figure 3.13: Perceived vowel quality as confusion matrix

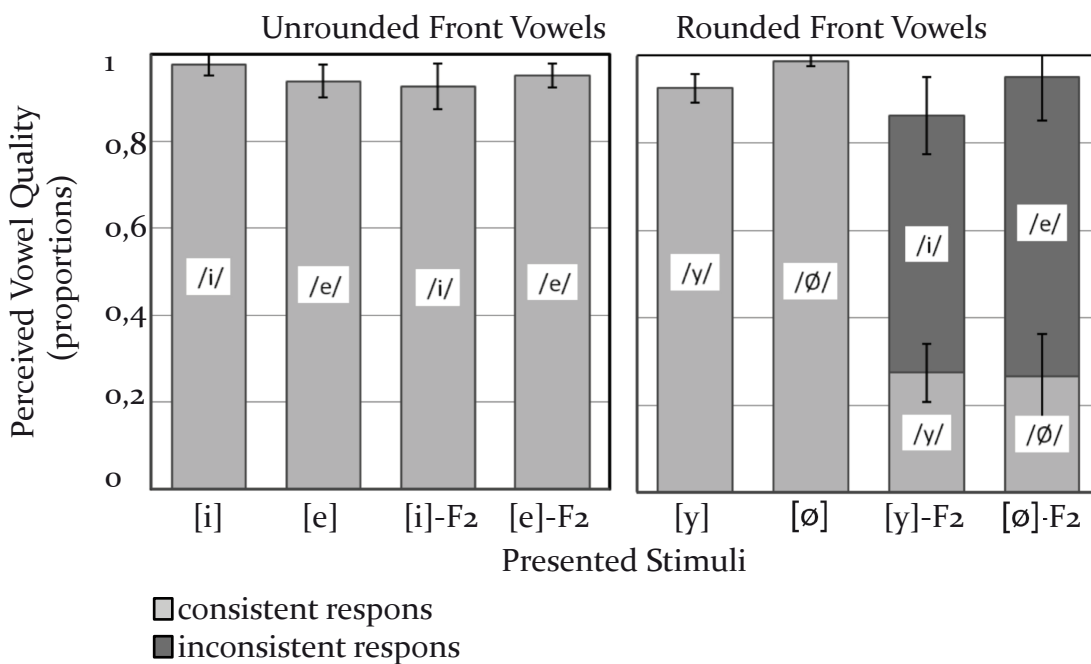


Figure 3.14: Perceived vowel quality averaged over the participant scores. Perception of both filtered (-F2 indicates suppressed second formant) as unfiltered primary (PCV) and secondary cardinal vowel (SCV). A light grey bar represents a correctly perceived vowel and a dark grey bar represents a vowel perceived as PCV-counterpart.

indicate the perceived original vowels. The scores are displayed vertically, depicting response proportions. The manipulated, rounded vowels are in some cases perceived as unrounded counterparts, which is depicted by dark grey bars for visual convenience. The results show that the unrounded vowels were perceived consistently with the original stimuli in both complete and in-complete conditions (mean score > 0.925 , standard error < 0.05 for all four stimulus types). Also, the unmanipulated rounded vowels were perceived consistently with the original stimuli (mean score > 0.92 , standard error < 0.03 for both vowels). However, the incompletely specified rounded vowels /y, ø/ were significantly more often perceived as their unrounded counterparts /i, e/ (mean scores 0.59 and 0.68, standard errors 0.09 and 0.1 respectively) than as the original rounded front vowels /y, ø/ (mean scores 0.28 and 0.27, standard errors 0.07 and 0.1).

A logistic model was fit to this consistency data to test the contribution of manipulation and vowel class (rounded or unrounded vowel) on correct recognition of the presented vowel. According to the model, presenting the stimulus in its normal (unmanipulated) form contributes significantly to the correct identification of a presented vowel ($\beta = -5.91$ with unmanipulated vowels as reference, $p < 0.001$). Also, vowel class adds significantly to the predictory value of the model ($\beta = -4.80$ with rounded vowels as reference, $p < 0.001$). In other words, manipulated, rounded vowels are less well identified than any of the vowels of the other conditions.

Certainty data

Figure 3.15 shows the average certainty of perceived vowel quality on a Likert scale (that ranges from 1 "very uncertain" to 5 "very certain"). The certainty score is lowest for both manipulated, rounded vowels /y/ and /ø/ with a Median of 4. All other stimuli have a Median certainty score of 5. A Kruskal-Wallis test indicated that the certainty of perceived vowel quality differed between some of the conditional groups ($\chi = 66.5$, $df = 7$, p -value < 0.001). Post-hoc comparisons with Wilcoxon Mann-Whitney revealed that participants reported significantly higher confidence for the unmanipulated rounded vowels /y, ø/ than for the manipulated rounded vowels; $W = 3173$ and 4092 , $p < 0.001$ (adjusted for Bonferroni correction) respectively. Confidence was not significantly different between the unmanipulated and manipulated unrounded vowels.

Experiment 6: Discussion

We argued that the investigation of speech processing in noise, with noise characterised by missing and superfluous features, can lead to new perspectives on noise robust speech processing and models for HSR. Here, we investigated human vowel processing when target segments were missing due to manipulation of vowels. This is done in spectral restoration experiments, where it is shown that spectrally induced gaps can be perceptually restored by adding noise (Warren et al., 1997). In the current experiment we showed that spectral distortion by silences leads to perceptual shifts in one direction, namely towards the unrounded front vowels.

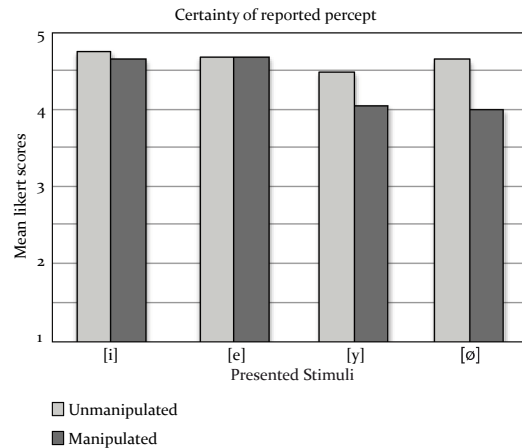


Figure 3.15: Mean reported certainty of the perceived identity of the vowel stimulus as obtained on a Likert scale ranging from 1 (very uncertain) to 5 (very certain)

To summarise, we presented the vowels /y, ø, i and e/ with and without second formant information in a listening experiment. As an adjustment to existing experiments (Ito et al., 2001; Kiefte & Kluender, 2005) we presented naturally-produced vowels instead of synthetically-produced vowels and rounded vowels in addition to unrounded vowels (as investigated by Ito et al. (2001)). It was hypothesised that if the influence of spectral shape is reduced for the perception of naturally-produced vowels (as suggested by Kiefte & Kluender (2005)), listeners are expected to rely on the available formant information which will lead to systematic identification changes when a formant is suppressed. This was true for the perception of /y/ and /ø/ but not for the perception of /i/ and /e/. These results suggest that, for the vowels /y/ and /ø/ the second formant is crucially relied on. It was hypothesised that unaltered vowel identifications may be due to perception mainly based on the amplitude information of the higher formants. This was true for the perception of /i/ and /e/. These findings replicate the findings of Ito et al. (2001) where unrounded vowels are perceived correctly when the second formant is energetically suppressed. The results suggest that, for these vowels, the first and third to higher formants are relied on.

Second formant as main indicator of roundedness

The acoustic characteristics of the filtered [i] and [y] (without second formant) hold cues for the identification as both /i/ and /y/ because the F1 of 300 Hz in the signal indicates that the perceived vowel definitely is a high vowel (in the Dutch vowel system this is /i/ or /y/). If an F2 of 2500 Hz were available in the signal, this would induce the sensation of /i/; if a lower F2 of 2000 Hz were available in the signal, this would result in the perception of /y/. Because we omitted the F2 information from the signal, we omitted the most important perceptual cue to distinguish between rounded /y/ and unrounded /i/ and other cues must be relied on. In case of a filtered [e] and [ø] the remaining acoustic cue of F1 indicates that the perceived vowel is a mid-vowel (in the Dutch vowel

system this is /e/ or /ø/). Roundness can not be discerned by F2 information because it is suppressed and therefore other cues must be relied on. We formulated three potential influences on the perception of manipulated vowels that together account for the data: formant substitution, a mismatch of formant amplitude and an educated guess.

Formant substitution

Figure 3.11 shows that the first and third formant of the vowels [y] and [ø] are similar to the first and second formant of [i] and [e] respectively (grey areas are added in Figure 3.11 to guide the eye). The substitution explanation assumes that, for the perception of the rounded vowels [y] and [ø], deletion of F2 can cause the still present F3 in the signal to be interpreted as F2. This is in line with the explanation of Aaltonen (1985) for the increased /i/ percepts when energy of the second formant was decreased in synthetically-generated [y]. In case of the manipulated unrounded vowels [i] and [e], the third formant might be interpreted as a second formant, albeit an extremely high one. This is also illustrated in Figure 3.11. An extremely high second formant does not lead to an altered percept for the vowels [i] and [e] because the Dutch vowel inventory (see Figure 3.2) does not provide an alternative vowel for these formant values.

Mismatch of formant amplitude

An additional account explains the variability in the data of perceived vowel identity for [y] and [ø]. Namely, the manipulated rounded vowels were not always interpreted as their unrounded counterpart, the percept was ambiguous. This suggests a conflict of different cues. The additional account focusses on the formant amplitude in the 3rd to highest formants as Kiefte et al. (2010) argued that formant amplitude plays a more important role in these formants. Although the still present formant frequencies of the manipulated rounded vowels [y] and [ø] are similar to the first two formants of the unrounded [i] and [e] respectively, the (relative) formant amplitude rather matches the characteristics of the rounded [y] and [ø] than those of the unrounded [i] and [e]. Thus, perceptual evidence is available for the perception of both vowels, leading to an ambiguous percept. The combined accounts explain the variability in perceived identity of the manipulated, rounded vowels as well as the significant decrease in certainty of perceived vowel quality.

Effects of mismatch of formant amplitude are likely to disappear when the silence is replaced by noise, such as is done by Warren et al. (1997). It can be expected that an experimental manipulation where noise of different energy levels is inserted at spectral gaps, provides the most stable percepts on the primary cardinal vowels such as [i] and [e] that were investigated here. This might extend to other primary cardinal vowels such as [a], [ɔ] and [u].

Educated guess

A third influence can explain the specific pattern for the results for manipulated [y] and [ø]; the perceptual shift being bigger for manipulated [ø] than for manipulated [y]. The ambiguity in the acoustic information in combination with the experimental forced response setup may have led to guessing between the activated rounded and unrounded vowel with frequency of occurrence as a guiding principle. Therefore, an educated guess

can result in a structured identification shift towards the more frequently occurring categories /i, e/. An educated guess is a guess based on knowledge; the imprint that the world made in our brain Anderson & Schooler (1991). A high type/token frequency results in a stronger mental activation, leading to a higher chance of perceiving in cases of uncertainty. The effect is most effective when the stimuli are ambiguous as in the experimental data presented in this study. Luyckx, Kloots, Coussé & Gillis (2007) show for Dutch that /i, e/ have a higher type/token frequency than /y, ø/. Therefore, we expect manipulated /y, ø/ to be more often perceived as /i, e/ than as /y, ø/ in case of ambiguity. Also, we expect this effect to be bigger for manipulated [ø] than for manipulated [y], because the type/token frequency ratio is higher for /e:/ø/ than for /i:/y/. Both expectations fit the trends in the data.

Naturally-produced vowels and real-life listening conditions

In this study we used naturally-produced Dutch rounded and unrounded vowels to investigate the effect of missing formants on vowel perception. Missing formant information was suggested to be a potential result of noise on vowel processing. We used naturally-produced vowels as it was questioned whether results found with synthetically-generated vowels hold for natural vowels (Jacewicz, 2005; Kiefte & Kluender, 2005). We found that (part of) the data obtained by Ito et al. (2001) and Aaltonen (1985) on synthetically-generated vowels, could be reproduced with naturally-produced vowels. Also, we found that the results can be extended to the rounded vowels [y] and [ø]. This finding serves as evidence that human listeners use formant frequency as well as other spectral characteristics from natural speech for the identification of vowels.

Noise might force the listener to rely more on formant frequencies than on other spectral characteristics, as formants, exhibiting relatively high energy levels, can be more stable in noise than spectral shape representations. The results can not be directly transferred to real-life listening conditions because different acoustical conditions may change the relative importance of different cues. However, we showed that when some formants are not available, a cluster of vowels that comply with the input can be activated, similar to the findings for the perception of manipulated [y] and [ø] that seemed to activate both /y, i/ and /ø, e/ respectively. In real life the context disambiguates such percepts but with no context available in the experiment an educated guess was made.

Experiment 6: Conclusions

Both Jacewicz (2005) and Kiefte et al. (2010) posed the question as to whether perceptual cues play the same role in the perception of naturally-produced vowels as they do in synthetically-generated vowels. With this experiment we showed that the perception of naturally-produced vowels is not only determined by formant frequency, but is influenced by other aspects of the spectral shape, similar to the findings of Ito et al. (2001) and Kiefte & Kluender (2005); Kiefte et al. (2010) and Jacewicz (2005) for synthetic vowels. Part of the results obtained with synthetically-generated vowels by Ito et al. (2001) and Aaltonen (1985) were reproduced with our naturally-produced vowels. This indicates that results from highly controlled conditions can be at least partly transmitted to more natural conditions. For our current work the most important conclusion is that the availability of part of the formants can lead to activation of clusters of formants. This aspect is not taken into account in modern approaches for ASR as these systems generally rely on whole spectral shape representations. Representations that represent parts of the spectrum can disambiguate part of the percept and this way reduce the search space.

3.4 Speech processing in noise needs local features and knowledge-driven processing

Modern approaches for ASR do not effectively process local representations, such as ECs that we found to be robust to noise and to aid segmentation (Section 2.4). The number of ECs is not fixed per time-segment whereas all commonly applied statistical methods based on linear algebra demand a fixed number of feature vectors. We showed that one of the effects of noise on local features is that features can be missing or superfluous (Section 2.2). Therefore, because human speech processing is at least partly based on local representations (Ito et al., 2001; Kiefte et al., 2010) we investigated vowel processing with human listeners to improve our understanding of speech processing with local features in noisy speech conditions. Participants were presented with vowels with additional features in the form of incongruent audiovisual vowels (Section 3.2) and missing features (Section 3.3). In the audiovisual perception experiment (Section 3.2) features of different auditory and visual vowel input merged into the perception of one vowel. We showed that visual features were more heavily weighted when auditory features were masked by acoustical noise and vice versa. These results are in line with the findings described by Ma, Zhou, Ross, Foxe & Parra (2009). Ma et al. (2009) demonstrated that human perception data can be effectively modelled in a computational model by integrating weighted featural input of different modalities into a single percept. Ambiguous input in auditory-only perception that was congruent with two vowels, resulted similarly in a single percept. We found evidence that the influence of other auditory characteristics such as formant amplitude (Section 3.3) or spectral shape (Kiefte et al., 2010) increases when local features are less reliable. This provides evidence for independent feature weighting within a perceptual modality, additional to demonstrated independence of feature weighting across modalities.

Independent weighting of local features creates the possibility to adjust the weighing to the reliability of the input. Because modern ASR systems use global instead of local representations of speech, frequency-wise weighting of features can not be performed. In contrast to approaches for ASR, the computational models for HSR have the possibility to process local features. (Scharenborg, Wan & Moore (2006) and Scharenborg (2007) adjusted the representations of an existing model for HSR such that articulatory features can be processed.) HSR models are developed with the goal to better understand the linguistic speech processing phenomena. Linguistic phenomena are effectively modelled with both bottom-up (Cutler & Norris, 1979; Norris et al., 2000) and top-down (McClelland & Elman, 1986) HSR models. However, non-linguistic phenomena, such as a speech-in-noise paradigm are not investigated with these type of models. A speech-in-noise paradigm, under the assumption of local representations, forces us to also deal with superfluous and missing features. In the current models for HSR an increasing number of superfluous features leads to inefficient processing because all input features are processed (illustrated in the left part of Figure 3.16).

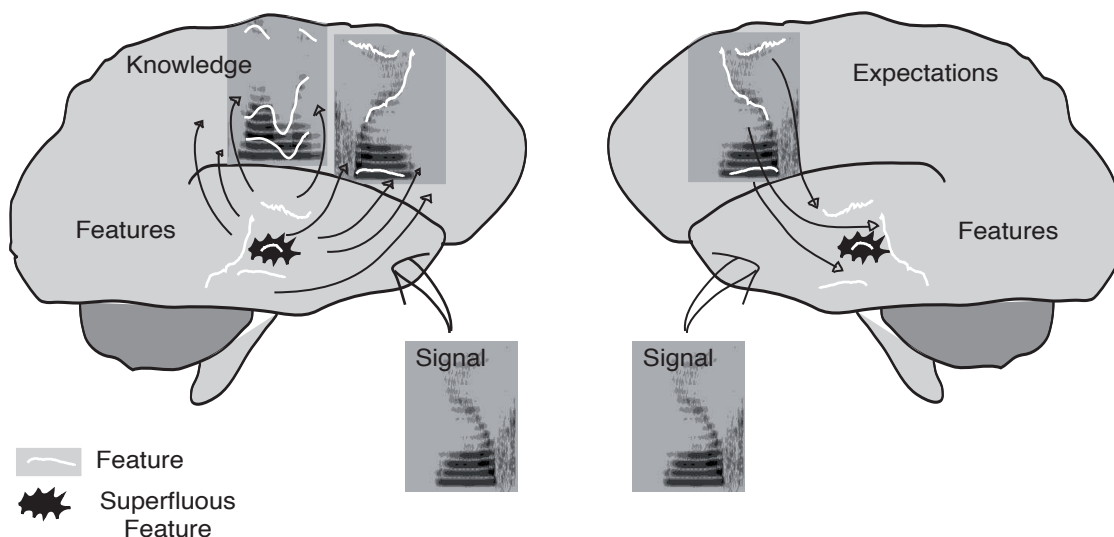


Figure 3.16: (a) In the original view of signal processing the signal is broken into elementary parts (features) that are recombined in a recognition model. (b) In knowledge guided signal processing the expected elementary parts activated from the knowledge level downwards to the feature level. This way processing load remains relatively low when superfluous features are extracted

Another group of models are developed with the goal to better understand speech in noise or mixtures of sounds. Barker et al. (2005) show how an iteration of signal-driven processing and knowledge-guided selections can lead to robust speech decoding. In this model knowledge based expectations are matched to the signal input. Subsequently hypotheses can be formed regarding the remaining noise components and once the noise components are estimated, the target expectations can be fine-tuned and a new iteration can be started. The bootstrapping knowledge that starts the process is suggested to be as simple as the pitch of speech. Because the knowledge based expectations are based on fragments, local representations of speech, the approach can profit from partial ex-

tractions. Another approach to modelling speech in noise is given by Yildiz, Kriegstein & Kiebel (2013). They use adaptive weighting of a knowledge-driven and a signal-driven component in a word recognition model. They developed and investigated a speech recognition model where input deviations (due to recording quality, dialects and competing speaker noises) lead to an increasing weight for expectations. This was modelled by allowing the precision of a match between input and knowledge representation to be relatively low. This weighting of bottom-up or top-down processing can be adjusted during recognition based on precision-weighted prediction errors. This model can explain input deviations such as recording quality, dialects and competing speaker noises. In my understanding the model of Barker et al. (2005) incorporates the main quality of Yildiz et al. (2013) model as the adjustable weighting of bottom-up and top-down processing is indirectly incorporated in the iterative process described by Barker et al. (2005).

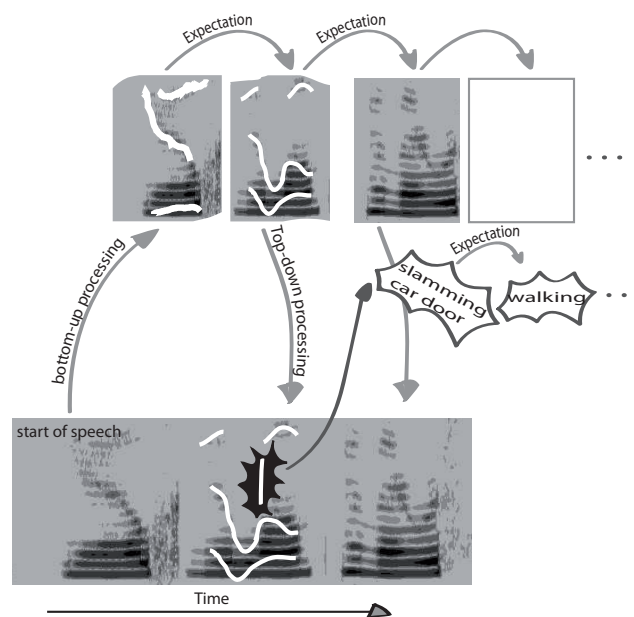


Figure 3.17: Knowledge based processing can not rely on expectations only. This figure illustrates how bottom-up feature processing activates expectations and also when the input does not meet the expectations, only bottom-up processing can keep the process running.

An interactive approach that is based on knowledge-driven feature weighting and signal-driven activation of expectations, such as described by Barker et al. (2005) solves the problem with inefficiency that we described for the models driven by linguistic purposes that process all input features. Figure 3.17 illustrates how the mismatch of expectations with the input can activate a new expectation cycle when signal-driven processing is integrated with knowledge-driven processing. In this interactive account not all features need to be processed before perception and is therefore expected to suffer less from inefficiency when noise leads to many superfluous features or disturbs the feature extraction process.

- From the currently obtained data we found supporting evidence that features can be weighted independently in human speech processing. This finding is in contrast with automatic speech processing that generally relies on global features that do not allow independent weighting of different frequency regions.
- Independent feature weighting allows for the processing of mainly expected, relevant features. Features that originate from noise instead of target-speech are not necessarily processed. We showed that HSR-models can explain human perception data for stimuli with missing and superfluous input features. However, because all features are processed in these models, efficiency decreases with the increase of superfluous features.

General discussion: Key-word spotting by humans and machines

4.1 Key-word spotting

In the current work we investigated two factors that are related to the segmentation and noise problem as identified in current ASR and automatic KWS approaches (Section 1.1 of this thesis). The first factor is the poor representation of speech in ASR (Moore, 2007; Li & Allen, 2011; Rabiner, 2003; Dusan & Rabiner, 2005; Li et al., 2014; Livescu, Fosler-Lussier & Metze, 2012; Cutajar et al., 2012). Therefore, we investigated automatic, robust, *signal-guided* speech representations (Chapter 2). The second factor is the structuring character of knowledge on perceptual processes. To obtain a better understanding of the influence of knowledge on perception, we investigated *knowledge-driven* speech processing by humans (Chapter 3). A consequence of *integrating* signal-guided (local) feature processing in ASR and automatic KWS is that knowledge needs to be applied in an expectancy driven manner. We will discuss this in the subsequent sections.

4.1.1 Signal-guided feature detection

The main contribution of Chapter 2 (Machine speech recognition: Signal-guided processing) of this thesis is that it clarifies the discrepancy between the pervasively used global representations and local, signal-guided representations. Global representations are optimised to represent speech in clean speech conditions, but are vulnerable to many types of noise and not correlated to the temporal characteristics of words which is related to the segmentation problem. Local, signal-guided representations are less thoroughly investigated but shown to be robust to noise and capturing segmentation information. These findings (as discussed in Section 2.4) demonstrate that a renewed focus on the featural representation of speech can be the first step in solving the problem of noise and explaining segmentation effects in automatic speech processing.

An evaluation of the literature on automatically extracted speech representations showed that *local, energetic features* address the segmentation and the noise problem (Section 2.4). However, we observed that the evaluated methods measure indirect characteristics of the signal as features, whereas many speech-related characteristics are directly available in the signal. Therefore, following Andringa (2002), we developed a method for *signal-guided feature detection* where local features follow directly from the signal. Signal-guided representations, such as harmonic complexes, tones (stemming from vowels and voiced phonemes), pulses (plosive phonemes) and noisy structures (fricative phonemes) are visible in the cochleogram representation of speech. Of these

structures, HCs can function as basic speech processing structures for human listeners as described by Bregman (1990). In line with this, we investigated local, energetic structures in the signal that were based on HCs. This resulted finally in a representation of voiced speech sounds that can be determined without prior segmentation and is highly noise robust (ECs, Chapter 2). Based on the underlying theory this representation can be considered similar to the glimpses described by Cooke (2006), but has the additional advantage that it can be established without prior knowledge of the noise. Subsequently, with the goal to obtain a better understanding of the relative usefulness of the algorithm, we investigated global representations based upon the same HC extraction method. The harmonic complexes can be determined relatively unaffected in noisy speech with SNR ranging from 30dB to 10dB. However, information is lost by using the spectral shape information based on HCs even in clean speech conditions. Also, partially extracted HCs are not always processed whereas they do provide information that can serve disambiguation of phonemes or words. Because the current back-end systems do not have the flexibility to process such partial representations, full profit of the detections is not taken. Although it might be possible to improve the HC-extraction algorithm, further evaluation of the HC-extraction does not help in our main aim to understand the representation and processing of speech.

We found that one of the problems with the investigation of new features is the integration of local features with existing automatic speech processing approaches. Existing back-end systems have the undesirable effect that features need to be adjusted to the back-end system. For example, the commonly applied methods, generally statistical methods based on linear algebra, demand a fixed number of input features. As the ECs do not match this criterion it restrains evaluation of the ECs. Therefore, in order to prevent stagnation, we decided to investigate the processing of local features further with *perceptual research*. By doing so, the perceptual system fulfils the function of a flexible back-end system that is needed to test the speech representations.

4.1.2 Knowledge-driven feature weighting

The results presented in Chapter 3 demonstrate the effectiveness of independent weighting of local features, based on their reliability, for vowel perception. We performed two perception experiments with superfluous and missing features respectively. Superfluous and missing features are characteristic mistakes made in the detection stage with local features (as illustrated in the right panel of Figure 4.1). This type of mistakes is made before mapping to a phoneme which contrasts the errors that are made with global, spectral shape features that are generally made in the classification stage (as illustrated in the left panel of Figure 4.1). From an analysis of the results we concluded that local features provide the possibility to weight the features independently (Section 3.4) which becomes especially relevant when not all input features are reliable, such as is the case in noisy speech conditions.

We evaluated back-end systems on the possibility to weight features independently in order to obtain a better understanding of speech processing in noise. Because ASR back-end systems demand a fixed number of input features, we focussed on models for

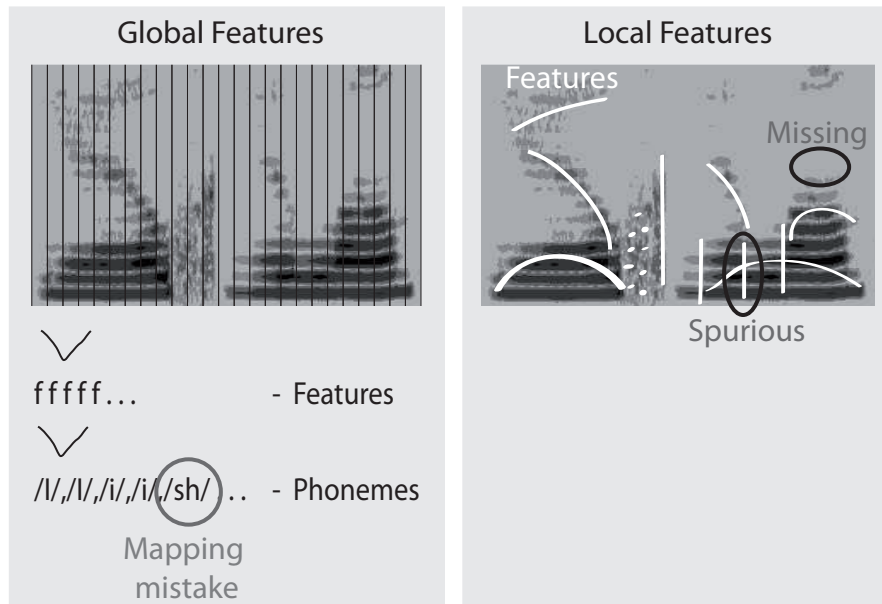


Figure 4.1: *The type of errors that are made by global features and local features are different in character. For global features (left panel) deviant mappings that are made after feature extraction are considered mistakes. For local features (right panel) some extractions themselves are deviant and considered errors. As a result of this different character of the mistakes also the engineering focus differs in character for both approaches.*

HSR. We analysed how existing models for HSR explain data with local features and how they apply feature weighting when noise leads to superfluous features. The commonly used models (Cutler & Norris, 1979; Norris et al., 2000; McClelland & Elman, 1986) can explain data with missing and superfluous phonemes (Norris et al., 2000) as demonstrated with the phoneme restoration task (Warren, 1970; Samuel, 1996) and phonemic decision making task (Ganong, 1980). We assume that the models can similarly explain data with missing and superfluous features when they are adjusted to process local features (Scharenborg et al., 2006; Scharenborg, 2007, showed that this is possible for articulatory features). We argued that an increasing ratio of superfluous to target features leads to inefficiency (Section 3.4) because both noise-related features and target features are processed in these models (Cutler & Norris, 1979; Norris et al., 2000; McClelland & Elman, 1986).

In line with these findings Yildiz et al. (2013) showed that noise based weighting (as exemplified by noise-adjusted matching precision of signal and knowledge) leads to effective, robust processing of words. The results presented by Yildiz et al. (2013) stress the relevance of an integrated approach of signal-based representations and knowledge-based expectations (see also Mattys et al., 2012). As the ECs are robust to noise and provide the opportunity to fine-grained weighting by the weighting of individual features, this may lead to a method to optimally profit from the local representations, even when noise leads to missing or superfluous features.

4.1.3 Integrating signal-guided and knowledge-driven processes

With Chapter 3 (*Human speech recognition: knowledge-driven and signal-guided processing*) we offered a new perspective on the role of knowledge in speech processing in noise. The integration of characteristics from signal-guided and knowledge-driven approaches in speech processing with local features can help to understand and deal with the problems of segmentation and noise. When a signal-guided mechanism processes all detected features and the number of superfluous features increases, the system is prone to become slow. Efficiency can be retained by integrating a knowledge-driven process that leads to feature processing based on expectancy (illustrated in Figure 3.17) where expectations can be driven by different knowledge levels; word expectations activate phoneme expectations and phoneme expectations activate feature expectations. This pre-activation of individual features that are congruent with knowledge-driven expectations can lead to fast processing. Additionally, all features can be processed in a signal-guided manner. Because the signal itself does not carry information to differentiate between expected and unexpected features this is presumably a relatively slow process. The main function of the signal-guided process is to prevent the system from premature commitment to expectations. When the target input is not congruent with the expectation, signal-guided processing is needed to process the target input.

Combining knowledge-driven processing of expected features to gain efficiency with signal-guided processing to prevent problems with premature commitment provides a useful framework to investigate speech processing with local features. It utilises the robustness and segmentation information that is captured in local features while dealing with the artefacts (superfluous and masked features) when used in noise. If demonstrated to be effective, this approach can evolve towards a system for knowledge-driven key-word spotting, for example by extending the criteria with expectations regarding pulse-components and noise-components that together comply to a target-word. This rule-based approach (rather than neural network or HMM based) fits the aim to understand speech recognition in noise which eventually can help to both improve ASR systems and to intervene in human speech processes.

4.2 The problem of spoken key-word spotting

With this work we proposed a new perspective on speech processing with the goal to solve problems with segmentation and noise in modern ASR and KWS. The newly introduced techniques have the potential to drag ASR out of a local maximum by providing new directions and perspectives. While new perspectives lead to sub-optimal short-term solutions on existing problems (Bourlard et al., 1995) they can be of instant use for new problems (as done for example by Strik & Cucchiraini (1999) and Strik et al. (2009) who performed phoneme recognition on distorted speech input). Similarly, the methods that were developed and tested in this work can be especially of use for the problem of phoneme or key-word detection in noisy speech conditions with no segmentation of the input stream before recognition.

Bibliography

- Aaltonen, O. (1985). The effects of relative amplitude levels of f2 and f3 on the categorization of synthetic vowels. *Journal of Phonetics*, 13, 1–9.
- Abdelatty Ali, A., van der Spiegel, J., & Mueller, P. (2001). Acoustic-phonetic features for the automatic classification of stop consonants. *Speech and Audio Processing, IEEE Transactions on*, 9(8), 833–840.
- Abdelatty Ali, A., Van der Spiegel, J., Mueller, P., Haentjens, G., & Berman, J. (1999). An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech. In *Circuits and Systems, 1999. Proceedings of the 1999 IEEE International Symposium on*, volume 3, (pp. 188–121).
- Abdelatty Ali, A. M., Van der Spiegel, A. M., & Mueller, P. (2001). Acoustic-phonetic features for the automatic classification of fricatives. *The Journal of the Acoustical Society of America*, 109(5), 2217–2235.
- Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of northern and southern standard dutch. *The Journal of the Acoustical Society of America*, 116(3), 1729–1738.
- Anderson, J. & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science*, 2(6), 396–408.
- Andringa, T. (2002). *Continuity preserving signal processing*. PhD thesis, University of Groningen.
- Barker, J. (1998). *The relationship between auditory organisation and speech perception: Studies with spectrally reduced speech*. PhD thesis, University of Sheffield.
- Barker, J., Cooke, M., & Ellis, D. (2005). Decoding speech in the presence of other sources. *Speech Communication* 45, 5–25.
- Başkent, D. (2012). Effect of speech degradation on top-down repair: Phonemic restoration with simulations of cochlear implants and combined electric-acoustic stimulation. *J Assoc Res Otolaryn.*, 13, 683–692.
- Başkent, D. & Bazo, D. (2011). Audiovisual asynchrony detection and speech intelligibility in noise with moderate to severe sensorineural hearing impairment. *Ear and Hearing*, 32, 582–595.

- Bladon, R. (1982). Arguments against formants in the auditory representation of speech. In Carlson, R. & Granström, B. (Eds.), *The representation of speech in the peripheral auditory system*. Elsevier Biomedical Press.
- Bladon, R. & Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *The Journal of the Acoustical Society of America*, 69(5), 1414–1422.
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International*, 5(9-10), 341–345.
- Bourlard, H., Hermansky, H., & Morgan, N. (1995). Towards increasing speech recognition error rates. *Speech Communication*, 18(3), 205–231.
- Bregman, A. (1990). *Auditory scene analysis: the perceptual organization of sound*. MIT press.
- Bush, M. (1983). Selecting acoustic features for stop consonant identification. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, volume 8, (pp. 742–745).
- Champoux, F., Lepore, F., Gagneú, J., & Théoret, H. (2009). Visual stimuli can impair auditory processing in cochlear implant users. *Neuropsychologia*, 47, 17–22.
- Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5), 975–979.
- Cohen, P. & Mercer, R. (1975). Automatic speech recognition: A critical survey and discussion of the literature. In *Human communication: A unified view*, (pp. 399–438). McGraw Hill, New York.
- Cole, R., Yan, Y., Mak, B., Fanty, M., & Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, (pp. 853–856).
- Cook, P. (2002). *Real sound synthesis for interactive applications*. A. K. Peters, Ltd. Natick, MA, USA.
- Cooke, M. (1993). *Modelling Auditory Processing and Organisation*. PhD thesis, Cambridge University Press, London.
- Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3), 1562–1573.
- Cooke, M., Green, P., Josifovski, L., & Vizinho, A. (2001). Robust automatic speech recognition with missing and uncertain acoustic data. *Speech Communication*, 34(3), 267–285.
- Cutajar, M., Gatt, E., Grech, I., Casha, O., & Micallef, J. (2012). Comparative study of automatic speech recognition techniques. *IET Signal Processing*, 7(1), 25–46.
- Cutler, A., Mehler, J., Norris, D., & Seguí, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, 19, 141–77.

- Cutler, A. & Norris, D. (1979). Monitoring sentence comprehension. In *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Walker, Erlbaum.
- Davis, M. & Johnsrude, I. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, 229(1-2), 132–147.
- De Mori, R. & Flammia, G. (1993). Speaker independent consonant classification in continuous speech with distinctive features and neural networks. *The Journal of the Acoustical Society of America*, 94(6), 3091–3103.
- Deng, L., Cui, X., Pruvencok, R., Chen, Y., Momen, S., & Alwan, A. (2006). A database of vocal tract resonance trajectories for research in speech processing. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (pp. 369–372).
- Diehl, R. & Lindblom, B. (2004). Explaining the structure of feature and phoneme inventories: The role of auditory distinctiveness. In Greenberg, S., Ainsworth, W., Popper, A., & Fay, R. (Eds.), *Speech Processing in the Auditory System*, (pp. 119). New York: Springer-Verlag.
- Dusan, S. & Rabiner, L. (2005). Can automatic speech recognition learn more from human speech perception? In *Trends in Speech technology*, (pp. 21–36). Romania; Romanian Academic Publisher, 2005.
- Ellis, D. (1996). *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT Press, Cambridge, MA.
- Ellis, D. (1998). Using knowledge to organize sound: The prediction driven approach to computational auditory scene analysis, and its application to speech / non-speech mixtures. *Speech Communication*, 27(3-4), 281–298.
- Evanini, K., Isard, S., & Liberman, M. (2009). Automatic formant extraction for sociolinguistic analysis of large corpora. In *Interspeech*.
- Field, J. (2003). *Psycholinguistics: A resource book for students*. United Kingdom, London: Routledge.
- Ganong, W. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 110–125.
- Gemmeke, J. & Cranen, B. (2009). Missing data imputation using compressive sensing techniques for connected digit recognition. In *DSP, Santorini, Greece*, (pp. 1–8).
- Gemmeke, J., Cranen, B., & Remes, U. (2011). Sparse imputation for large vocabulary noise robust asr. *Computer Speech and Language*, 25, 462–479.
- Gläser, C., Heckmann, M., Joublin, F., & Goerick, C. (2010). Combining auditory preprocessing and bayesian estimation for robust formant tracking. In *Audio, Speech, and Language Processing, IEEE Transactions on*, volume 18, (pp. 224–236).

- Gong, Y. (1995). Speech recognition in noise environments: A survey. *Speech Communication, 16*, 261–291.
- Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditoryvisual integration. *The Journal of the Acoustical Society of America, 103*, 2677–2690.
- Green, P., Cooke, M., & Crawford, M. (1995). Auditory scene analysis and hmm recognition of speech in noise. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (401-404).
- Grossberg, S. & Kazerounian, S. (2011). Laminar cortical dynamics of conscious speech perception: neural model of phonemic restoration using subsequent context in noise. *J Acoust Soc Am, 130*(1), 440–460.
- Gussenhoven, C. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, chapter Dutch, (pp. 74–77). Cambridge University Press.
- Heckmann, M., Domont, X., Joublin, F., & Goerick, C. (2008). A closer look on hierarchical spectro-temporal features (hist). In *Proc. Interspeech 2008*, (pp. 894–897).
- Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical Society of America, 97*, 3099 – 3111.
- Hinton, G., Deng, L., Dong, Y., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine, 29*(6), 8297.
- Holmes, H. & Holmes, W. (2002). *Speech synthesis and recognition* (2 ed.). Taylor and Francis, London and New York.
- Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM, 57*(1), 94–103.
- Irino, T. & Patterson, R. (1997). A time-domain, level-dependent auditory filter: The gammachirp. *The Journal of the Acoustical Society of America, 101*(1), 412 – 419.
- Ito, M., Tsuchida, J., & Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *The Journal of the Acoustical Society of America, 110*(2), 1141–1149.
- Jacewicz, E. (2005). Listener sensitivity to variations in the relative amplitude of vowel formants. *Acoustics Research Letters Online, 6*(3), 118–124.
- Jacobi, I. (2009). *On variation and change in diphthongs and long vowels of spoken Dutch*. PhD thesis, Amsterdam.

- Jones, D. (1963). *The Pronunciation of English*. Cambridge University Press, Cambridge.
- Kewley-Port, D., Burkle, Z. T., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearingimpaired listeners. *The Journal of the Acoustical Society of America*, *122*, 2365–2375.
- Kiefte, M., Enright, T., & Marshall, L. (2010). The role of formant amplitude in the perception of /i/ and /u/. *The Journal of the Acoustical Society of America*, *127*(4), 2611–2621.
- Kiefte, M. & Kluender, K. R. (2005). The relative importance of spectral tilt in monophthongs and diphthongs. *The Journal of the Acoustical Society of America*, *117*(3), 1395–1404.
- Kienast, M. & Sendlmeier, W. (2000). Acoustical analysis of spectral and temporal changes in emotional speech. In *ISCA Workshop on Speech and Emotion, Northern Ireland*.
- Kirchhoff, K., Fink, G., & Sagerer, G. (2002). Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, *37*, 303–319.
- Kleinschmidt, M. (2002a). Methods for capturing spectro-temporal modulations in automatic speech recognition. *Acustica united with acta acoustica*, *88*, 416–422.
- Kleinschmidt, M. (2002b). *robust speech recognition based on spectrotemporal processing*. PhD thesis, universitaet oldenburg.
- Kleinschmidt, M. (2003). Localized spectro-temporal features for automatic speech recognition. In *Proc. Eurospeech 2003*.
- Krijnders, J., Niessen, M., & Andringa, T. (2010). Sound event recognition through expectancy-based evaluation of signal-driven hypotheses. *Pattern Recognition Letters*, *31*, 1552 – 1559.
- Krijnders, J. D. (2010). *Signal-driven sound processing for uncontrolled environments*. PhD thesis.
- Ladefoged, P. (1982). *A course in phonetics (6th edition)*, chapter English vowels, (pp. 85–105). New York, NY: Harcourt Brace Jovanovich.
- Lee, K. & Hon, H. (1989). Speaker independent phone recognition using hidden markov models. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *37*(11), 1641–1648.
- Leinonen, T. (2010). *An acoustic analysis of vowel pronunciation in Swedish dialects*. PhD thesis, Groningen.
- Li, F. & Allen, J. (2011). Manipulation of consonants in natural speech. *Audio, Speech and Language Processing, IEEE Transactions on*, *19*(3), 496 – 503.

- Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *Audio, Speech and Language Processing, IEEE Transactions on*, 22(4), 745–777.
- Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1), 1–15.
- Lisker, L. & Rossi, M. (1992). Auditory and visual cueing of the [\pm rounded] feature of vowels. *Language and Speech*, 35, 391–417.
- Liu, S. (1996). Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America*, 100(5), 3417–3430.
- Livescu, K., Fosler-Lussier, E., & Metze, F. (2012). Subword modeling for automatic speech recognition. In *IEEE Signal Processing Magazine*, (pp. 44–57).
- Luyckx, K., Kloots, H., Coussé, E., & Gillis, S. (2007). Klankfrequenties in het nederlands. In Sandra, D., Rymenans, R., Cuvelier, P., & Petegem, P. V. (Eds.), *Tussen Taal, Spelling en Onderwijs*, (pp. 141–154). Gent: Academia Press.
- Ma, W., Zhou, X., Ross, L., Foxe, J., & Parra, L. (2009). Lip-reading aids word recognition most in moderate noise: A bayesian explanation using high-dimensional feature space. *PLoS ONE*, 4(3).
- Maddieson, I. (1984). *Patterns of Sounds*. Cambridge, Massachusetts.
- Massaro, D. & Stork, D. G. (1995). Speech recognition and sensory integration: A 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist*, 86, 236–244.
- Massaro, D. W. (1987). Single versus multiple sources of speech information: The contribution of visible speech. In *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.
- Massaro, D. W. (1989). Multiple book review of speech perception by ear and eye: A paradigm for psychological inquiry. *Behavioral and Brain Sciences*, 12, 741–794.
- Massaro, D. W. & Cohen, M. M. (1990). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9, 753–771.
- Mattys, S., Davis, M., Bradlow, A., & Scott, S. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7/8), 953–978.
- McClelland, J. & Elman, J. (1986). The trace model of speech perception. *Cognitive Psychology*, 18, 1–86.
- McGrath, M. & Summerfield, Q. (1985). Intermodal timing relations and audiovisual speech recognition by normal hearing adults. *The Journal of the Acoustical Society of America*, 77, 678–685.

- McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264* (December 23), 746–748.
- McQueen, J., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*(621-638).
- Miller, G. A. & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *77*, 338–352.
- Miller, L. M. & D’Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience*, *25*, 5884–5893.
- Molis, R. (2005). Evaluating models of vowel perception. *The Journal of the Acoustical Society of America*, *118*(2), 1062–1071.
- Moore, B. (1996). A revision of zwicker’s loudness model. *Acustica*, *82*(2), 335 – 345.
- Moore, R. (2007). Spoken language processing: Piecing together the puzzle. *Speech Communication*, *49*(5), 418–435.
- Mury, T. & Sigh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical Society of America*, *68*(1294-1300).
- Mustafa, K. & Bruce, I. (2006). Robust formant tracking for continuous speech with speaker variability. *Audio, Speech and Language Processing, IEEE Transactions on*, *14*(2), 435 – 444.
- Nakatani, T. & Irino, T. (2004). Robust and accurate fundamental frequency estimation based on dominant harmonic components. *The Journal of the Acoustical Society of America.*, *116*, 3690–3700.
- Niessen, M., Krijnders, J., & Andringa, T. (2009). Understanding a soundscape through its components. In *Proceedings of Euronoise*.
- Norris, D., McQueen, J., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–370.
- O’Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, *41*(10), 2965 – 2979.
- Otsu, N. (1979). A threshold selection method from gray-level histograms. *Systems, Man, and Cybernetics, IEEE Transactions on*, *9*(1), 62–66.
- Pickett, J. M. (1957). Perception of vowels heard in noises of various spectra. *The Journal of the Acoustical Society of America*, *29*, 613–620.
- Plomp, R., Pols, L. C. W., & van der Geer, J. (1967). Dimensional analysis of vowel spectra. *The Journal of the Acoustical Society of America*, 707–712.
- Pols, L. C. W., Tromp, H. R. C., & Plomp, R. (1973). Frequency analysis of dutch vowels from 50 male speakers. *The Journal of the Acoustical Society of America*, *53*, 1093–1101.

- Rabiner, L. (2003). The power of speech. *Science*, *301*(5639), 1494–1495.
- Rath, T. & Manmatha, R. (2007). Word spotting for historical documents. *International Journal on Document Analysis and Recognition*, *9*(2), 139–152.
- Remez, R., Rubin, P., Pisoni, D., & Carrell, T. (1981). Speech perception without traditional speech cues. *Science*, *212*(4497), 947–950.
- Rietveld, A. C. M. & van Heuven, V. J. (2009). *Algemene fonetiek [General phonetics]*, chapter Productie van spraakklanken [Production of speech sounds], (pp. 55–91). Bussum, the Netherlands: Uitgeverij Coutinho.
- Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of french oral vowels in noise. *The Journal of the Acoustical Society of America*, *103*, 3677–3689.
- Rosner, B. S. & Pickering, J. B. (1994). *Vowel perception and production*. New York, NY: Oxford University Press.
- Rouger, J., Fraysse, B., Deguine, O., & Barone, P. (2008). McGurk effects in cochlear-implemented deaf subjects. *Brain Research*, *1188*, 87–99.
- Rubin, P., Turvey, M. T., & van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in nonwords. *Perception and Psychophysics*, *19*, 394–398.
- Samuel, A. (1996). Phoneme restoration. *Language and Cognitive Processes*, *11*, 647–654.
- Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication - Special Issue on Bridging the Gap between Human and Automatic Speech Processing*, *49*, 336–347.
- Scharenborg, O., Wan, V., & Moore, R. K. (2006). Capturing fine-phonetic variation in speech through automatic classification of articulatory features. In *Speech Recognition and Intrinsic Variation*, (pp. 77–82).
- Schorr, E., Fox, N., van Wassenhove, V., & Knudsen, E. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 102, (pp. 18748–18759).
- Seltzer, M., Yu, D., & Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (pp. 7398–7402).
- Simpson, G. B. (1984). Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin*, *96*(2), 316–340.
- Strik, H. & Cucchiraini, C. (1999). Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication*, *29*(1), 225–246.

- Strik, H., Truong, K., de Wet, F., & Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 51(10), 845–852.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B. & Campbell, R. (Eds.), *Hearing by eye: The psychology of lipreading.*, (pp. 3–51). Hillsdale, NJ: Erlbaum.
- Tabossi, P. & Zardon, F. (1993). Processing ambiguous words in context. *Journal of Memory and Language*, 32(3), 359–372.
- Traunmüller, H. & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, 35, 244–258.
- Valkenier, B., Duyne, J., Andringa, T., & Baskent, D. (2012). Audiovisual perception of congruent and incongruent dutch front vowels. *Journal of Speech, Language and Hearing Research*, 55(6), 1788–1801.
- Valkenier, B., Krijnders, J., van Elburg, R., & Andringa, T. (2011). Psycho-acoustically motivated formant feature extraction. In *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011*, volume 11, (pp. 218–223).
- van de Vooren, H., Violanda, R. R., van Elburg, R., & Andringa, T. (2010a). Time-frequency tracks as segments in computational auditory scene analysis: Architectural and evaluative aspects.
- van de Vooren, H., Violanda, R. R., van Elburg, R. A. J., & Andringa, T. C. (2010b). Unpublished Matlab code, can be obtained via a request to t.c.andringa@rug.nl.
- Van der Zant, T., Schomaker, L., & Haak, K. (2008). Handwritten-word spotting using biologically inspired features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11), 1945–1957.
- van Hout, R., Adank, P., & van Heuven, V. J. (2000). Akoestische metingen van nederlandse klinkers in algemeen nederlands en in zuid-limburg [acoustical measures of dutch vowels: General dutch and south limburg]. *Taal and Tongval*, 52, 151–162.
- van Oosten, J. & Schomaker, L. (2014). A reevaluation and benchmark of hidden markov models. In *Proc. Int. Conference on Frontiers in Handwriting Recognition, Crete, Greece, IEEE Computer Society*, (pp. 531–536).
- van Son, R. (1994). A method to quantify the error distribution in confusion matrices. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 18, (pp. 41–63).
- Violanda, R. R., van de Vooren, H., van Elburg, R. A. J., & Andringa, T. C. (2009). Signal component estimation in background noise. In *Proceeding of NAG/DAGA*, (pp. 1588–1591).
- Wang, D. & Brown, G. (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEEpress/ Wiley-Interscience.

- Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393.
- Warren, R., Hainworth, K., Brubaker, B., Bashford, J., & Healy, E. (1997). Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps. *Perception and Psychophysics*, *2*, 275–283.
- Wester, M. (2003). Syllable classification using articulatory-acoustic features. In *Proc. Eurospeech 2003.*, (pp. 233–236).
- Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S., & Boulard, H. (2004). Evaluation of formant-like features on an automatic vowel classification task. *The Journal of the Acoustical Society of America*, *116*(1781 - 1791).
- Whiteside, S. (1998). Identification of a speaker's sex: a study of vowels. *Perceptual and Motor Skills*, *86*(2), 579–584.
- Witten, I. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, volume 2nd Edition. Morgan Kaufmann, San Francisco.
- Wood, N. & Cowan, N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology*, *21*(1), 255–260.
- Yan, Q., Vesghi, S., Zavarehei, E., Milner, B., Darch, J., White, P., & Andrianakis, I. (2007). Formant tracking linear prediction model using hmms and kalman filters for noisy speech processing. *Computer Speech and Language*, *21*(543 - 561).
- Yildiz, I., Kriegstein, K., & Kiebel, S. (2013). From birdsong to human speech recognition: bayesian inference on a hierarchy of nonlinear dynamical systems. *PLoS Computational Biology*, *9*(9), 1–16.
- Young, S. (1992). The general use of tying in phoneme-based hmm speech recognisers. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '92.*, (pp. 569–572).

Samenvatting (summary in Dutch)

Het uiteindelijke doel van het automatisch herkennen van gesproken woorden (Automatic Keyword Spotting; AKS), zoals we het in het huidige werk hebben gedefinieerd, is om woorden onafhankelijk van de context te herkennen. In ASR lost het gebruik van context-informatie veel ambiguïteiten in de herkenning op, maar daarmee zijn deze ambiguïteiten ook onzichtbaar. De verschuiving in AKS naar het herkennen van individuele woorden zonder gebruik van context-informatie leidt er daardoor toe dat twee problemen in de huidige systemen voor spraakverwerking zichtbaar worden.

Ten eerste is er de moeilijkheid om de stroom van spraakgeluiden, die in het akoestisch domein een relatief continu karakter heeft, te segmenteren in delen die bruikbaar zijn voor herkenning zonder de doelwoorden op te breken. In de huidige systemen wordt dit niet als een probleem aangemerkt en wordt het input-geluid opgedeeld in gelijke delen, onafhankelijk van het begin en het einde van woorden. Ten tweede is er sprake van een grote gevoeligheid voor verstoringen van niet-doel geluiden in bestaande systemen voor spraakverwerking die tot problemen leidt wanneer er geen (zins)context beschikbaar is die helpt bij disambigueren van de input.

Beide problemen worden gerelateerd aan de gebrekkige representatie van spraak enerzijds en het niet gebruiken van verwachting bij de verwerking van de geluidsrepresentaties anderzijds. Deze twee gelieerde factoren worden in dit proefschrift onderzocht. We concentreren ons allereerst op de *representatie van spraak*. Het doel is daarbij om "akoestische kenmerksvectoren" te vinden waarbij de segmentatie en de ruisgevoeligheid verbeterd is ten opzichte van bestaande representaties (veel gebruikte representaties worden geëvalueerd in o.a. Li et al., 2014). Evaluatie van bestaande spraakrepresentaties (andere dan de gebruikelijke MFCC features die niet erg ruisrobuust zijn) leidt tot de conclusie dat sommige spraakrepresentaties een deel van het tijd-frequentie domein selecteren die energie-rijk en daardoor ook ruis-robust zijn. Ook concluderen we dat spraak-representaties die gerelateerd zijn aan articulatie-karakteristieken (fonetische features) de potentie hebben om de stroom van spraakgeluiden te segmenteren in de tijd.

We presenteren een spraak-representatie methode die een lokaal karakter combineert met articulatie-karakteristieken. Deze methode is gebaseerd op de selectie van energetische componenten (ECs) van een harmonisch complex (HC). We laten zien dat deze selecties zowel de ruisrobuustheid als de segmentatie faciliteren. Ze representeren stemhebbende componenten van spraak zonder dat context-informatie nodig is. De ECs zijn vergelijkbaar met de zogenaamde "glimpses" (Cooke, 2006) waar elementen worden geselecteerd die een hoge lokale signaal ruis verhouding hebben. De ECs hebben, ten opzichte van de "glimpses", het voordeel dat de karakteristieken van de ruis niet bekend worden verondersteld om ze te extraheren. Een nadeel van de ECs is dat een deel van de

spraak-gerelateerde informatie verloren lijkt te gaan wanneer HCs worden geëxtraheerd van schone spraak. Dat dit verlies niet erger wordt met toevoeging van milde ruis, waar dit bij standaard-representaties (zonder disambiguatie met behulp van zinscontext) leidt tot onherkenbare input, maakt de ECs toch erg aantrekkelijk als representatie, vooral in omstandigheden met veel niet-doel geluiden. Doordat ECs ruisrobuust zijn en segmentatie van het spraaksignaal mogelijk maken zal de zinscontext niet meer in dezelfde mate nodig zijn voor het disambigueren van de input. De bruikbaarheid van ECs voor spraakherkennings-doeleinden is niet expliciet getest met de bestaande ASR technieken. De reden hiervoor is dat het aantal extracties varieert, wat het ongeschikt maakt als input voor bestaande ASR systemen. Dat de set van extracties varieert is het gevolg van ruis. Door toedoen van de ruis worden sommige ECs niet geëxtraheerd terwijl er ook extra ECs worden geëxtraheerd die een ruis-element representeren in plaats van een articulatie-element. Omdat Cooke (2006) heeft laten zien dat de elementen met een relatief hoge lokale signaal-ruis-verhouding sterk gerelateerd zijn aan de elementen in menselijke spraakperceptie, veronderstellen we dat mensen wel kunnen omgaan met een variabele hoeveelheid input features. Daarom hebben we de menselijke perceptuele processen en het effect van kennis en verwachting op perceptie onderzocht als een tweede factor, om zo een beter begrip te krijgen van het spraakverwerking in ruis.

De tweede factor is de structurerende invloed van *kennis en verwachting* op de waarneming. De rol van kennis in automatische spraakverwerking verandert als spraak wordt gerepresenteerd door lokale representaties zoals die we onderzocht hebben. Omdat deze veranderde rol van kennis in spraakverwerking niet kan worden onderzocht met behulp van bestaande ASR systemen, onderzoeken we de structurerende invloed van kennis en verwachting bij menselijke waarneming van spraakgeluiden. De evaluatie van computationele modellen voor menselijke spraakverwerking suggereert dat zowel de "bottom-up" als "top-down" systemen inefficiënt worden als geluiden die voor de taak niet van belang zijn, leiden tot een toename van input features. In beide typen systemen worden namelijk alle input features verwerkt. Omdat de evaluatie van de computationele modellen ons niet verder brengt hebben we twee perceptie-experimenten uitgevoerd om zo tot een beter begrip van menselijke verwerking van spraak met missende en overvloedige input features te komen. In beide experimenten hebben we stimuli aangeboden van gemanipuleerde Nederlands gesproken klinkers, om een situatie te creëren waarin lokale input features missen danwel toegevoegd zijn. In het *eerste experiment* hebben we de energie onderdrukt rondom de frequentie die hoort bij de tweede formant. We hebben gevonden dat de waarneming van ongeronde klinkers (zoals bijvoorbeeld de /i/) onveranderd blijft terwijl de geronde klinkers (zoals bijvoorbeeld de /y/) worden waargenomen als de ongeronde klinker met dezelfde eerste formant (/i/ in dit voorbeeld). Deze resultaten suggereren dat menselijke waarneming om kan gaan met gedeeltelijke informatie. In het *tweede experiment* hebben we audio-visuele stimuli aangeboden waarbij het geluid en het beeld van twee verschillende Nederlands gesproken klinkers afkomstig waren. In dit experiment namen de participanten één klinker waar per stimulus. De waargenomen klinker was in de meeste gevallen de klinker met de sterkste karakteristiek van beide aangeboden klinkers. Zo is ronding een sterke visuele karaktereigenschap terwijl de plaats van articulatie, de hoogte, een sterke auditieve karaktereigenschap is. In gevallen waar

een sterke visuele eigenschap in één van de twee klinkers werd gecombineerd met een sterke auditieve eigenschap in de andere klinker vond vaak vermenging plaats. De combinatie van de auditieve plaats van articulatie (bijvoorbeeld de hoogte voor /e/) en de visuele manier van articulatie (bijvoorbeeld de ronding voor /y/) leidde in veel gevallen tot de waarneming van de hoge, geronde klinker /ø/. De resultaten van het tweede experimenten suggereren dat menselijke waarneming om kan gaan met overvloedige informatie. We concluderen dat het menselijk waarnemingssysteem flexibel kan omgaan met een variërend aantal extracties. Kennis van klanken in een taal leidt bij mensen tot integratie van het deel van de features die samen tot een coherente waarneming leiden. Omdat de twee besproken typen computationele modellen voor menselijke spraakverwerking beiden tot inefficiëntie leiden wanneer het aantal input features toeneemt, maar mensen hier wel mee lijken te kunnen omgaan, stellen we een alternatief model voor. Dit alternatieve model kan de experimentele data van het huidige werk verklaren en wordt, in theorie, niet inefficiënt wanneer het aantal input features toeneemt. Het model is een "knowledge-driven" model, waar perceptuele verwachtingen worden gepreactiveerd door kennis. Wanneer verwachtingen overeenkomen met (delen van) de "signal-driven" input features kunnen de resterende, vaak irrelevant features onverwerkt blijven. Input die congruent is met de verwachtingen kan zo snel en efficiënt worden verwerkt, ook wanneer ruis-structuren leiden tot input features die niet bij de doel-spraak horen.

Het huidige werk verklaart de discrepantie tussen de algemeen gebruikte globale representaties en de minder goed bekende lokale representaties. Daarnaast geeft het een nieuw perspectief op de rol van kennis in spraakperceptie wanneer gebruik wordt gemaakt van lokale representaties.

(1) Lokale, signaal gedreven spraakrepresentaties zijn ruisrobuust en maken segmentatie van het spraaksignaal mogelijk waardoor context niet meer op dezelfde manier nodig is voor het disambigueren van de input. Dit betekent dat als dergelijke features worden geïntegreerd in systemen voor ASR, dat deze systemen veel flexibeler kunnen worden ingezet. Om dit te bereiken is het noodzakelijk dat systemen voor automatische spraakverwerking kunnen omgaan met een variabel aantal input features, wat in de huidige technologie nog niet het geval is. Lokale representaties van spraak, zoals in dit proefschrift beschreven, zijn een opening naar flexibeler toepasbaar ASR.

(2) Als gevolg van het veel gebruikte paradigma wat gericht is op de taalkundige verwerking van spraak verklaren spraakverwerkings-modellen data die is verkregen in gecontroleerde omgevingen, terwijl spraakherkennings-systemen moeten fungeren in niet gecontroleerde omgevingen. Deze discrepantie benadrukt het belang van perceptie-onderzoek in minder gecontroleerde settings. Het onderzoeksparadigma waarbij spraak in ruis wordt gezien als spraak met extra, irrelevante features biedt daarbij nieuwe mogelijkheden voor het begrijpen en modelleren van spraakverwerking.

Summary (summary in English)

The ultimate goal of Automatic spoken Key-word Spotting (AKS) as defined in the current work, is to recognise words without the strict reliance on contextual information. Modern systems (evaluated in e.g. Li et al., 2014) often rely on contextual information such as the sentence a word appears in, or the task a word is presented in. With this definition AKS differs from Automatic Speech Recognition (ASR) in the focus of AKS on recognition of individual words. The reliance on contextual information in ASR solves ambiguities but as a result it also hides the weaknesses of the computational approaches. Therefore, the limited reliance on context information in ASR elucidates two problems in speech processing.

The first problem is the segmenting of the acoustically continuous stream of speech sounds without destroying information by breaking target words into tiny parts. In current systems this is not considered a problem and input sound is cut into pieces, irrespective of the beginning and end of words. Sentence context allows for an estimation of the lost information of target-words. The second problem is the sensitivity to disturbances from non-target sound that lead to problems when context is not available to solve emerging ambiguities.

Both problems are associated with the poor representation of speech and the ignorance of knowledge-based expectancy in speech processing. These two factors are investigated further in the current work. We first concentrate on automatically extracted acoustical speech features with the goal to improve segmentation and robustness for speech representations. Evaluation of existing speech representations other than the commonly used MFCC features led to the conclusion that the local character of these representations helps to select high energetic, and thus noise robust, elements. Also, we concluded that the speech representations that are related to pronunciation characteristics (phonetic features) have the potential to segment the speech stream.

We present a speech representation method that incorporates the locality aspect and is directly related to pronunciation characteristics. The method is based on the selection of energetic components (ECs) from a harmonic complex (HC). We show that these extractions facilitate noise-robustness and segmentation. They represent the energetic segments of voiced speech components without the need of context information. The ECs are similar to glimpses (Cooke, 2006) where elements are selected that have a high local SNR relative to the noise. The ECs have the advantage that the noise characteristics do not need to be known before feature extraction. A disadvantage of the ECs is that not all information is retained when HCs are extracted from clean speech conditions. The ECs are attractive for the representation of speech because the loss of information in clean speech does not increase when mild noise is added, where the addition of mild

noise leads to unrecognisable input (when no context is used) when MFCCs are used to represent speech. Also, because ECs are noise robust and aid segmentation of the signal, the context of words or sentences is not necessary in the same way as it is for standard, global representations. A problem for the ECs is that the usefulness for speech recognition can not be explicitly tested with modern ASR techniques. The reason for this is that the number of extractions varies which does not fit current systems for ASR. The set of extractions varies as a result of noise, some extractions are missing while other extractions represent the noise instead of the speech. Because (Cooke, 2006) showed that the elements with a relatively high local SNR are strongly related to human perceptual elements we assume that humans are able to process a variable number of input features. We investigated human perceptual processes and the effect of knowledge based expectation on perception as a second factor to obtain a better understanding of speech processing in noise.

The second factor is the structuring character of knowledge and expectation on perception. The role of knowledge in ASR changes when speech is represented by local features. Because local features do not comply with the demands of existing back-end systems we investigate the structuring effect of knowledge in human perception of speech sounds. The evaluation of computational models for knowledge in human speech processing suggests that both "bottom-up" and "top-down" systems become increasingly inefficient when task irrelevant sounds lead to an increase of input features because all input features are processed. To obtain a better understanding of human processing of speech representations we performed two perceptual experiments. In both experiments we presented manipulated Dutch spoken vowels. In the first experiment we found that, after energetic suppression of one of the formant frequencies of the presented vowels, participants still perceived the auditory input as vowels, but the perceived identity changed for the articulatory rounded vowels (for example, /y/ was perceived as /i/). In the second experiment a formant was perceptually induced by presenting audio-visually incongruent vowels. In this experiment participants perceived only one vowel per presentation, where the strongest feature determined the percept. Additionally merging, similar to merging phenomena in the McGurk effect in the perception of consonants, took place for those instances where the combination of auditory place of articulation (for example, the height of /e/) and visual manner of articulation (for example the rounding of /y/) led to an existing vowel in the Dutch vowel system (the high, rounded vowel /ø/). The results suggest that human perception can handle information that represents only part of the frequency domain. We propose an alternative to existing models for human speech processing that explains the experimental data in the current work and does not necessarily become inefficient when the number of input features increases. This model is a knowledge driven model where perceptual expectations are pre-activated. When knowledge driven expectations comply to (part of) the signal driven input, irrelevant input can remain unprocessed. Input that is congruent with the expectations can be processed fast and efficiency of processing is retained also when noises lead to noise-related task-irrelevant input features.

The current work clarifies the discrepancy between the generally used global and the

less well known local representations and offers a new perspective on the role of knowledge in speech perception in noise when local representations are used. While models and theories on human speech recognition focus on data obtained in laboratory conditions to explain linguistic aspects of speech processing, systems for automatic speech recognition have to function in relatively uncontrolled conditions. The speech-in-noise paradigm, where speech in noise is represented as local features that can be both target and noise-related extends the field to less well controlled conditions. Our research suggests that local representations, by addressing the robustness and segmentation problems can lead to a more flexible application field for ASR where key-words can be detected in a variety of task settings and auditory conditions.

Author publications

- Lobanova, A., Spenader, J. and Valkenier, B. (2007). Lexical and perceptual grounding of a sound ontology. In: Matousek, V., and Mautner, P. (Eds.): *Text, Speech and Dialogue*, 4629, 180-187. Berlin: Springer-Verlag, doi:10.1007/978-3-540-74628-7_25
- Boone, M., De Vries, D., Andringa, T.C., Schlesinger, A., Van Dorp Schuitman, J., Valkenier, B., and Van de Vooren, H. (2008). Modelling of the cochlea response as a versatile tool for acoustic signal processing. *The Journal of the Acoustical Society of America* 123(5), 3722. doi:10.1121/1.2935188
- Valkenier, B., Krijnders, D.J., Van Elburg, R.A.J., and Andringa T.C. (2011). Psycho-acoustically motivated formant feature extraction. In *Proceedings of the 18th Nordic Conference of Computational Linguistics*, 11, 218-223.
- Schmidt, T.P., Wiering, M.A., van Rossum, A.C., Elburg, R.A.J., Andringa, T.C., and Valkenier, B. (2010). Robust real-time vowel classification with an echo state network. *Cognitive and Neural Models for Automated Processing of speech and text*.
- Valkenier, B. and Gilbers, D.G. (2013). The effect of formant manipulation on the perception of primary and secondary cardinal vowels. in Gooskens, G. and Bezooijen, R. van. (Eds.): *Phonetics in Europe: Perception and Production*, 303-315.
- Valkenier, B., Duyne, J.Y., Andringa, T.C., and Baskent, D. (2012). Audiovisual perception of congruent and incongruent Dutch front vowels. *Journal of Speech Language and Hearing Research* 55, 1788-1801. doi:10.1044/1092-4388(2012/11-0227)
- Hendriks, P., Van Rijn, H., and Valkenier, B. (2007). Learning to reason about speakers alternatives in sentence comprehension: A computational account. *Lingua*, 117 (11), 1879-1896. doi:10.1016/j.lingua.2006.11.008

