USING LEXICAL STRESS IN CONTINUOUS SPEECH RECOGNITION FOR DUTCH

David van Kuijk (1+2), Henk van den Heuvel (1), Loe Boves (1)

(1) Department of Language and Speech, University of Nijmegen, The Netherlands
(2) Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands
Snail-mail: P.O. Box 310, NL-6500 AH Nijmegen, The Netherlands
E-mail: kuijk@mpi.nl

ABSTRACT

The acoustic realization of vowels with lexical stress generally differs substantially from their unstressed counterparts, which are more reduced in spectral quality, shorter in duration, weaker in intensity and tend to have a flatter spectral tilt. Therefore, in an automatic speech recognizer it would appear profitable to train separate models for the stressed and unstressed variants of each vowel. A problem is how to define the mapping from the theoretical stress of words to the actual realization of stress in fluent speech. We compared several hypotheses about this mapping applied in both training and testing of the recognizer. The recognition results on an independent test-set showed that recognition rates did not increase by this use of stress in our ASR. Possible explanations are discussed and future research plans are outlined.

1. INTRODUCTION

Modern automatic speech recognizers do not usually model stressed and unstressed variants of vowels separately. Nonetheless, speech recognition might be enhanced by including this distinction since a variety of studies has demonstrated that stressed and unstressed vowels differ with respect to various acoustical dimensions.

Van Bergem (1993) compared the characteristics of Dutch stressed and unstressed vowels in read sentences. He found that in content words the unstressed variant of a vowel (e.g. /æ/ in canTEEN) was on average shorter than the stressed variant (e.g. /æ/ in CANdy), while the variant that was the rhyme of a function word (e.g. /æ/ in can) was still shorter. Furthermore, in polysyllabic words the vowels from stressed syllables were generally more clearly pronounced (with respect to F1 and F2) and therefore closer to their target form than the vowels from unstressed vowels.

Sluijter and van Heuven (1996) studied the occurrence of stress in minimal stress-pairs (e.g. kaNON - KAnon 'cannon' - 'canon'). They found that the duration of a vowel is a good predictor for the factor lexical stress. Spectral slope (defined by energies in 4 bands in the 0-2 kHz domain) was as good a predictor as duration, but the intensity levels of stressed and unstressed variants of the same vowel were almost the same. Furthermore perception-experiments (Sluijter, van Heuven, & Pically, 1996) showed that the subjects were more inclined to base their judgements about the stressedness of syllables on the spectral slope than on the overall intensity.

We compared the durations and intensities of the stressed and unstressed variants of each vowel in 2500 utterances from the Polyphone-corpus (Damhuis et al., 1994). The durations were significantly different on a t-test for all vowels except /øː/, /œy/, and /ɔ/. The energies of the vowels were significantly different on a t-test for all vowels except /ɛi/.

We may conclude that there are differences in the durations, in the intensity, and in spectral balance between the unstressed and stressed variants of the same vowel. In conventional HMM speech recognizers duration is difficult to model, but spectral differences and energy differences are generally modeled in a straight-forward manner. Therefore, it seems reasonable that vowels can be better modeled in an automatic speech recognizer if the distinction between stressed and unstressed vowels is taken into account in creating the models.

For Dutch ASR systems there are no reports on attempts to use stress in ASR, but for English there are a number of papers on the subject. Dumouchel and O'Shaughnessy (1993) included lexical stress (and other prosodic factors) as an extra information stream in the Bayesian decision logic of their HMM-based speech recognizer (during recognition only). The acoustic features used to model the prosodics were the logarithms of the duration, intensity and F0. They tested the effect of stress assignment in their recognizer as follows. First recognition was performed without prosodic knowledge. Then for each falsely recognized utterance two segmentations were made, which were rescored on the basis of prosodic information. Stress assignment would favor recognition if the addition of knowledge about stress lowered the probability of the recognized word string, but increased that of the actual word string. In this experiment lexical stress information did not improve recognition results.

Hieronymus, McKelvie, and McInnes (1992) included a hybrid prosodic component in their speech recognizer which determined the sentence level stress (accent) and marked the vowel of stressed (accented) syllables in the phoneme lattice. In the recognition lexicon the stress was marked on all content words. A 65% reduction in word error rate (WER) and 45% reduction in sentence error rate (SER), relative to a baseline system without prosodic information, was reported.

In both studies the authors suggested that stress should also be used in the training phase to (further) improve recognition results. Exactly this approach was chosen in the present paper.

2. METHOD

We use an off-the-shelf monophone based, large vocabulary continuous speech recognizer trained for telephone speech. The feature set comprises 14 Mel-scaled filter log power values in the range of 350 to 3400 Hertz, their deltas, log energy over all filter bands, with delta and delta-delta energy (in all 31 features per 10 ms-frame). Each phone model consists of three states, while each state is subdivided into two substates with tied observation distributions. First, the recognizer is trained for the baseline condition where there is no distinction between stressed and unstressed vowels indicated in the lexicon (stress0 condition). This minimal configuration contains 38 phone models. Then we train the recognizer anew while indicating the primary stress in each lexicon entry, thus making separate HMM-models for vowels in syllables with and without lexical stress.

The mapping from the primary stress as indicated in the dictionary to the realization of stress in fluent speech is not very clear. It is assumed that normally short (monosyllabic) function words are not pronounced with stress (see also van Bergem's results), but it is not clear whether this is only true for function words with a schwa, or also for other frequently used function words with non-schwa vowels. Therefore we introduce two experimental conditions with different hypotheses about the mapping from citation form stress to stress realized in fluent speech.

In the first experimental condition (stress1) we adopt the lexical stress as indicated in the CELEX-database (Baayen, Piepenbrock & van Rijn, 1993), but remove the stress mark from monosyllabic words that contain a schwa-vowel. In addition, we introduce a second experimental condition (stress2) in which, apart from monosyllabic words with a schwa, also a subset of the Dutch function words (articles, conjunctions, and pronouns) are considered to be unstressed, while all other words bear stress as indicated in CELEX. As a result, the proportion of stressed vowels is higher in condition stress1 than in stress2. We regard condition stress2 as the most realistic one. Both recognizers distinguishing stressed and unstressed vowels contain 54 phone models.

In all conditions the recognizer is trained with 5000 phonetically rich sentences from the Polyphone corpus (Damhuis et al., 1994). The Polyphone corpus was designed in such a way that the five phonetically rich sentences of each of the speakers contained at least one example of each Dutch phoneme (not taking stress into account). All five utterances of a selected speaker are included, if possible. But some of the sentences are discarded due to bad quality. Thus, the train set covers utterances from 551 male and 551 female speakers. The speakers are selected from all Dutch provinces, and the number of speakers from each part of the country is equal. The resulting train set of 5000 sentences comprises 9 hours and 6 minutes of speech (including silences), and is composed of 9,012 words and 53,268 word tokens. This corresponds to an average of 10.7 words per sentence. The minimum frequency of occurrence of a vowel in the stress0 condition is 1,078 (phoneme /øː/). For the stress1 condition the unstressed vowel /øː/ is the least frequent vowel, occurring 231 times. The least frequent vowel for stress2 is also unstressed /øː/, occurring 231 times. The number of words in the train-lexicon which is stressed in the stress1-condition, but not in the stress2-condition, is 103. This is not much, but due to the high frequency of occurrence of those words the difference between the two conditions can result in different models.

The test set consists of 480 sentences from the development-test part of the Polyphone corpus of phonetically rich sentences. Again, we tried to include all five utterances of a selected speaker. The utterances in the test set stem from 60 males and 60 females. It contains 52 minutes of speech, and is composed of 1886 words (which is also the size of the test lexicon) and 5,143 word tokens. The perplexity of the test set is 30.38. The recognizer operates with a unigram and bigram language model which is trained on all 2299 sentences of the development-test part of the Polyphone database.

It can be argued against our experimental set-up that a decrease in WERs could be due to doubling the number of mixtures available for modeling vowels (by taking two acoustic models instead of one for each vowel), rather than to significant differences between the unstressed and stressed variants of vowels. In a control experiment we will swap the stress marks in the recognition lexicon: stressed vowels are marked as unstressed and vice versa. If the recognition results deteriorate for this control then we have collected evidence that the recognizer has learned at least some of the acoustic differences between stressed and unstressed vowels.

3. RESULTS

The word error rates (WERs) on the test set are summarized in Table I. This table also lists the WERs for the conditions in which the stress marks in the test lexicon were swapped. The WERs are shown as a function of the number of (LaPlacian) components in the mixture densities with which the recognizer was trained.

Table I shows that the increase in the number of LaPlacians in the mixture improved the recognition results (as expected), but that the distinction between stressed and unstressed vowels did not, despite the fact that the number of vowel models is doubled in the stress1 and stress2 conditions.

Further, it can be seen from the table that swapping the stress marks in the test lexicon yields deteriorated recognition results. A computation of confidence intervals (95% level) for WERs reveals that the difference in WERs between the normal stress and the swapped stress conditions is significant for all (six) comparisons.

4. DISCUSSION

Our results did not confirm the hypothesis that the distinction between stressed and unstressed vowels implemented both in training and recognition would improve recognition results.

It could be argued that after splitting the vowel models into a stressed and unstressed variant too few observations remained for at least some of the vowels to be sensibly modeled. We do not think this is the case, since we had over 230 observations for each of the vowels left and for most of them considerably more.

Table I shows that the training of separate models for the stressed and unstressed variants of each vowel does not lead to better recognition results, even though effectively the number of mixtures

	Normal Conditions			Swapped Conditions	
Number of mixtures per state	**stress0**	**stress1**	**stress2**	**stress1**	**stress2**
4	29.85	29.39	29.13	32.76	32.89
8	24.63	24.51	24.21	28.11	27.81
16	20.96	21.06	20.98	25.60	24.61
32	19.96	19.90	19.88	24.83	23.57

Table I: The word error rates over the test-set of 480 utterances for the various stress-conditions and for various numbers of mixtures. The WERs for the conditions where the stress marks in the test lexicon were swapped are shown in the two right-most columns.

which is spent for each of the vowels is doubled. However, doubling the number of mixtures for the total set of phonemes strongly improves recognition performance. This suggests that the improvement of recognition rates obtained by doubling the number of mixtures for all phonemes is mainly due to a better modeling of the consonants.

The finding that swapping the stress marks in the test lexicon leads to worse recognition results favors the idea that the vowel models do contain meaningful stress information: if the models for stressed and unstressed vowels were similar then swapping the stress marks would not make any difference. Therefore, it seems that the recognizer has built different models for stressed and unstressed vowels, but that it could not profit from these separate models to increase recognition performance.

However, another explanation for the swap results is possible. We examined the phonemic contexts in which the stressed and unstressed vowels occurred both during training and recognition. It is possible that stressed vowels tend to appear more often in a certain context than their unstressed counterparts. If the same contextual bias appears in the training and test set then the results for the swapped stress conditions did not only reflect stress-based differences in the models, but also effects of the vowel context. By counting triphone occurrences for the vowels we found that certain contexts were typical for stressed vowels whereas other contexts were typical for unstressed vowels, and that the same biases were present in the training and the test sentences. For example the triphone {n}i{t} occurred 518 times and the triphone {n}'i{t} only 8 times in the training set. In the test set {n}i{t} was observed 59 times, and {n}'i{t} 3 times. This suggests that stress information may have been confounded with contextual information in the vowel models of the recognizer. It will be worthwhile to investigate whether the same contextual distributions are observed in much larger corpora like CELEX, and whether similar results are found if context-dependent models are used for training and testing the recognizer.

As far as the stress-related information in the models is concerned the lack of improvement in the WERs may have two underlying causes. (1) The recognizer needs more explicit stress-related information in its acoustic vector to be able to utilize this information to improve its recognition performance. (2) The acoustical correlates of stress in fluent speech are more variable than suggested by the phonetic studies discussed in the introduction, due to higher level phenomena, such as the assignment of sentence accents and rhythmic patterns. In that case, lexical stress in an automatic speech recognizer can only meaningfully be exploited in a way that was reported by Hieronymus et al. (1992).

5. CONCLUSIONS

In this study the effect of training separate models for the stressed and unstressed variants of each vowel in an HMM-recognizer was investigated. Based on acoustic-phonetic studies we hypothesized that a continuous speech recognizer could benefit from creating separate models for stressed and unstressed vowel variants, especially if this distinction was made not only during recognition but also during the training of the models.

This hypothesis was not confirmed by our results, because the WERs for our recognizer did not decrease by the inclusion of stress. However, by swapping the stress in the recognition lexicon we established that the models for the stressed and unstressed vowels did differ. Further analysis showed that stressed vowels and unstressed vowels tend to appear in different contexts, so that stress and context information are confounded in the statistics of the models.

Future work will investigate the inclusion of more explicit stress-modeling in HMM-recognizers, in which the duration of the vowels (which is an important factor in distinguishing between stressed and unstressed vowels), will be modeled as well. To this end it is necessary that normalizations are carried out for duration (on the basis of speaking rate) and intensity (on the basis of the intensity of surrounding vowels), so that the properties of lexical stress as a relational phenomenon will be taken more into account than in the present study.

6. REFERENCES

Baayen, R.H., Piepenbrock, R., and van Rijn, H. (1993). *The CELEX lexical database (on CD-ROM).* Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

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

Damhuis, M., Boogaart, T., in't Veld, C., Versteijlen, M., Schelvis, W., Bos, L., and Boves, L. (1994). Creation and analysis of the Dutch Polyphone corpus. *Proceedings ICSLP 94*. Yokohama.

Dumouchel, P., and O'Shaughnessy, D. (1993). Prosody and continuous speech recognition. *Proceedings Eurospeech 93*. Berlin, III, 2195-2198.

Hieronymus, J.L., McKelvie, D., and McInnes, F.R. (1992). Use of acoustic sentence level and lexical stress in HMM speech recognition. *Proceedings ICASSP-92*, San Francisco, I, 225-229.

Sluijter, A.M.C., and van Heuven, V.J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *JASA*, in press.

Sluijter, A.M.C., van Heuven, V.J., and Pacilly, J.J.A. (1996). Spectral balance as a cue in the perception of linguistic stress. *JASA*, in press.