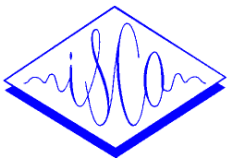


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

INFLUENCE OF DURATION ON STATIC AND DYNAMIC PROPERTIES OF GERMAN VOWELS IN SPONTANEOUS SPEECH*

Britta Wrede, Gernot A. Fink, Gerhard Sagerer
Bielefeld University, Faculty of Technology, 33594 Bielefeld, Germany
e-Mail: {bwrede, gernot, sagerer}@techfak.uni-bielefeld.de

ABSTRACT

Changes in speech rate severely affect the performance of continuous speech recognition systems. In order to better understand the underlying effects of speech rate changes an analysis was carried out on the influence of duration on the spectral properties of vowels in a large German corpus of spontaneous speech. The results show a strong centralisation effect of the vowel formant frequencies due to shorter duration while the formant movements are only slightly affected. The data suggest that the movement velocity is not changed in vowels with a limited duration. As the means of the on- and offset frequencies also remain stable only the middle part of the vowels are affected by the centralisation effect. These results are discussed in the light of the modelling of varying speech rate in automatic speech recognition systems.

1. INTRODUCTION

In automatic speech recognition (ASR) the variability of speech rate is a major source of recognition errors. Therefore, different approaches have been taken to account for variability in speech rate (e.g. [8], [10], [12]). These approaches usually consist of a preliminary estimation of the speech rate which is assumed to be constant over the whole utterance. During the recognition phase rate specific models are selected according to the estimated speech rate. These rate specific models are trained with samples of a specific rate class and thus contain implicit information about the spectral properties of speech at this rate. However, no use is made of the regularities of spectral changes that co-occur with changes in speech rate and duration. If such regularities can be extracted from the speech data they can enhance the training of rate specific models or even replace them by using transformations of the speech signal without classifying it into a discrete speech rate.

A well known phenomenon occurring with vowels in fast speech is that of *vowel reduction*. The formant frequencies of the German cardinal vowels /a:/ /i:/ and /u:/ describe a vowel triangle in the space spanned by the first and second formants F1 and F2 (cf. Fig. 1). If these vowels are reduced due to shorter duration the formant frequencies tend to become more centralised within this triangle. Shorter duration is not only a result of faster speech but also of unstressed speech. Therefore, many experiments on reduction are based on tempo as well as on stress.

In [4] tempo and stress, both, had a strong influence on the vowel space area. The area was largest with slow stressed vowels and smallest with fast unstressed vowels. Van Bergem [11] confirmed in a listening test the perceptual significance of the vowel space area. The larger the vowel space area the less recognition errors are made by human listeners.

Another factor found to influence the formant frequencies of vowels is the speaking style. Moon et al. compared clear speech such as used in noisy environments to normal speech [9]. Again, the experiments showed a centralisation effect due to shorter duration. However, the reduction was more limited for clear speech where an increase in rate of formant change compensated for the time compression and thus preserved the absolute formant values.

Only few experiments are concerned with the effects of duration on the formant movements. A series of experiments reported by Gay in [5] and [6] indicate that the rate of formant movement is not influenced by a change of speaking rate.

However, most of these experiments have been carried out on read speech that was explicitly recorded for the analyses. It seems possible that the different instructions given to the speakers elicited different speaking styles with different degrees of reduction and formant movement changes. For more general predications of the influence of tempo on the spectral characteristics it is desirable to perform experiments on the basis of more natural speech. Therefore, an analysis was carried out on a large corpus of spontaneous speech from many different talkers.

* This research was partly supported by a graduate grant of the German Research Foundation (DFG) within the Graduate Program 'Task-oriented Communication'.

2. CORPUS

The basis of the experiment constituted the large German corpus of spontaneous speech recorded in the Verbmobil project [7]. It consists of dialogues with two participants negotiating one or more appointments in accordance with their diaries. Although the scenario is an artificial setup for the purpose of speech recording no instructions concerning the speaking style were given to the subjects. Therefore the corpus contains the usual features known to occur in spontaneous speech such as hesitations, corrections, slips of the tongue, high variability in speech rate etc. In summary, it consists of 13,910 utterances from 654 speakers from different regions of Germany and contains 303,746 tokens of 6,258 different words.

The recordings were performed in an ordinary office environment with close-talking microphones and with a sampling-rate of 16 kHz.

3. METHODS

The segmentation of the speech signal was done automatically with the speech recognition system as described in [3] using the official orthographic transcription of the Verbmobil project. The signal processing of the speech recogniser computes 12 mel-frequency cepstral coefficients [2] and their smoothed first and second order derivatives. By adding one energy coefficient this results in a 39-dimensional feature vector which is computed every 10 ms on a 16 ms frame of speech. The derivatives are calculated over a series of five frames.

The formant analysis was performed automatically with the formant tracker of the ESPS software[1]. The formant tracker used 20th order LPC-coefficients that were computed every 10 ms on a 16 ms frame of speech. The tracker was instructed to scan the spectrum for five formants from which only two were used for further analysis.

In accordance with the segmentation of the speech signal the following values were computed for the vowels /a:/ /i:/ and /u:/:

1. The average first and second formants over the whole cosine-weighted vowel segment
2. The first and second formants at the first, middle and last frame of the vowel segment
3. The first derivatives of the first and second formants at the first, middle and last frame of the segment
4. The duration of the segment

The first order derivatives of the formant frequency values were computed as a first order regression over the preceding two, the actual and the following two formant values of the intended frame. As the speech recogniser performs a frame based segmentation of the speech signal boundaries can only occur at the end of a frame leading to segment durations that are multiples of 10 ms.

[1]The Entropic Signal Processing System (ESPS) is a commercial software package distributed by Entropic Research Laboratory, Inc.

4. RESULTS

The segmentation produced 38,168 instances of /a:/, 23,464 instances of /i:/ and 7,927 instances of /u:/. For the following analyses the samples of each vowel were divided independently into four duration classes. Each of the classes contained approximately 25% of the data of one vowel.

4.1. Static formant frequencies

For a general picture of the behaviour of the formant frequencies with different durations the means of the average first and second formants over the whole weighted segment were computed according to their duration class. The means of the first two formant frequencies plotted against each other are shown in Fig. 1. The vowel triangle shows a clear trend of centralisation for shorter vowels.

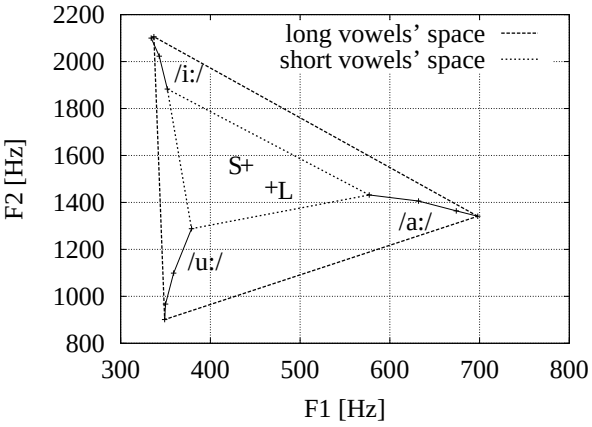


Figure 1: Means of formant frequencies by duration classes of the German vowels /a:/ /i:/ /u:/. Only the longest and shortest vowels are shown. L shows the vowel space centre of the longest, S of the shortest vowels.

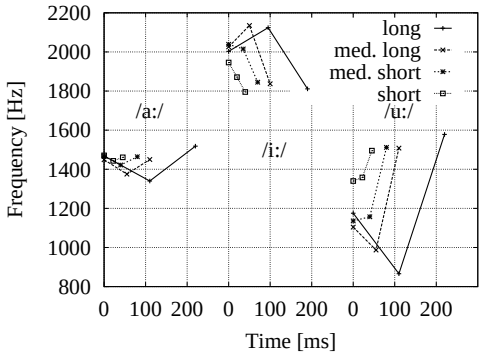
For the statistic analysis of this trend the centres of the vowel triangles of each duration class were computed. As the formant frequencies for each vowel and duration class are approximately normally distributed the computation of the distances was carried out on the mean values of each class instead of the distances for each token. By this procedure a too strong influence of the variances is avoided which would occur if the distances were computed item by item. As the variances represent the high variability of the speech sample caused by the large amount of different male and female speakers with individual vowel spaces the

comparison of means was chosen as an abstraction over the speaker specificities.

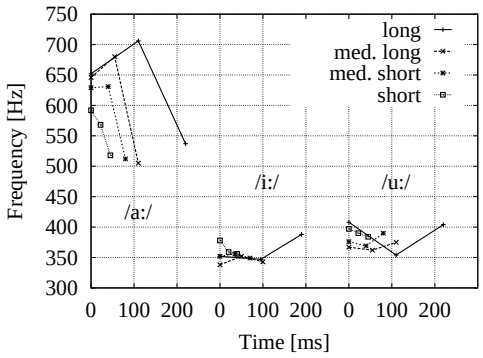
The distances of the mean values of the three vowels to the duration specific centres were computed according to their duration class. While the mean distance of all long vowels to the corresponding centre was 485 Hz this distance was reduced to 239 Hz for the shortest vowels which represents a reduction of 51% of the distances and 75% of the vowel space area.

4.2. Formant frequency movements

For the analysis of the changes in the formant movements the means of the formant frequencies measured at the beginning, middle and end of the segments were computed. The results are shown in Fig. 2.



(a) Movements of second formants



(b) Movements of first formants

Figure 2: Movements of the second (a) and first (b) formants of the German vowels /a:/ /i:/ /u:/ as measured at the begin, middle and end of the vowels. The duration is chosen according to the mean duration values of the corresponding duration classes.

As can be seen from Fig. 2 while the on- and offset formant frequencies almost remain invariant the target values in the middle of the segments are clearly reduced for the shorter segments. This indicates that the total amount of reduction as seen in Fig. 1 is mainly due to the middle segment. The steepness of the connections of the measured frequencies as shown in Fig. 2 suggest that the rate of change stays the same over the different duration classes. Indeed, the first derivatives computed at the three different points of the vowel segments show that the steepness of the formant movements at least for the middle and last frame are roughly the same for the different duration classes. However, the values for the first derivatives of the second formants of the initial frame of all three vowels indicate a flatter movement.

5. DISCUSSION

The presented data confirms the general tendency of vowel reduction due to shorter duration in spontaneous speech over a large number of speakers. Since the direction of the vowel reduction points towards the centre of the vowel space this centre becomes an important reference point for speech recognition. It is generally assumed that this centre is speaker specific and represents the spectral characteristics of the neutral vocal tract of a speaker. Consequently the speaker specificity increases with shorter segmental duration as the formant frequencies tend to become more and more neutral.

However, formants proved not to be successful features in signal processing ([11](p.6)). Therefore, to account for this reduction of the vowel space area in ASR the behaviour of the whole spectrum has to be known. Since the formant frequencies are defined as energy peaks in the spectrum ([1](p.221)) a tendency for centralisation means that the energy of the spectrum is shifted towards the points of the central formant frequencies. Thus, to compensate for vowel reduction a non-linear transformation that shifts the spectral energy in the opposite direction is needed.

Furthermore, these results can be used for the estimation of speech rate or reduction. If most of the spectral energy is distributed near the centre frequencies it can be assumed that the speech segments are rather reduced, thus spoken fast or unstressed. Additionally, a measure of the spectral variability over time can give information about the velocity of movements and thus about the estimated rate of speech when a clear speaking style is assumed.

The relative constancy of formant movements stands in contradiction to results from the modelling of speech rate as reported in section one. Martínez et al. [8] found that the features that were most affected by different speech rates were the dynamic features which consist of the first and second order derivatives of the cepstral coefficients. Their results suggest that the velocity of formant movement changed significantly with increasing speech rate which mirrors the results of Moon et al. [9] for clear speech. As the corpus of Martínez et al. consisted of speech data recorded with the purpose of gathering data of different speech rates the instructions given to the subjects might

have elicited a clear speaking style which might thus have caused the different results. The speakers of the corpus of the current experiment seem to use an informal speaking style where less effort is made to achieve the targets of the intended vowels with short duration. Therefore, a heavy centralisation effect as shown in Fig. 1 occurs because the articulatory movements are not speeded up as can be seen from the almost stable formant frequency movements in the short and long duration classes (cf. Fig. 2). However, the on- and offset frequencies remain stable indicating that only the middle part of the vowel is reduced and not the whole segment as reported by van Bergem [11]. From these results it can be assumed that the dynamic features in ASR will only be affected to a small degree by varying speech rates if no normalisation of the spectral centralisation is performed. If a compensation of the spectral reduction of the formant frequencies is done for example by a non-linear transformation as suggested above the spectral change should become faster which is mirrored in higher values of the derivatives.

6. SUMMARY

The variability of speech rate especially in spontaneous speech causes severe problems for ASR systems. Current approaches to model these changes do not take the regularities of the co-occurring spectral changes into account. However, experiments concerning the influence of prosodic features such as speech rate and stress often are performed on small corpora of read speech in very restricted situations. Therefore, an analysis was carried out on a large corpus of spontaneous speech. The results showed a clear centralisation effect of the formant frequencies of the German cardinal vowels /a:/ /i:/ /u:/ when uttered with shorter duration. This centralisation effect could be measured by means of a decreasing distance to a duration specific centre of the vowel space which leads to a reduction of 75% of the vowel space area. A more detailed analysis of different parts of the vowel segments showed that while the on- and offset frequencies almost remained stable over the different duration classes the middle part was heavily affected by a centralisation tendency. The derivatives as measure for the steepness of the formant movements indicated that no systematic change in the velocity of the articulatory movements took place due to shorter duration. A comparison of experiments reported in the literature suggests that in the analysed spontaneous speech a more informal speaking style is chosen where no effort is taken to compensate for shorter time by faster movements to obtain the target frequencies of the intended vowel.

The results show that the influence of speaking rate on the spectral properties of vowels is very regular and therfore predictable. Thus, the centralisation effect might be compensated by a non-linear transformation of the whole spectrum depending on the measured extent of reduction. Therfore, a continuous adaptive normalisation of speaking rate is possible.

7. REFERENCES

1. J. Clark and C. Yallop. *An introduction to Phonetics & Phonology*. Blackwell, Oxford, 1990.
2. S. Davis and P. Mermelstein. Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.
3. G. A. Fink. Developing HMM-based recognizers with ESMERALDA. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Berlin Heidelberg, 1999. Springer.
4. M. Fourakis. Tempo, stress, and vowel reduction in american english. *Journal of the Acoustical Society of America*, 90(4):1816–1827, 1991.
5. T. Gay. Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America*, 44:1570–1573, 1968.
6. T. Gay. Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America*, 63:223–230, 1978.
7. K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, and W. Thon. Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL – 3.0. Technical Report 11, Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel, 1994.
8. F. Martínez, D. Tapias, and J. Álvarez. Towards speech rate independence in large vocabulary continuous speech recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 725–728, 1998.
9. S.-J. Moon and B. Lindblom. Interaction between duration, context, and speaking style in english stressed vowels. *Journal of the Acoustical Society of America*, 96(1):40–55, 1994.
10. N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proc. European Conf. on Speech Communication and Technology*, pages 2079–2082, 1997.
11. D. R. van Bergem. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12:1–23, 1993.
12. J. Verhasselt and J. Martens. A fast and reliable rate of speech detector. In *International Conference on Spoken Language Processing*, pages 2258–2261, 1996.