SUMMARIES OF PH.D. THESES DEFENDED IN 1995

ACOUSTIC AND LEXICAL VOWEL REDUCTION

author: Dick R. van Bergem
promotor: Louis C.W. Pols
co-promotor: Florien J. Koopmans-van Beinum
date of defence: 9 maart 1995

Summary

Speech is probably the most important means of human communication. In the present thesis we approach the phenomenon of speech from a functional point of view. That is, we believe that in general speakers strive to achieve optimal transfer of messages with a minimal amount of articulatory effort. Of course, there are bounds to the tendency for ease of articulation of a speaker. Recognition of the message will be harder, as speech is articulated more sloppily. On the other hand, listeners have several information sources at their disposal to help them in restoring acoustically fuzzy messages. Especially semantic and pragmatic knowledge sources may play an important role in this respect. A certain degree of sloppiness in the articulation can thus be tolerated. The phenomenon of vowel reduction, which is the subject of this thesis, is interpreted by us as a tendency towards ease of articulation. Actually, we investigate two types of vowel reduction: lexical reduction and acoustic reduction.

Lexical vowel reduction stems from linguistic research and is defined as the substitution of a full vowel with a schwa in specific words. An example of a Dutch word in which this phenomenon can occur is the word "miNUUT" (minute). The unstressed vowel in the first syllable of the word "miNUUT" (word stress indicated with capitals) can be realized as a full vowel /i/, but also as a schwa /ə/. In the latter case the schwa has become a characteristic (generally accepted) part of the word and therefore we call this phenomenon *lexical vowel reduction.*

Suppose a speaker has the *intention* to produce a full vowel. This does not necessarily mean that a neatly articulated full vowel will be realized. There appears to be an enormous variability in 'vowel quality'. Some vowels are pronounced much more carefully than others. Speakers are usually not aware of these differences in pronunciation, and listeners normally do not notice the variation. However, if several different vowels are segmented from their natural context and presented to listeners, it appears that some specimens are identified much better than others (assuming that speakers intend to produce phonologically 'correct' vowels). Another way to demonstrate differences in vowel quality is through a spectral analysis.

Vowel timbre is to a large extent determined by characteristic resonances in the vocal tract, which arise when the articulators (jaw, tongue, lips, etc.) take on specific positions. These resonances in the vocal tract are called *formants.* In the spectrum of the acoustic vowel signal peaks occur at the formant frequencies. Standard techniques are available to determine the location of these peaks in the vowel spectrum. Usually the formant frequencies of vowels are measured at just one specific point in a representative part of the vowel. Vowel quality is mainly determined by the first and

second formant frequency. The first formant frequency (F_1) roughly corresponds with the degree of jaw opening. The Dutch vowel /a:/ has the largest jaw opening and thus the largest F_1. The second formant frequency (F_2) roughly corresponds with the position of the tongue body. When the tongue body is front, as for /i/, a high F_2 emerges; when the tongue body is back, as for /u/, a low F_2 emerges. By plotting the first and the second formant frequency of Dutch vowels, a vowel triangle appears with the vowels /i/, /u/, and /a:/ at its corners.

In normal connected speech the articulators shift in a continuous motion from one sound production to another. The beginning of a vowel often shows characteristics of preceding sounds, whereas the end of the vowel can show characteristics of following sounds. If formant frequencies are measured throughout the vowel, *formant tracks* emerge that reflect the movements of the articulators. Formant tracks of vowels usually reveal a (parabolic) curvature; the degree of curvature of formant tracks depends per vowel on the (consonant) context in which the vowel is uttered. The maximum or minimum in the formant curves point to the 'ideal' target position of the vowel. This position in the formant tracks can be regarded as most representative for the vowel. The formant frequencies in this vowel part are called steady-state formant frequencies, because they vary little. In our study of vowel reduction we investigated changes in steady-state formant frequencies, as well as changes in the dynamics of formant tracks.

If steady-state formant frequencies of vowels from normal speech utterances are compared with formant frequencies of vowels pronounced in isolation (which can be regarded as 'ideal' vowels), it appears that all kinds of formant shifts have emerged. The formant frequencies of some vowels from normal speech are relatively close to their target position, whereas the formant frequencies of other vowels have shifted considerably away from their target position. These formant shifts can partly be ascribed to *coarticulation* effects: neighbouring sounds influence each other to a certain extent because of limitations of the articulators. However, the formant shifts are often much larger than would be expected on the basis of coarticulation effects alone. Such extra shifts emerge, for instance, when the vowel occurs in an unstressed syllable, or when a 'spontaneous' speaking style is used. In such cases we speak of *acoustic vowel reduction.*

The primary goal of the present thesis was to get a better insight in the phenomenon of vowel reduction through empirical research. Our investigation was restricted to the Dutch monophthongs /u, ɔ, o:, ɑ, a:, ɛ, ɪ, e:, i, y, œ, ø:/. The diphthongs /ɑu, ɛi, ʌy/ were excluded. Three main experiments were carried out. The first experiment, described in chapter 2, focused on the phenomenon of acoustic vowel reduction. In the second experiment, described in chapter 3, we extensively studied the schwa in order to better understand its role in acoustic and lexical vowel reduction. And the third experiment, described in chapter 4, focused on the phenomenon of lexical vowel reduction. These three experiments will be summarized below.

In the first experiment we systematically investigated the effect of sentence accent, word stress, and word class (function words versus content words) on acoustic vowel reduction. A list of sentences was read aloud by 15 male speakers. Each sentence contained one syllable of interest. This could be a monosyllabic function word, an unstressed syllable of a content word, or a stressed syllable of a content word. The same syllable occurred in all three conditions, as for instance in the series "kon" (/kɔn/, could), "konSULT" (/kɔnsœlt/, consultation), and "KONsul" (/kɔnsœl/, consul). In total, there were 33 triplets of words. We only wanted to study the vowel in these syllables (the /ɔ/ in the example above). All speakers first read the list of test sentences and afterwards they were asked to pronounce the test syllables ("kon" in the example above) in isolation; vowels from the latter syllables were defined as

'ideal' reference vowels. Sentence accent was manipulated with questions that preceded the test sentences.

Steady-state formant frequencies of all test vowels were measured and compared with the formant frequencies of the reference vowels. It appeared that the formant frequencies were closest to those of the reference positions for vowels from stressed syllables. The shift of formant frequencies was much larger for vowels from unstressed syllables, and largest for vowels from function words. Acoustic vowel reduction was stronger in non-accented words than in accented words. We also investigated how the dynamics of formant tracks changed under the influence of the factors mentioned above. It was found that formant tracks became flatter, as the steady-state formant frequencies were deviating more from the target positions, suggesting a decrease in articulatory gestures. And finally, a group of listeners was asked to identify the vowels (segmented from their context). As expected, the number of vowel confusions increased, as the acoustic reduction of vowels became stronger.

In the literature two different explanations for the phenomenon of acoustic vowel reduction are given: contextual assimilation (adaptation to consonant context) and centralization. The theory of centralization is based on the idea that the schwa is produced with a 'neutral' vocal tract (a straight lossless tube). According to acoustic theory, the formant frequencies of such a neutral vowel take a position in the centre of the vowel space. The neutral vocal tract can be interpreted as a resting position of the articulators. In order to achieve articulatory economy, people might have a tendency to approach this neutral position, when vowels are produced. This would result in centralization of formant frequencies. However, our experimental results showed that acoustic vowel reduction can be better interpreted as contextual assimilation, because for some consonantal contexts formant frequencies of vowels did not centralize at all. In our view centralization is a *by-product* of contextual assimilation. On the basis of a number of formant measurements on schwas that had been uttered by the 15 speakers in the test sentences, we strongly questioned the idea that the schwa is produced with a neutral vocal tract. This triggered the setup of a second experiment.

In this second experiment we systematically investigated the acoustic properties of the schwa in various consonantal contexts. Three male speakers read aloud a list of nonsense words of the form C_1əC_2V and of the form VC_1əC_2, in which the schwa occurred in an open syllable and in a closed syllable, respectively. In these nonsense words C_1 and C_2 could be any of the consonants /p, t, k, f, s, χ, m, n, ŋ, r, l, j, ʋ/ and V was taken from the vowel set /i, a:, u/. Consonants and vowels were systematically varied in all possible combinations. Formant frequencies, measured at the centre of the schwas, showed a considerable spread, especially F_2, dependent on the surrounding consonants C_1 and C_2 and the vowel V. This means that the schwa is strongly influenced by surrounding phonemes. We also studied the dynamics of F_2-tracks of the schwas. Normally, F_2-tracks of vowels show a certain degree of curvature, pointing to the F_2 target of the vowel. Our experimental results revealed, however, that the F_2-tracks of the schwas were hardly curved.

Based on the experimental results, the most important conclusion of this investigation was that the schwa is a vowel *without articulatory target* that is completely assimilated with its phonemic context. Thus, in our interpretation vowel reduction does not result in centralization, but in a shift of formant frequencies to a schwa position that can be almost anywhere in the vowel plane, dependent on the phonemic context. The widespread view that the schwa is produced with a neutral vocal tract is not very accurate. This view is based on assumptions about *static* articulatory positions, whereas in connected speech the *dynamics* of articulatory gestures have to be taken into account. According to us, a schwa production should be seen as the most economical (direct) movement from the preceding phoneme (consonant) to the following

phoneme (consonant).

The third experiment focused on the phenomenon of lexical vowel reduction. We asked 20 male speakers to read aloud a list of sentences. Each sentence contained one word of interest. The speakers were also asked to read these words without sentence context. In addition, the test words had to be named by them through the presentation of pictures (if possible). Each test word contained one specific vowel in unstressed position that we wanted to investigate. All test words from the conditions "words", "pictures", and "sentences" were aurally presented to a group of 20 listeners who were asked to identify the vowel of interest in each test word. The vowel responses of listeners were recoded into two broad categories: "full vowel" or "schwa". According to us, lexical vowel reduction could occur in part of the test words, for instance in the words "miNUUT" (/minyt/, minute), and "baNAAN" (/ba:na:n/, banana). Especially for such words we wanted to find out to what extent listeners were able to unambiguously identify the test vowel as either a full vowel or a schwa. In addition, we wanted to investigate the influence of the frequency of occurrence of words on lexical vowel reduction. For that purpose words like "baNAAN" and "miNUUT", that have a relatively high frequency of occurrence, were matched with words like "baNIER" (/ba:nir/, banner) and miNIEM (/minim/, marginal), that have a similar structure but a much lower frequency of occurrence. And finally, we wanted to investigate the influence of speaking style on lexical vowel reduction by comparing the conditions "words", "pictures", and "sentences".

The experimental results showed in the first place that listeners often disagreed about the classification of test vowels as either a full vowel or a schwa; it also appeared that the total number of schwa responses per speaker varied strongly. In the second place the number of schwa responses was much higher for vowels in words with a relatively high frequency of occurrence. And in the third place the number of vowel responses increased, as the speaking style became more casual. That is, the largest number of schwa responses occurred in the condition "sentences", followed by the condition "pictures", and the smallest number of schwa responses occurred in the condition "words". An acoustic analysis of the test vowels showed a strong relation with the perceptual results.

The final question that was posed in the present thesis was: What is the relation between acoustic and lexical vowel reduction? In our view these phenomena are two intermediate stages in the process of the sound change 'full vowel –> schwa'. A full vowel that is often subject to a strong acoustic reduction in a particular word may be confused with a schwa by listeners. In the next stage two variant forms of the word occur: a variant form with a full vowel and a variant form with a schwa (lexical vowel reduction). At a certain point in time the variant form with the schwa could become dominant and ultimately it could remain as the only accepted form of the word. Such a sound change appears to have taken place in the Dutch word "bETON" (/bətɔn/, concrete). This French loan word was originally pronounced with the full vowel /e:/ in the first syllable, but in modern Dutch the only accepted form of the word is the one with a schwa. Such a sound change can of course not occur in languages such as Italian and Japanese, that do not include the schwa in their phonological system.

Although the preconditions for the sound change 'full vowel –> schwa' in several Dutch words are excellent, the actual completion of the sound change is in our view to a large extent blocked by the rather close correspondence between Dutch vowel sounds and their orthographic representations. As soon as this orthographic barrier would be cleared, we expect a spreading of schwa substitutions throughout the entire Dutch lexicon. The growing influence of English on modern Dutch might speed up this process.

PERCEIVING DYNAMIC SPEECHLIKE SOUNDS
psycho-acoustics and speech perception

author: Astrid van Wieringen
promotor: Louis C.W. Pols
date of defence: 11 april 1995

Summary

Stop consonants in speech are cued by several properties including short and rapid vocalic transitions. These rapid transitions result from the changing configurations of the vocal tract and they are, therefore, more often produced than the release burst or the vocal murmur of the plosive. Despite the perceptual importance of these dynamic cues much less is known about the dynamics of speech stimuli than about stationary vowel sounds. This is partly because of the difficulty of examining perceptual cues of sounds which consist of three covarying dimensions, namely frequency (extent), duration, and rate of frequency change, and partly because other physical variables, such as amplitude or bandwidth, are difficult to control.

By increasing the complexity of the stimulus (from tone transitions to interpolated speech-based stimuli) and by varying the cognitive load of the task (from same/different discrimination to phoneme classification) a series of conditions are tested which seem to fit on a perceptual continuum. The experiments provide fundamental knowledge on the auditory and perceptual processing of glides, and several findings can be applied to speech synthesis and speech recognition.

Perceptual resolution varies with stimulus and task, but the global pattern of responses appears to be similar for single, complex, and interpolated speech-based stimuli, and possibly also for speech sounds. In the first part of the thesis auditory sensitivity is examined for different kinds of speechlike sounds (part I), in the second part the perceptual importance of the physical cues is examined in speech paradigms (part II).

Part I describes the (cues underlying the) minimal detectable changes in frequency, duration and rate of frequency change for short and rapid speechlike formant transitions. Despite the existing knowledge on the perception of dynamic sounds (chapter 2), discriminability of short and rapid tone and formant transitions had not been examined systematically for different kinds of speechlike stimuli. The data obtained from the psycho-acoustical experiments served as a basis for the following 'speech' experiments. With more limited stimuli it would be difficult to extend the psycho-acoustical results to speech perceptual research.

Just noticeable differences in endpoint frequency were determined for isolated transitions (chapter 3), as well as for transitions, which were preceded or followed by a stationary vowellike part (chapter 4). The global pattern of discrimination functions was similar for these stimuli in as far as difference limens in endpoint frequency decrease as the transitions become longer. As for temporal changes, discrimination was markedly affected by a change in transition duration, even when the total duration of the transition plus the steady-state remain constant (chapter 4).

The stimuli were not only made more speechlike by adding a stationary (vowel) part, but also by adding (fixed) formants and a short vocal murmur. Auditory sensitivity decreased as the stimulus became more complex. In order of sensitivity: tone sweeps (chapter 5) yielded the smallest difference limens in endpoint frequency, followed by isolated formant transitions (chapter 3), while single formant transitions yielded smaller difference limens in endpoint frequency than complex transitions (chapter 4). The difference in sensitivity is explained in terms of the difference in stimulus structure of tone sweeps versus formant glides, and in terms of masking effects caused by surrounding formants, as well as by the speechlike quality of the complex stimuli. The psycho-acoustical experiments suggest that discrimination is based more on endpoint frequency or on frequency extent, rather than on rate of frequency change. The discriminability for (consonant-vowel) CV-like transitions was worse than for (vowel-consonant) VC-like ones as a result of the steady-state following the transition (chapters 4 and 5). Detailed psycho-acoustical experiments with tone glides indicated that transients and recency effects may account for the fact that short and rapid initial transitions followed by a stationary part (CV-like) are not clearly heard as gliding tones, while final ones preceded by a steady-state (VC-like) are (chapter 5): it is easier to remember the varying (and probably discriminative) parts of the transition when they occur at the end of the information entering the ear than when they occur at the beginning.

The perceptual asymmetry was also analysed by a model of the inner ear. With this model the excitation patterns of the CV-like and VC-like stimuli showed some evidence that the perceptual asymmetry already occurs on a peripheral level of processing.

Part II deals with the perceptual importance of the psycho-acoustical cues in 'speech' tasks, i.e., the extent to which perceptual resolution varied with stimulus complexity and task. The perceptual experiments, which were based on the psycho-acoustical data determined in part I, showed that the perceptual importance of detailed acoustical cues depended on the complexity of the stimulus: the more speechlike the stimulus the more difficult it was to differentiate between the stimuli of the continua (due to masking effects and/or attentional constraints).

For instance, listeners matched similar endpoint frequencies and similar durations (despite rather different timbres), probably because the isolated transitions were too remote from speech to interfere with higher-order processes (chapter 6). Similarly, some of the CV-like and VC-like formant stimuli were also perceived analytically in speech tasks. The perceptual results of those stimuli which were perceived analytically, clearly reflected the initial-final perceptual asymmetry that was highly significant psycho-acoustically. As the stimuli became more complex such effects were less noticeable, because perception was increasingly affected by, for example, partial masking of acoustical cues of the speechlikeness of the sound.

In the absolute identification paradigm we measured the listener's optimal labelling ability for the different kinds of speechlike stimuli (chapter 7) and we found that the number of identifiable categories did not decrease with increasing stimulus complexity. Most of the pairs of stimuli were identifiable from each other and the perceptual processes reflected general sensory restrictions (due to masking, attention or the speechlike quality of the sound). Neither these experiments nor the ABX discrimination experiments (chapter 8) showed clear evidence that perceptual resolution was determined by a speech-specific labelling mechanism based on long-term linguistic experience: In our study all the stimuli, including the interpolated speech-based ones, were discriminated better than predicted from the 2-AFC classification tasks,

suggesting that listeners make use of additional acoustical cues (chapter 8). If a phoneme labelling mechanism interferes in the processing of the complex or speech-based stimuli, discrimination would be at chance level for stimuli which are classified similarly, while sounds which are classified differently would be highly discriminable.

We conclude that, in our study, the perception of vocalic transitions is, for a large part, based on general auditory properties and that it is not limited by a speech-specific mechanism based on long-term linguistic experience. The speechlike stimuli in the absolute identification and ABX discrimination tasks are probably stored temporarily. That is to say, listeners create internal representations of the stimuli, and each of the incoming sounds are compared to the perceptual anchors of that continuum in memory. It is possible that the quality of the stimuli is not natural enough to be compared to language-specific anchors in memory.

Plosive identification from initial and final natural speech transitions did not reflect the perceptual asymmetry found in the experiments with formant stimuli. Natural Dutch CV and VC transitions, excised from both CVC and VCV syllables, were perceptually equally strong (chapter 9) for both Dutch and American-English subjects, possibly because the speech signal contains redundant cues for plosive identification. In natural speech the VC transition may be perceptually more salient than the CV one whenever the release burst is not produced. However, the original and time-reversed transitions in this 6-AFC classification experiment did not reflect the perceptual asymmetry in a carefully balanced classification task. In short, our experiments showed that the perceptual importance of the psycho-acoustical cues decreased as the stimuli and task became more complex. In speech communication, listeners (fortunately) do not need to perceive all these detailed acoustical cues. However, if the characteristics of the signal are not masked, listeners can try to zoom in on certain levels of processing and discriminate ambiguous or new (e.g., foreign language) cues. Our study illustrates that many perceptual aspects can be explained by general auditory and cognitive properties (chapter 10). Further study is necessary to examine the origins of perception in detail and to understand how higher-order phenomena, such as knowledge and expectations, influence the perception of vocalic transitions in speech.