First publ. in: Prosody 2000 : speech recognition and synthesis; 2 - 5 October 2000, Kraków, Poland / ed. by Stanisław Puppel... (Eds.). Poznań : Adam Mickiewicz University, Faculty of Modern Languages and Literature, 2001. - pp. 37-43

The Effect of Accentuation on Vowel Recognition

Bettina Braun, Jacques Koreman, Jürgen Trouvain

University of the Saarland, Institute of Phonetics, Saarbrücken, Germany
{bebr,koreman,trouvain}@coli.uni-sb.de

ABSTRACT

This study focuses on the implementation of the phonetic effects of vowel accentuation in automatic speech recognition (ASR). The durational and spectral effects of accentuation are investigated separately by manipulating the transition and observation probabilities in hidden Markov models. We also attempt to implement the undershoot hypothesis [1], which describes spectral reduction as a direct consequence of shortening. Our findings support the widespread belief that the transition probabilities, which indirectly model durational effects, are negligible, and that the distinction between accented and unaccented vowels is determined by the observation probabilities.

1. Introduction

Accented and unaccented vowels differ in both durational and spectral properties [2]. The undershoot hypothesis [1] describes undershoot of the target position for unaccented vowels as a direct consequence of vowel shortening. More recently [3], undershoot of the target position was described as an effect of speaking style and conversational demands. Since the conversational demands on discriminability are high for accented parts of the utterance, which carry the main information, these must not be reduced, while unaccented parts of the utterance may show a high degree of undershoot.

The importance of accentuation effects on vowel realisation are evaluated in a vowel recognition experiment by separately modelling accented and unaccented vowels and comparing the results with those when vowel models pooled for accentuation are used (section 3.1). In section 3.2 the undershoot hypothesis is implemented by shortening the self-loop probability of the middle state of each hidden Markov model (HMM) for unaccented vowels. Since this decreases the recognition rates, we have investigated the effects of separately modelling durational and spectral reduction in HMMs (section 3.3). Finally, we shall show in section 3.4 that the degree of accentuation can be determined better than we should predict on the basis of the frequency of occurrence of accented and unaccented vowels.

2. Method

As training and testing material we used the vowels in the prosodically labelled part of the KielCorpus of German spontaneous speech [4] (sampled with 16kHz/16Bit) – 88 minutes of 7 appointment-making dialogues of 32 speakers. Since the behaviour of vowel monophthongs in accented and unaccented conditions is well understood and, particularly, since we want to attempt to implement the undershoot hypothesis, which was developed on the basis of monophthongs, only these are used in our experiments (29,565 tokens). Four levels of accentuation have been manually labelled in the database: 0 for 'unaccented', 1 for 'partially accented', 2 for 'accented' and 3 for 'reinforced' vowels. These levels constitute sentence level prominence – not lexical stress, although this is obviously implied by accentuation (the reverse is not true). Although all vowel monophthongs were used in the experiments, only those vowels were considered for *interpretation* which provide enough data for testing and modelling our ideas. Vowels which have less than 90 occurrences in either the accented or unaccented category were excluded from the presentation (all front rounded vowels, /@/ and /6/).

The speech signals were parameterised using a 25.6ms Hamming window with a preemphasis of 0.97 and step size 5ms. For each frame 12 mel-frequency cepstral coefficients (MFCCs), their first derivatives and energy were extracted. The experiments were performed using the HTK toolkit [5]. Only self-loops and transitions to the next state were allowed in the HMMs, which consisted of 3 states. For training only the first six out of the seven dialogue games were used, testing material consisted of the seventh, i.e. training material as well as testing material contains speech samples of the same speakers. By using the same speakers for training and testing, speaker variation is better modelled. This allows us to concentrate on the effects of accentuation.

Our experiments differ from standard ASR experiments in that no lexicon or language model was used. *Phone* classification rather than *word* recognition was performed to ensure that we isolate the acoustic influence of accentuation on recognition.

3. Experiments

3.1. Accentuation and baseline experiments

In an accentuation experiment (Acc4) separate HMMs with 4 mixtures in each state were trained for each of the accented (levels 1, 2, and 3 accentuation labels) and unaccented (level 0) vowels. The four mixtures were used to model variation due to other factors than accentuation.

As a baseline for comparison, two experiments were carried out in which one overall HMM was trained for each vowel, i.e. the vowels were pooled for accentuation. In the first baseline experiment (Base4), four mixtures per state were used, while in the second (Base8) eight mixtures were used.

The results from the accentuation experiment were expected to lie between those from the two baseline experiments: on the one hand, the Base4 experiment has to model the complete variability in the signal with only four mixtures per state, where the Base8 experiment has double the number of mixtures to model the different realisations of each vowel. In the Acc4 experiment, the total number of mixtures across all the vowels is the same as in the Base8 experiment, but by separately modelling accented and unaccented vowels, four of the mixtures are explicitly assigned to the accented, the other four to the unaccented vowels. Since this may be less effective than when the eight mixtures are left free to model any source of variability in the signal, the Base8 experiment sets the recognition ceiling for the Acc4 experiment. On the other hand, since in both the Base4 and the Acc4 experiments we have four mixtures per state, but fewer data (less variability) are modelled by each HMM in the Acc4 experiment, the Base4 experiment should determine the minimum recognition.

	a	I	a:	i:	E	e:	O	U	o:	u:	Tot:
Base4	42.3	42.8	52.8	69.7	60.3	**65.4**	51.5	56.2	39.1	67.8	48.0
Acc4	**50.2**	47.0	**54.5**	68.3	62.4	60.3	**55.6**	57.7	**40.2**	**72.9**	49.4
Base8	49.1	**49.4**	51.6	**71.5**	**65.4**	60.7	52.7	**64.3**	39.5	64.6	**50.8**

Table 3.1. Recognition results (total rates for *all* monophthongs)

Recognition rates are shown in table 3.1. Please note that the recognition rates of Acc4 reflect recognition of vowel quality *irrespective of degree of accentuation.* Numbers in bold print indicate the highest recognition rate in this set of three experiments. Most vowels are best recognised in either the Acc4 or Base8 experiment, as was expected. Five of the vowels are best recognised in Acc4, four in Base8. The fact that dividing the vowels into two accentuation groups does not lead to a general deterioration of the recognition rates shows how important a factor accentuation is in determining variability among the vowels.

3.2. Modelling undershoot

Since there is a significant difference in the durations of the accented and unaccented vowels and since the undershoot hypothesis predicts the automatic spectral reduction due to time constraints, a first attempt to implement the undershoot hypothesis was made by deriving HMMs for unaccented vowels *from accented ones*. Unaccented HMMs were derived by decreasing the self-loop probability of the middle state of the HMM for the corresponding accented vowel (and increasing the exit probability of this state). This allows for modelling a reduction in the duration of the vowel and *can* at the same time model spectral undershoot. We argue that modelling durational effects by changing the transition probabilities is admissible, since there is a significant correlation between mean durations and mean self-loop probabilities (r=0.969).

Experiment AbsUS strongly reduces the self-loop probability of the middle state of all unaccented vowels to 20%. In a second experiment (RelUS) the self-loop probability is reduced according to the relative reduction in duration. The proportions of the mean vowel durations of accented to unaccented vowels were significantly correlated (r=0.901) with the proportion of the mean self-loop probabilities in the HMMs. The self-loop probabilities of the middle state were shortened using the regression formula y=0.72+0.28*R (R = durational proportion of each accented vowel to its unaccented counterpart). The self-loop probability for unaccented vowels is calculated as $p_{L2}*y^{3,}$, (p_{L2} = self-loop probability of the middle state of accented vowels).

Vowel recognition rates were equally low in both experiments (see table 3.2) compared to Acc4, in which accented and unaccented vowels were modelled separately. This shows that the chosen implementation of the undershoot hypothesis cannot be applied to derive unaccented from accented HMMs.

Interestingly, the recognition rates of the two undershoot experiments are comparable, although the differences in the self-loop probabilities of the middle state were enormous (mean probability for AbsUS 20%, for RelUS 70%). This leads us to believe that transition probabilities only play a minor role in the recognition of accented and unaccented vowels.

3.3. Influence of transition and observation probabilities

In a third set of experiments we try to reveal the influence of transition and observation probabilities on the recognition rates. As in the undershoot experiments presented in the previous section, HMMs for unaccented vowels were derived from those for accented ones.

In experiment Trans this is done by replacing the *transition* probabilities trained for accented vowels with those trained for unaccented vowels in Acc4. Likewise, in experiment Observ the *observation* probabilities trained for accented vowels were replaced with those trained for unaccented vowels in Acc4 (see also table 3.2).

Changing only the transition probabilities (Trans) leads to similarly poor results as in the undershoot experiments, again showing the relative unimportance of the transition probabilities for the recognition of accented and unaccented vowels. Replacing the observation probabilities to derive HMMs for unaccented vowels from those for accented vowels (Observ) led to comparably good results as Acc4.

Exper.	Acc-model	Unacc-model	rate
Acc4	Acc-vow	Unacc-vow	49.4
AbsUS	Acc-vow	Acc-vow, but self-loop prob. of 2nd state red. to 20%	41.7
RelUS	Acc-vow	Acc-vow, but self-loop prob. of 2nd state red. relative to dur. proportion of acc : unacc	41.9
Trans	Acc-vow	Acc-vow, but transition prob. of unacc-vow	41.7
Observ	Acc-vow	Acc-vow, but observation prob. of unacc-vow	49.1

Table 3.2. Recognition rates for all accentuation experiments

3.4. Accent Recognition

We have modelled accentuation in ASR, because it is an important prosodic means to transport information structure. In experiment Acc4, all accented vowels (except /I/ and /O/) are recognised better than unaccented ones (not shown in table 3.1). This is what we expected given the unreduced forms of accented vowels. Although considerable overlap should be expected between accented and unaccented vowels due to other sources of variation, the accent recognition rate is 59.2% for accented and 72.1% for unaccented vowels. This compares favourably with chance level, based on the frequency of occurrence of accented and unaccented vowels, which is 36.3% for accented and 63.7% for unaccented vowels in our corpus.

4. Discussion

It is shown that dividing our corpus into accented and unaccented vowels (Acc4) leads to results which are comparable to those of modelling the data with the same number of mixtures, but leaving it up to HMM how the variability in the signal is modelled (Base8). This proves the importance of accentuation as a source of variation in the signal. The advantage with a controlled splitting of the data is that we also obtain information about the degree of accentuation, which can be used for higher-level processing.

Table 3.1 shows that recognition rates decrease when we try to derive models for unaccented vowels from accented models by simply decreasing the self-loop probability of the middle state in order to implement the undershoot hypothesis. This may be due to the fact that the states in a HMM are related to, but certainly do not exactly correspond to phonetic “categories” like transitions and steady states. A different type of HMMs which does not model the datapoints as independent observations may be more appropriate for this purpose. Further, both methods of reduction of the self-loop probabilities of the middle state (AbsUS, RelUS), performed equally poorly. This led us to believe that transition probabilities only play a minor role in recognition of accented and unaccented vowels.

The third set of experiments, in which HMMs for unaccented vowels were again directly derived from those for accented vowels, corroborates this hypothesis. When unaccented vowels are modelled by replacing the *transition* probabilities by those from the unaccented vowel HMMs in Acc4 (Trans), this leads to an equally poor performance as the undershoot experiments. Replacing the *observation* probabilities with those from the unaccented vowels to create HMMs for unaccented vowels (Observ), however, leads to recognition rates comparable to those of Acc4, in which *both* the observation *and* the transition probabilities were different for accented and unaccented HMMs. Therefore, only the observation probabilities seem to be of importance in modelling the effects of accentuation on vowels. In so far, our findings deviate from those of experiments on speaking rate [6], where fast speech was successfully modelled by increasing the exit probabilities (= decreasing the self-loop probabilities) of phones spoken at a normal rate.

Normally, human listeners use multiple cues to recognise whether a word is highlighted or not. We show that accent can be predicted correctly to 67.4% from the vocalic portions only. Information about accentuation may be usable for “higher-level” linguistic processing, such as decoding information packaging, resolving lexical and part-of-speech ambiguities.

5. Conclusion

Durational and spectral reduction is caused by deaccentuation, and is explained by a greater amount of undershoot. As was shown, this cannot be modelled by simply increasing the transition probabilities of the middle state of HMMs for vowels (at least not for the type of HMMs used here). It was shown that transition probabilities in general cannot model the effects of deaccentuation and that spectral properties are decisive in recognising accented and unaccented vowels.

Being able to determine accentuation is of great advantage for subsequent linguistic analysis. Since distinguishing accented and unaccented vowels in hidden Markov modelling does not deteriorate recognition results, this approach should be preferred over models with simply more mixtures to model the variation in the signal.

REFERENCES

[1] Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. In: *Journal of the Acoustic Society of America (JASA)* 35. pp. 1773-1781.

[2] van Bergem, D.R. (1993). Acoustic vowel reduction as a function as a function of sentence accent, word stress and word class. *Speech Communication* 12. pp. 1-23.

[3] Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In: W.J. Hardcastle and A. Marchal (eds.) Speech Production and Speech Modelling. Kluwer. pp. 403-439.

[4] Kohler, K.J., M. Pätzhold, A.P. Simpson (1995). From scenario to segment: the controlled elicitation, transcription, segmentation and labelling of spontaneous speech. *Arbeitsberichte Phonetik Kiel* 29.

[5] Young, S. (1996). *The HTK Book*. Cambridge University Press.

[6] Morgan, N., E. Fosler, N. Mirghafori (1997). Speech Recognition using on-line Estimation of Speaking Rate. In: *Proc. Eurospeech Rhodes*. pp. 2079-2082.