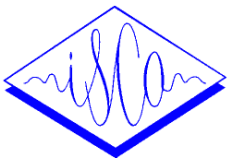

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

STRATEGIES OF VOWEL REDUCTION – A SPEAKER-DEPENDENT PHENOMENON

Christina Widera

Institut für Kommunikationsforschung und Phonetik (IKP), University of Bonn
Poppelsdorfer Allee 47, 53115 Bonn, Germany
e-mail: widera@ikp.uni-bonn.de

ABSTRACT

In natural speech, a lot of inter- and intra-subject variation in the realisation of vowels is found. Perception experiments show that listeners can discriminate speaker-independent levels of vowel reduction. The question is whether all speakers used the same acoustic cues for signalling reduction levels. The acoustic realisation of reduction levels of the vowels [u:], [i:], and [a:] of three speakers were tested for their separability and their predictability using linear discriminant analysis, support vector machines, and artificial neural networks. The results indicate a speaker-dependent realisation of vowel reduction.

1. INTRODUCTION

Investigations of vowel reduction are well established in the literature. Variation in vowel realisations is not only found between different speakers, but also in different utterances of the same speaker. Vowels spoken in isolation or in a neutral context are considered to be ideal vowel realisations. Vowels differing from these ideal vowels are regarded to be reduced. Acoustically, vowel reduction is described by smaller spectral distances between the sounds. Vowel reduction frequently coincides with shorter segmental duration [1], [2].

In German, tense vowels are opposed to lax vowels (e.g. [i:] vs. [ɪ], [u:] vs. [ʊ]). Tense and lax vowels mainly differ in quality. However, the difference between the tense and lax vowel /a/ is mostly due to quantity ([a:] vs. [a]; [3]). Vowel reduction is described by a change in quality. Depending on the strength of reduction, reduced tense vowels sound similar to their lax counterparts and strongly reduced ones are similar to [ə].

In this investigation we focus on the acoustic realisation of vowel reduction of different speakers. Here, the strength of vowel reduction is described by levels. The reduction levels of the tense vowels [i:], [a:], and [u:] are analysed according their separability and predictability.

2. DATABASE

The database (“Bonner Prosodische Datenbank”, [4]) consists of isolated sentences, question and answer pairs, and short stories read by three speakers (two female, one male). One female speaker can be regarded as a professional one. All three subjects are native speakers of Standard German. The utterances were labelled manually on the segmental level (SAMPA [5]). Each vowel was annotated with following acoustic features: duration, fundamental frequency (F0), relative intensity in four frequency bands (0-1 kHz, 1-2 kHz, 2-4 kHz,4-8 kHz), and the first three formant frequencies (F1, F2, F3). For each vowel, the frequencies of the first three formants were computed every 5 ms [6]. The values of each formant were estimated by a third order polynominal function fitted to the formant trajectory. The formant frequency of a vowel is defined here as the value in the middle of that vowel [7]. Additionally, the Hz-scaled formant values were converted to the mel-scale. All values of all vowels within one phoneme class were standardised so that their mean value becomes 0 and their deviation 1 (z-scores). Furthermore, some vowels were labelled with respect to their strength of reduction. The strength of reduction is described by discrete levels. The method is outlined in the following section (for a detailed description see [8]).

3. PERCEIVED REDUCTION LEVELS

The initial hypothesis is that clustering of the vowels would indicate potential prototypes of reduction levels. Since F1 and F2 are assumed to be the main factors determining vowel quality [9], prototypes of reduction levels are established by clustering of F1 and F2 values (mel-scaled and standardised) of vowels for each phoneme class. The test material was taken from the utterances of the professional speaker (Speaker1). On the basis of a pre-test, these vowels are grouped in eight clusters (mean cluster analysis). For each cluster, one prototype was determined whose formant values are closest to the cluster centre.

The hypothesis was tested for each phoneme class, separately. Subjects were asked to arrange the prototypes by strength of reduction from unreduced (reduction level 1) to reduced. Then, the relevance of the prototypes for reduction levels was examined by assigning further vowels to these prototypes according to their qualitative similarity. The results show that not all prototypes can be regarded as representative of reduction levels. These prototypes were excluded and the remaining prototypes were evaluated by further experiments. The outcome is that listeners reliably discriminate three reduction levels for [a:], four levels for [i:], and five reduction levels for [u:]. The agreement between the subjects is above 70%.

Next, it was investigated whether the reduction levels and their prototypes can be transferred to other speakers. Subjects had to classify vowels of two other speakers (the naive female and naive male speaker) according their qualitative similarity to the

prototypes. The analyses show that subjects compensate for speaker differences. The agreement between listeners is comparable to that in the previous experiments.

4. ACOUSTIC REALISATION OF REDUCTION LEVELS

The perception experiments showed that subject can reliably describe vowel reduction by discrete levels. The number of levels depends on the vowel. For [i:] four reduction levels, for [a:] three levels, and for [u:] five reduction levels were found.

Now, we address the question of the acoustic realisation of reduction levels by different speakers. The acoustic realisation of the reduction levels are tested according their separability and predictability. Linear discriminant analyses (LDAs) and support vector machines (SVMs) were used for examining the separability of reduction levels.

LDA describes correlation between dependent (acoustic features) and independent variables (reduction levels) with linear functions. The maximal separation between groups is specified by the weight-coefficients of the function [10]. The objective of SVMs is to find optimal hyperplanes that correctly classify as much data as possible and to maximise the distance of these hyperplanes. Optimal hyperplanes are found by the Structural Risk Minimisation Principle. A hyperplane is a linear separator, however SVMs can be used for separating non-linear distributed data by mapping them into a higher-dimensional space [11]. The reduction levels were divided into two classes. One group consists of vowels of a certain reduction class, the other one contains the vowels of the remaining levels. The SVMs were trained per reduction level. For LDAs as well as SVMs, the training set is equal to the test set.

The predictability of reduction levels was additionally investigated with artificial neural networks (NNs). Feed-forward NNs with two hidden layers were taken. A backpropagation learning algorithm was used (supervised learning). NNs are trained by repeated presentations of pairs consisting of input- and output-patterns. During the training phase, NNs try to find common regularities of the input patterns for the discrimination of output-patterns [12]. The performance of the NNs was tested by five-fold cross-validation. Output of the NNs was the pertinent reduction level.

All classifiers were trained with the data of all speakers (complete data set) and with the data of each speaker separately (speaker-specific data set). For [u:] and for [i:] the training sets contain 60 vowels per speaker. For [a:] 50 vowels per speaker were used.

The input of all classifiers consists of fundamental frequency, formant frequencies, the mean energy of the four frequency bands, and duration (standardised by z-scores).

4.1. Separability of reduction levels

The LDA classifications (Figure 1) show that for [a:] the accuracy hardly differs depending on the data set. It is over 70%. The classifications are comparable to the agreement between human judgements.

In contrast to [a:], there is a obvious decrease of accuracy with respect to the different data sets for the vowels [i:] and [u:]. The classification rates of the data of the two female speakers (Speaker1 and Speaker3) are above 70%. However, the classification of each reduction level is better for Speaker1 than for Speaker3.

The functions of LDAs indicate that for [a:] the main acoustic correlate of the reduction levels is duration (complete data set: cc=-0.907, Speaker1: cc=0.790, Speaker2: cc=0.841, Speaker3: cc=0.825). Most of the variance is explained by these functions (complete data set: 85.0%, Speaker1: 68.0%, Speaker2: 90.3%, Speaker3: 87,6%). For [i:] and [u:] the main correlates of reduction levels differ depending on the data set. The classification of [i:] is mainly determined (cc>=0.7, except for the data set of Speaker1 cc>0.5) by energy of the second frequency band (complete data set, Speaker1, Speaker3) and F2 (Speaker1, Speaker2) More than 70% of variance is explained (complete data set: 77.4%, Speaker1: 83.0%, Speaker2: 83.9%, Speaker3: 71.6%). The classification of [u:] bases on energy of the first (complete data set, Speaker1, Speaker2) and the second frequency bands (Speaker1, Speaker2), duration (complete data set, Speaker1, Speaker2) and F2 (Speaker3). The values of the explanation of variance (complete data set: 66.6%, Speaker1: 52.8%, Speaker2: 61.8%, Speaker3: 60.9%) indicate that the classification of [u:] is more complex than those of the other two vowels with less reduction levels.

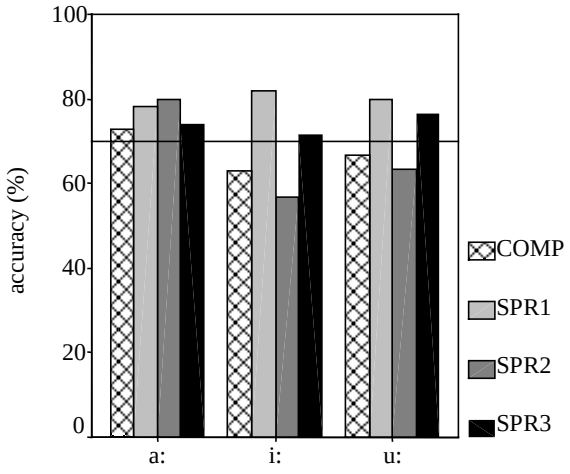

Figure 1: Accuracy (%) of LDAs depending on the data set (COMP: complete data set; SPR1: data set of Speaker1; SPR2: data set of Speaker2; SPR3: data set of Speaker3) and the vowel (a:, i:, u:). The line marks the agreement between human judges.

In contrast to the classifications by LDAs, the classifications by SVMs are based on two classes. Here, we used recall instead of accuracy, since the number of vowels with a certain reduction level is always far lower than the number of vowels with the other levels. Thus the accuracy does not say anything about the performance of SVMs to classify one reduction level.

$$\text{recall} = \frac{\text{no. of vowels with reduction level R correct}}{\text{no. of vowels with reduction level R}}$$

Compared to the classification of the speaker-specific data sets of the vowels [a:] and [u:], the recall of SVMs decreases in the complete data sets (Table 1). For the speaker-specific data sets of [a:] the average recall is over 70% for all speakers and comparable to the results of the LDAs. For the vowels [i:] and [u:] similar results are only found for Speaker1.

vowel	data set	Red1	Red2	Red3	Red4	Red5
a:	COMP	**76.1**	43.5	**74.1**		
a:	SPR1	**84.6**	**90.9**	**73.1**		
a:	SPR2	**100.0**	**81.3**	**71.8**		
a:	SPR3	**90.0**	63.2	**72.7**		
i:	COMP	68.3	68.4	0.0	46.9	
i:	SPR1	**86.7**	**78.3**	**90.0**	**100.0**	
i:	SPR2	**78.6**	15.4	0.0	0.0	
i:	SPR3	0.0	**80.0**	0.0	33.0	
u:	COMP	**81.9**	0.0	50.0	50.0	0.0
u:	SPR1	**95.8**	53.9	**80.0**	**87.5**	**100.0**
u:	SPR2	**87.5**	20.0	14.3	28.6	**100.0**
u:	SPR3	**83.3**	38.9	**100.0**	**77.8**	50.0

Table 1: Recall (%) of SVMs for each reduction level (RED: reduction level) depending on the data set (COMP: complete data set; SPR1: data set of Speaker1; SPR2: data set of Speaker2; SPR3: data set of Speaker3) and the vowel (a:, i:, u:).

The results of both LDA and SVM show that the acoustic realisation of the reduction levels of the vowel [a:] is quite similar between the speakers. Furthermore, the LDAs indicate that the reduction levels are mainly influenced by duration. This supports the claim that in German the realisation of the tense vowel [a:] and its lax counterpart differs in quantity.

For the vowels [i:] and [u:], only the classification of data set of Speaker1 is comparable to the agreement of human judgements. The lower accuracy and recall for the complete data sets might account for a different acoustic realisation of reduction levels. The distribution of the acoustic parameters could be one possible explanation for the classification results of the data taken from Speaker2 and of Speaker3. However, ANOVA finds no significant difference between the mean values of the speaker-specific data sets. According to the variance, significant difference are only found for F2 of [a:] (Levene-test=3.829, df=2, $\alpha<0.04$), for F2 and for energy in the second frequency band of [i:] (Levene-test=11.660, df=2, $\alpha<0.01$ and Levene-test=3.113, df=2, $\alpha<0.05$, respectively), and for energy in the fourth frequency band of [u:] (Levene-test=8.122, df=2, $\alpha<0.01$).

The sums of absolute differences between the mean values of each parameter for each reduction level indicate that the reduction levels of Speaker2 and of Speaker3 are closer distributed in the acoustic space than those of Speaker1 (c.f. Figure 2). The sums of the standard deviations show that the acoustic parameters of the vowels [i:] and of [u:] of Speaker2 and Speaker3 vary hardly more than those of Speaker1 (Figure 3).

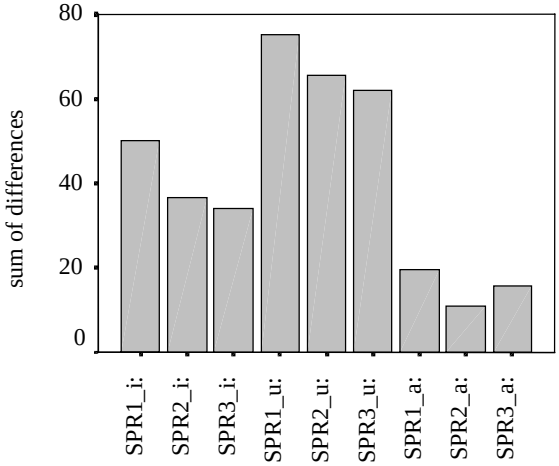

Figure 2: Sum of absolute differences between the mean values of each parameter for each reduction level depending on the data set (SPR1: data set of Speaker1; SPR2: data set of Speaker2; SPR3: data set of Speaker3) and the vowel (a:, i:, u:).

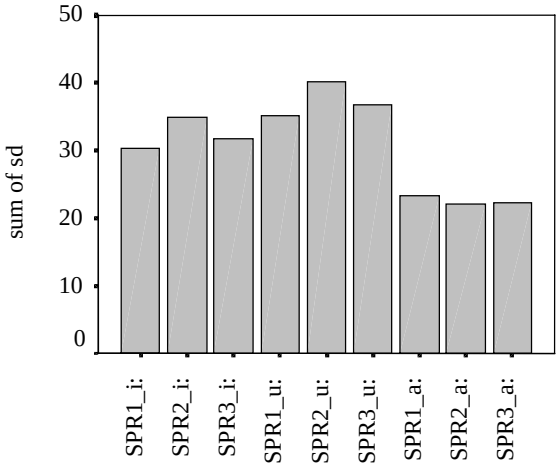

Figure 3: Sum of standard deviation for each parameter of each reduction level depending on the data set (SPR1: data set of Speaker1; SPR2: data set of Speaker2; SPR3: data set of Speaker3) and the vowel (a:, i:, u:).

4.2. Predictability of reduction levels

Using NNs, the predictability of reduction levels on the basis of the acoustic parameters were tested. Only for the data set of Speaker1 the mean accuracy of classification is over 70% (five-fold cross-validation). For all vowels higher classification rates of the speaker-specific data sets are found in contrast to the complete data sets (Figure 4).

The higher accuracy of the speaker-specific data sets indicates inter-speaker differences with respect to acoustic realisations of reduction levels. However, the question comes up why NNs fail to classify the data sets of Speaker2 and Speaker3. One possible reason is that the data of these speakers are noisy, another one is that the NNs are unable to find regularities in the patterns for the description of reduction levels of Speaker2 and Speaker3. However, ANOVA only shows slight differences between the speaker-specific data sets according to the variances (c.f. section 4.1). Therefore, we suggest that for Speaker2 and Speaker3 the results of NNs base on intra-subject differences according to the acoustic realisation of reduction levels. This is also supported by the results of LDAs and SVMs. The reduction levels of the professional speaker (Speaker1) seem to be signalled more consistently than those of the naive speakers.

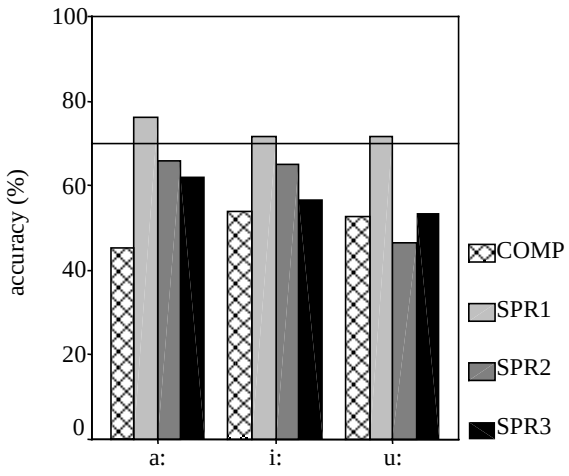

Figure 4: Mean accuracy (%) of NNs depending on the data set (COMP: complete data set; SPR1: data set of Speaker1; SPR2: data set of Speaker2; SPR3: data set of Speaker3) and the vowel (a:, i:, u:). The line marks the agreement between human judges.

5. CONCLUSION

The aim was to investigate the acoustic realisation of reduction levels of the vowel [u:], [i:], and [a:] of three speakers. LDAs and SVMs were used for examining the separability of reduction levels. Predictability of reduction levels was additionally examined by NNs.

For the vowel [a:], LDA performs as well as human judges. However, we find a better performance on speaker-specific data sets than on the complete data set. For [u:] and [i:], the results of LDAs are comparable to human judgements for two speakers. When trained on the speaker-specific set, the SVMs perform as well as LDAs. For NNs, only the performance of those trained with the data set from the professional speaker is comparable to the agreement between listeners.

The results indicate, that the realisation of vowel reduction is speaker-dependent. Furthermore, it seems that the professional speaker used a more consistent strategy for signalling reduction levels.

6. ACKNOWLEDGEMENTS

I would like to thank Maria Wolters for fruitful discussions and her helpful advice concerning support vector machines. For support vector machines SVMlight [13] was used. This work was funded by the Deutsche Forschungsgemeinschaft (DFG) under grant HE 1019/9-1.

7. REFERENCES

[1] Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35, 1773-1781.

[2] Lindblom, B.; Brownlee, B.D. and Moon, S.-J. (1992): Speech transforms. *Speech Communication*, 11, 357-368.

[3] Kohler, K. J. (1995). *Einführung in die Phonetik des Deutschen* (2nd ed), Berlin: Erich Schmidt Verlag.

[4] Heuft, B.; Portele, T.; Höfer, F.; Krämer, J. Meyer, H.; Rauth, M. and Sonntag, G. (1995). Parametric description of F0-contours in a prosodic database. *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 2, 378-381, Stockholm: KTH.

[5] http://www.phon.ucl.uk/home/sampa/german.htm.

[6] ESPS (Version 5.0), Entropic Research Laboratory.

[7] Stöber, K.-H. (1997). Unpublished software.

[8] Widera, C. and Portele, T. (1999). Levels of reduction for German tense vowels. *Proceedings of Eurospeech*, 4, 1695-1698.

[9] Pols, L. C. W.; van der Kamp, L. J. T. and Plomp, R. (1969). Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America*, 46, 458-467.

[10] Bortz, J. (1979). *Lehrbuch der Statistik. Für Sozialwissenschaftler* (2nd ed), Berlin: Springer-Verlag.

[11] Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2),1-43. http://svm.first.gmd.de/.

[12] Zell, A. (1996). *Simulation neuronaler Netze.* Bonn: Addison-Wesley.

[13] Joachims, T. (1999). Making large-scale SVM learning practical. In: *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (eds.), Cambridge: MIT-Press.