ELSEVIER

Speech Communication 27 (1999) 95–111

SPEECH COMMUNICATION

Acoustic characteristics of lexical stress in continuous telephone speech

David van Kuijk [a,b,*], Loe Boves [a]

[a] A^2RT, Department of Language and Speech, University of Nijmegen, Nijmegen, The Netherlands
[b] Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands

Received 10 May 1998; received in revised form 3 November 1998; accepted 17 November 1998

Abstract

In this paper we investigate acoustic differences between vowels in syllables that do or do not carry lexical stress. In doing so, we concentrated on segmental acoustic phonetic features that are conventionally assumed to differ between stressed and unstressed syllables, viz. Duration, Energy and Spectral Tilt. The speech material in this study differs from the type of material used in previous research: instead of specially constructed sentences we used phonetically rich sentences from the Dutch POLYPHONE corpus. Most of the Duration, Energy and Spectral Tilt features that we used in the investigation show statistically significant differences for the population means of stressed and unstressed vowels. However, it also appears that the distributions overlap to such an extent that automatic detection of stressed and unstressed syllables yields correct classifications of 72.6% at best. It is argued that this result is due to the large variety in the ways in which the abstract linguistic feature 'lexical stress' is realized in the acoustic speech signal. Our findings suggest that a lexical stress detector has little use for a single pass decoder in an automatic speech recognition (ASR) system, but could still play a useful role as an additional knowledge source in a multi-pass decoder. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Dutch; Prosody; Lexical stress; Automatic vowel classification; Automatic speech recognition

1. Introduction

Most research on automatic speech recognition (ASR) of the last few decades has focused on the segmental level. Of the units on 'higher' levels of linguistic description only words and word sequences have received due attention, but only in terms of Language Models, which can be estimated independently of the acoustics and phonetics of actual speech. This focus has certainly contributed to the enormous progress of the field. However, it has never been disputed that speech is more than just a concatenation of segments that make up words which, in turn, make up phrases. Suprasegmental features have been studied intensively in phonetics, phonology and linguistics, but the results of these investigations have had very little impact on (and so far have hardly made any contribution to) the progress in ASR.

Two related suprasegmental features that have received much attention in phonology and linguistics, and also in phonetics, are stress and accent. Languages like English and Dutch are known

[*] Corresponding author. Address: Roukensstraat 40, 6521 BP Nijmegen, The Netherlands. Tel.: +31 24 3812217; e-mail: david@lcn.nl

PII: S 0 1 6 7 - 6 3 9 3 (9 8) 0 0 0 6 9 - 7

to have words which consist of essentially identical phoneme sequences, yet can be distinguished, thanks to the fact that word stress is on different syllables. Conventional examples include English word pairs like 'COMment' and 'comMENT', 'REcord' and 'reCORD', etc., but also less well-known pairs like FORbear and forBEAR. Examples in Dutch are word pairs like 'VOORkomen' (exist) and 'voorKOmen' (prevent), 'Overleggen' (present) and 'overLEGgen' (discuss).

Although both 'stress' and 'accent' relate to this distinction, it is now generally accepted that the two terms should be used for two distinct phenomena. Stress is an abstract feature on the level of the lexicon: stress is assigned to that syllable of a word that will stand out more conspicuously if that word is produced with an accent. By implication, accent is a phonetic feature, with measurable correlates in production, acoustics and perception. In principle, each word, including monosyllabic function words, can be accented. Thus, each and every word in the lexicon has a stress-hook assigned to one of its syllables. For monosyllabic words this seems redundant, but if such a hook were not present, accent could never be attached to monosyllables, which is clearly inadequate. It has been known for a long time that there is substantial and putatively systematic interaction between stress and the acoustic–phonetic features of the vowel in stressed syllables. Of course, it would have been surprising if this had not been the case, given the definitions of stress and accent. Therefore, one would expect that clever use of the linguistic feature stress should contribute to the performance of ASR algorithms: if vowels in stressed and unstressed syllables are really different, then training separate models for the two 'contexts' should help to reduce the amount of 'systematic' variance in the models.

In the past a small number of attempts have been made to use stress in the development of ASR (e.g., Waibel, 1986; Dumouchel and O'Shaughnessy, 1993; Hieronymus et al., 1992). However, up to now the contribution of stress to the performance improvement of ASR has been equivocal. In this paper we investigate the relation between stress and acoustic–phonetic features of the vowel sounds. The results should help us understand why previous attempts to deploy stress in ASR have met with little success, and show possible routes to using stress in future ASR algorithms.

1.1. Stress in ASR

Van Kuijk et al. (1996) reported on experiments with lexical stress in an off-the-shelf HMM-recognizer. In these experiments, the recognizer was trained with different models for the stressed and unstressed variants of each vowel. To this end the lexical stress was indicated in each word in the lexicon derived from the training corpus. Although the concepts stress and accent, as well as their distinction, have been very fruitful in linguistic theorizing, they may not be optimally suited for use in ASR: In the lexicon the abstract feature stress is not only attributed to the most prominent syllable in words like *TElephone*, where the stressed syllable is likely to be acoustically different from the surrounding unstressed syllables, but also to monosyllabic function words like *the*, *a*, *is*, etc., which are normally pronounced much like the unstressed syllables in TElephone. This discrepancy between linguistic theory and phonetic 'reality' might very well interfere with the systematic difference in the acoustic–phonetic properties of vowels in stressed and unstressed syllables. Therefore, Van Kuijk et al. (1996) decided to experiment with different mappings from theoretical stress to 'phonetic' stress in the lexicon, to see whether this would influence the recognition scores. In the first mapping stress was removed from all monosyllables with a schwa, and in the second mapping also a subset of the Dutch function words were considered to be unstressed. Their results showed that none of the mappings significantly improved the performance of the recognizer.

Van Kuijk et al. (1996) mentioned several possible explanations for the failure of the feature stress to enhance recognition performance. The first possible explanation is that the number of Gaussian mixtures used for the unseparated hidden Markov models (HMM) is high enough to capture all acoustic–phonetic variation between stressed and unstressed variants of a vowel. They controlled for this possibility by varying the number of mixtures from 4 up to 32 per state. Even with 4 mixtures per state the recognizers which distinguished between models for stressed and unstressed vowels did not perform better. However, increasing the number of mixtures *did* significantly decrease the Word Error Rates. Thus, this first explanation is not very likely.

The second possible explanation is to assume that – at least in Dutch – the linguistic concept stress has no systematic acoustical correlates after all. In this case no recognizer could ever profit from training separate models for stressed and unstressed vowels (and any improvement found in the recognition scores would simply be due to an increase in the number of parameters). This issue is the focus of the present paper.

The third explanation assumes that the acoustic features that are conventionally used in HMM and Neural Net ASR algorithms (log band filter coefficients, cepstral representations of the log spectra, their delta's, log-energy, delta-log-energy, and delta-delta-log-energy) are not appropriate to distinguish stressed vowels from unstressed ones. Specifically, HMM decoders do not explicitly model duration, while Van Bergem (1993), and Sluijter and Van Heuven (1996) showed that duration is (one of) the most important indicators for lexical stress in Dutch. The present paper will also address this explanation by investigating whether features which are not commonly used in ASR algorithms can discriminate between stressed and unstressed vowels.

Then, the speech material used in Van Kuijk et al. (1996) was recorded over the telephone. The distortion due to transmission over the Public Switched Telephone Network (PSTN) might have affected the acoustic–phonetic properties that support the distinction between stressed and unstressed syllables. However, since humans have no difficulty in distinguishing words like COMment and comMENT in telephone speech, this hypothesis is not very likely.

To set the scene for the investigation into the relation between word stress and acoustic–phonetic properties of vowels, we will summarize the most important findings from the literature. For two reasons much attention will be given to stress in Dutch: first, the speech material under investigation is Dutch; second, a relatively large proportion of the phonetic research into stress is based on Dutch as the target language.

1.2. Acoustical differences between stressed and unstressed vowels

Several researchers have studied the acoustic correlates of stress in Dutch. Van Bergem (1993) – who used sentences read by 15 male speakers in response to audio prompts that were designed to put an accent on the syllable of interest or, alternatively, to shift accent to another syllable – found that stressed vowels have a longer duration, and that in polysyllabic words the vowels in stressed syllables have formant patterns that resemble the formant patterns of vowels produced in isolation much better than do vowels in unstressed syllables. Sluijter and Van Heuven (1996) – using one word pair, embedded in two short sentences – found that duration is the most important acoustic correlate of stress. Stressed vowels are longer, and the duration of a vowel is a good predictor of its stress-level. Spectral tilt is almost as good a predictor of stress as duration. Stressed vowels tend to have a flatter negative spectral slope than unstressed vowels, due to the fact that they are produced with more vocal effort. The energy in the lower frequency regions (0–0.5 kHz) is hardly affected by stress, but the energy in the other regions (0.5–4 kHz) is higher for stressed vowels than for unstressed vowels. They also found that, compared to duration and spectral tilt, intensity and vowel quality (in terms of formant patterns) are less powerful predictors of stress.

1.3. Automatic classification of stressed/unstressed vowels

Another approach to the study of differences between stressed and unstressed vowels is to see whether it is possible to tell the two classes apart automatically. Ying et al. (1996) have attempted automatic classifications of stressed and unstressed vowels for American English. In their study, for syllables in bi-syllabic stress-minimal word pairs (e.g. 'OBject–obJECT') very good classification results (10% errors) were obtained on the speech of a single male speaker for the words embedded in 112 different carrier phrases. The features used for classification were normalized duration and normalized energy. For the same set of words, spoken by the same speaker, in more natural sentences, 4% classification errors were obtained with a three-feature classifier (using two different normalizations for Duration and one for Energy). Adding three more male speakers and one female speaker to the database and retraining and testing with this material, even decreased the error rate to 2%. It is interesting to note that Energy appears to be instrumental in automatic classification of stressed/unstressed vowels, despite the fact that Sluijter and Van Heuven (1996) found that it is not the most efficient feature in human classification.

Research works like those of Ying et al. (1996), Sluijter and Van Heuven (1996) and Van Bergem (1993) certainly have contributed to our understanding of the differences between stressed and unstressed vowels. However, the artificial character of the speech material they used makes generalization of the findings to speech produced under less controlled conditions questionable. Waibel (1986) already used continuous speech from a small database of 50 sentences to train a Bayesian classifier for the stressed/unstressed dimension. On a test set of 192 sentences (including the training set) he obtained correct classification results of about 88%. The feature which gave the best results was the peak-to-peak sound power integral over the sonorant region of the syllable, which is a kind of combined duration and energy measure over the vowel. Although the classification scores obtained by Waibel (1986) are clearly worse than the performance of the classifier trained by Ying et al. (1996), they are still encouraging.

1.4. Aims and design of this study

In this study we want to investigate whether stressed and unstressed vowels can be distinguished automatically in speech material that has NOT specifically been designed for this purpose. Therefore, we will use recordings of arbitrary (but relatively short) sentences adapted from newspaper text. We will investigate to what extent it is possible to distinguish stressed and unstressed vowels from each other in a Bayesian classifier on the basis of the acoustic–phonetic features duration, energy, and spectral tilt (i.e., the features that have been shown to be most directly related to the distinction between stressed and unstressed vowels). We decided not to try to use formant patterns as features for classification, because we expected that any attempt at automatic formant extraction would generate so much noise as to invalidate the features. However, if there are systematic spectral differences between stressed and unstressed vowels that can be expressed in terms of filter bank energy values or cepstral coefficients, that information will always be reflected in whatever models are built from these parameters. We also tried to diminish the variance resulting from varying recording conditions of our speech material and from different speaking styles, by studying several different techniques to normalize the raw acoustic feature values.

The next section will explain the experimental design in more detail.

2. Method

For our analyses we used phonetically rich sentences from the Dutch Polyphone corpus (Den Os et al., 1995a). These sentences were automatically segmented into phoneme-like units, using the baseline version of an HMM recognizer trained on an independent part of the Polyphone corpus. From the segmentation we could derive for each vowel the raw acoustic features, and derive normalized measures for those features. The distributions of the raw features in stressed and unstressed vowels were analyzed, and the raw and normalized features were used in our classification experiments.

2.1. Speech material

The Polyphone corpus contains speech from 5050 speakers from all regions of The Netherlands. These speakers had to answer a number of questions and read digits and sentences over the telephone. Among the text read out by each participant were five phonetically rich sentences, i.e., sentences selected in such a way that the canonical transcription of each complete set of five sentences contained all phonemes of the Dutch language. No attempt was made to ensure that all vowels would occur at least twice, viz. in a stressed and in an unstressed syllable. Nor did we attempt to balance diphone sequences. The five sentences in a set were chosen from a total of 12,500 different sentences used in the corpus. The sentences were adapted from newspaper text. In order to enhance readability, none of the sentences in the corpus contained more than 80 printing characters in the orthographic transcription.

From these phonetically rich sentences we composed a training set and a test set, each consisting of 5000 sentences. In our choice of sentences we applied the following constraints:

- use only sentences which are of reasonable acoustic quality (as marked by the producers of Polyphone);
- try to use equal numbers of sentences by men and women;
- try to use an equal number of speakers from each Dutch province;
- try to keep the five sentences of each speaker together in one set (so that for each speaker at least one occurrence of each phoneme is present in the set).

These constraints maximize the variability in terms of regional accent, male–female distinction, and phoneme-variability.

The lexical stress for each word was adopted from the CELEX database (Baayen et al., 1993), but we removed the stress mark from the Dutch function words listed in (Van Wijk and Kempen, 1980). Schwas were excluded from all analyses, because they can never be stressed.

The total number of words in the 10,000 sentences was 104,310. Of these words 7% were compounds. The proportion of content words (Nouns, Verbs, and Adjectives) was 52%, so it was about equal to the proportion of words from other word classes (48%). There were less monosyllables in the content words than in the other word classes: 31% of the content words were monosyllabic, while in the other word classes 89% was monosyllabic. As can be expected, many of the content words contained a stressed syllable (87%), while for the other word classes only 14% of the words comprised a stressed syllable, according to the criteria set out above. The reason that not all content words have a stressed syllable is that the auxiliary verbs are counted as unstressed in Van Wijk and Kempen's list. Furthermore, this list does not contain all the Dutch adverbs, which explains why also words which are not a content word can be stressed.

2.2. Preprocessing of the speech data

The speech material was pre-processed by applying an FFT with a 10-ms frame shift and a Hamming window with a length of 16 ms. From the FFT coefficients we computed 14 Mel-scaled filter log power values in the range of 350–3400 Hz. The log energy for each frame was computed as the mean over all filter bank values in that frame.

We trained our off-the-shelf continuous mixture density HMM-recognizer with the 5000 training sentences, and then carried out a forced segmentation on both the training and the test set. Segmentation was accomplished by limiting the search space of the recognizer to exactly the canonical transcription of each individual utterance, and storing the trace back information of the Viterbi search. Part of the resulting segmentation of the speech material was checked manually by trained phoneticians, who agreed that for the vowels the segmentation was similar in quality to a hand-segmentation by a phonetician. Gross discrepancies between the canonical transcription and the actually produced speech signal would lead to a complete failure of the recognizer to obtain an alignment. Such cases were not observed in the material used in this study. Nevertheless, local discrepancies may have gone unnoticed, but we are confident that these did not significantly affect our results.

2.3. Features

In this study we used a comprehensive set of acoustic features (related to duration, energy and spectral tilt) that have been reported to be related to stress. The raw features were derived from the results of the forced segmentation. From these raw features we computed a number of normalized features which take the phonetic and linguistic contexts of the individual vowels into account. The computations for the raw and normalized features are explained below.

2.3.1. Raw features

Duration. The feature DURATION is straightforwardly determined by the forced segmentation. Measurement accuracy is determined by the 10-ms frames. The minimum DURATION of the vowel tokens in the speech material is 30 ms, because in our recognizer a minimum of three states per acoustic model must be visited. This minimum length of 30 ms could have been too long for reduced vowels, but does not cause a problem for our experiment since all vowels in this experiment are non-reduced.

Energy. The energy of a particular speech frame is computed as the average over the spectral features. We computed two energy features for each vowel: the maximum energy (MAXENE), and the total energy (TOTENE). The MAXENE of a vowel is defined as the log energy of the frame which has the highest energy value. The TOTENE of a vowel is an integration of the energy over all frames within that vowel. So TOTENE is implicitly also sensitive to the duration of the vowel. In a pilot study we found that the Energy features became better predictors of stress if we applied spectral mean subtraction (with the means computed over the whole utterance) to the spectral features before computing the energy, so we did this.

Spectral tilt. The spectral tilt of a vowel was estimated as the difference between the energy in the lower spectral bands and that in the higher spectral bands. Different definitions of 'lower' and 'higher' frequency bands lead to different estimates of the 'tilt' feature. The spectral tilt was computed over the spectral bands in three frames: the frame with the highest energy in the vowel (so the same frame from which we took the MAXENE), and both the frames preceding and following that frame. Spectral tilt was expressed in two ways: TILT1000 was computed from the three frames by subtracting the sum of the log-energies in the bands from 1170 up to and including 3400 Hz from the sum of the log-energies in the spectral bands ranging from 350 Hz up to and including 1000 Hz. TILT570 was computed in the same way by summing the log-energies in the bands from 350 up to and including 570 Hz, on the one hand, and summing the log-energies in the bands from 700 Hz to and including 3400 Hz on the other hand, and then subtracting the latter total from the former. Because of this way of computing TILT570 we can expect a considerable absolute correlation between MAXENE and TILT570.

2.3.2. Normalized features

The raw features defined above are known to be highly context dependent. For instance, for a fast speaker the duration of a stressed vowel may well be shorter than the duration of the same vowel in an unstressed syllable spoken by a slower speaker. Therefore, we defined several transformations of the raw features, which are all meant to reduce the context dependence. The categories of normalizations used are summarized in Table 1. In all cases, 'normalization' amounts to expressing the feature value of a vowel relative to the values in its neighboring vowels. Different definitions of 'neighbors' lead to different normalizations. One set of normalizations takes a complete sentence as the context (numbered 1 and 2 in Table 1). Another set (3, 4) relates the feature values to the values of only the preceding vowels in the same sentence; this kind of normalization is inspired by the observation that humans can detect stressed syllables in on-line tasks (i.e., it is not necessary to wait for the sentence or even for the word to finish), as appears from e.g. shadowing tasks (Radeau and Morais, 1990). Yet another set of normalizations (5) takes only the immediate left and right neighboring vowels into account, while (6) only normalizes for the vowel immediate left of a vowel. The normalizations 7 and 8 in Table 1 are specifically designed for the energy-features and the DURATION feature. They will be discussed below.

Normalized duration. The duration of a vowel is known to be influenced by many factors, like the intrinsic duration of the vowel, speaking rate, lexical stress, position in the utterance in which it

Table 1
A short description of each of the normalizations applied to the raw features in our classification experiments

Description of different values used for normalizing feature F in vowel V	DURATION	MAXENE	TOTENE	TILT1000	TILT570
1. Average value of F as computed over all vowels in the utterance.	X	X	X	X	X
2. Average value of F as computed over all phonemes in the utterance	X				
3. Average value of F as computed over all vowels left of V in the utterance	X	X	X	X	X
4. Average value of F as computed over all phonemes left of V in the utterance	X				
5. Values of F of the vowels immediately preceding and following V	X	X	X	X	X
6. Value of F of the vowel immediately preceding V	X	X	X	X	X
7. The maximal value of F which appeared immediately left of V		X	X		
8. Just for duration: A complex estimate of speaking rate which takes intrinsic duration into account (Wightman, 1992)	X				

An "X" in one of the feature columns indicates that the normalization was indeed applied to that feature.

appears, and word class of the word in which it appears (Wang et al., 1996). We have not attempted to normalize DURATION for word class, mainly because this factor is already taken into account by the fact that we defined the vowels in a number of function words as unstressed. The factor 'position in the utterance' is, at least to some extent, covered by the normalizations in which only the left context is taken into account. It is evident that we cannot accurately model the effects of phrase final lengthening in this way. Although this does increase the unaccounted variance, the number of utterance final syllables is small relative to the non-final ones. However, the main factor we wanted to compensate for was articulation rate. Fig. 1 shows that articulation rate, computed by dividing the number of phonemes by the total duration of an utterance (excluding non-speech), in our material roughly varies between 10 and 15 phonemes per second. The normalizations for duration listed in Table 1 attempt to compensate for this variance.

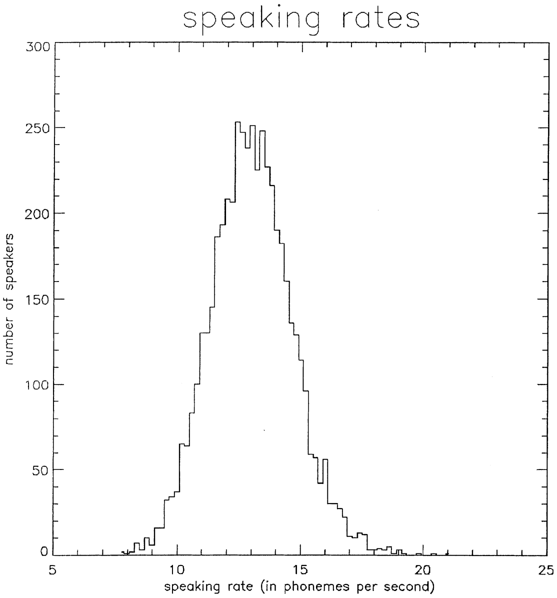

Fig. 1. Distribution of speaking rate in our material.

Apart from the normalizations which were applied to all features, Table 1 shows three normalizations which were applied only to duration. Normalizations 2 and 4 were added here because we wanted to compare two kinds of computations for speaking rate. In normalizations 1 and 3 speaking rate is computed over the durations of the vowels in the utterance, whereas normalizations 2 and 4 also use the durations of the consonants to compute speaking rate.

Normalization 8 is a complex estimate of speaking rate developed by Wightman (1992). The normalization takes the mean and variance of the duration of a vowel into account, while also compensating for the speaking rate in the utterance.

Normalized energy. The energy of a vowel is known to be dependent on factors like the degree of openness (Lehiste and Peterson, 1959), and the position in the utterance (especially vowels following the last sentence accent are known to have a much lower energy than the vowels preceding the last accent; Pierrehumbert, 1994). The overall effort with which an utterance is produced can vary within a broad range. Furthermore, the telephone channel may have influenced energy; there is the possibility that some of the switches that connect the digital trunk network to the analog local loop apply automatic gain-control (AGC), although we have not seen clear indications in our signals that AGC has been applied. We defined several normalizations for energy, which were all applied to both the MAXENE and TOTENE features. They are listed in Table 1.

The normalizations for energy are mostly similar to those for duration, including normalizations for the average energy in the utterance and several conditions for varying context-windows.

Normalization 7 is similar to normalization 6, but it normalizes an energy feature for the value of the previous local maximum of that feature in the utterance. Often, but not always, this value will be identical to (and stemming from) that of the previous vowel in the utterance. However, normalization 7 does not require that the location of the previous vowel is known, so it could theoretically be used in a classifier which does not have this information.

Normalized spectral tilt. According to Sluijter and Van Heuven (1996) the spectral slope of stressed vowels will tend to be flatter than the spectral slope of unstressed ones. The explanation for this would be that stressed vowels are produced with more effort than unstressed vowels, and this extra effort will yield a higher energy-increase in the higher spectral bands than in the lower spectral bands. So, like the other features, tilt will also be context-sensitive. Therefore, the same normalizations which were applied to the energy-features were also applied to the tilt-features.

2.3.3. Free normalizations vs. calculated normalizations

We have attempted to automatically separate stressed and unstressed vowels by means of linear classifiers, based on individual raw features, individual normalized features, as well as on a large number of sets comprising combinations of features. In addition to the (sets of) *calculated* normalized features mentioned above, we have also performed classification experiments in which the same underlying set of raw features was used (*free normalizations*). To give an example: a *calculated* normalization for the MAXENE of a vowel is computed by dividing the MAXENE of that vowel by the MAXENE of the vowel immediately preceding it. In the *free* normalization corresponding with this, simply both the MAXENE of the vowel and the MAXENE of the vowel immediately preceding it are fed into the classifier. The difference with the case of *calculated* normalized features is that we now leave the optimum use of the data for normalization purposes to the classifier, instead of imposing a predetermined, and deterministic use of the data. In the results section we will return to this difference between *calculated* and *free* normalizations.

2.4. Statistics

We have carried out t-tests on the feature values of the vowels in the training set, to check whether the sample means for stressed and unstressed vowels differ at all. We also looked at the distributions of the feature values in the two classes. Next, we computed the correlation between those features, to check for interdependencies that should be taken into account when combining individual features to improve the separation of the classes. Detailed results of these statistical tests are given in Section 3.1.

2.5. The classifier

We have performed a large number of tests to investigate the extent to which stressed and unstressed vowels can be distinguished. In all experiments the identity of the vowels is assumed to be known. This condition holds in HMM and artificial neural network (ANN) recognizers, which compute the likelihood of top–down generated hypotheses about the words that were spoken. For these experiments we use the same simple Bayesian classifier as Ying et al. (1996) and Waibel (1986). We assume that the features can be jointly modeled by an N-dimensional normal distribution for each of the two classes (stressed and unstressed vowels).

Then the likelihood $p(x|\omega_{\mathrm{S}})$ that a specific vowel from the test set is stressed will be given by

$$p(x|\omega_{\mathrm{S}}) = (2\pi)^{-N/2}\left|\Sigma_{\mathrm{S}}\right|^{-1/2}\exp\left[-\frac{1}{2}(x-\mu_{\mathrm{S}})^{\mathrm{T}}\Sigma_{\mathrm{S}}^{-1}(x-\mu_{\mathrm{S}})\right], \qquad (1)$$

where N is the number of features, Σ_{S} the covariance-matrix over the feature vectors from all stressed vowels in the training set, μ_{S} the mean over the feature vectors from all stressed vowels in the training set, x the feature vector belonging to the vowel from the test set.

The likelihood $p(x \mid \omega_{\mathrm{U}})$ that a specific vowel from the test set is unstressed will be given by

$$p(x \mid \omega_{\mathrm{U}}) = (2\pi)^{-N/2}\left|\Sigma_{\mathrm{U}}\right|^{-1/2}\exp\left[-\frac{1}{2}(x-\mu_{\mathrm{U}})^{\mathrm{T}}\Sigma_{\mathrm{U}}^{-1}(x-\mu_{\mathrm{U}})\right]. \qquad (2)$$

Since in this study we will ensure that the priors are equal, we can leave them out of the equations, and then our stress classifier reduces to a maximum likelihood classifier which can be written as

$$(x-\mu_{\mathrm{S}})^{\mathrm{T}}\Sigma_{\mathrm{S}}^{-1}(x-\mu_{\mathrm{S}}) + \log|\Sigma_{\mathrm{S}}| < (x-\mu_{\mathrm{U}})^{\mathrm{T}}\Sigma_{\mathrm{U}}^{-1}(x-\mu_{\mathrm{U}}) + \log|\Sigma_{\mathrm{U}}|. \qquad (3)$$

3. Results

In this section we summarize the results of our experiments.

3.1. Statistics

Lexical statistics. Although we do not use the prior probability for a vowel to occur in a stressed syllable, it is still worthwhile investigating whether each vowel has roughly the same probability of being stressed. Phonetic arguments seem to support the hypothesis that diphthongs (which might be considered as two vowels in a single syllable) and long, open vowels can be expected to be inherently more conspicuous than short, closed vowels. Thus, it is interesting to see whether there is a tendency for stressed syllables to contain ‘heavy’ instead of ‘light’ vowels. We have investigated three databases of words, viz. the 10,000 Polyphone sentences (Den Os et al., 1995a) comprising the training and test sets in the present study, the frequency lexicon of the SpeechStyles corpus (Den Os et al., 1995b) and the CELEX lexical database (Baayen et al., 1993) which contains data on the frequency of use of the words. CELEX is based on word counts in over 40 million words of text, and on the largest general purpose dictionaries of the Dutch language. Thus, it does not reflect word frequencies in a specific domain. The Polyphone corpus is not too different from CELEX, in that it is composed of 25,000 unrelated sentences, extracted from a general newspaper (and selected so as to maximize phonemic coverage). SpeechStyles, on the other hand, only contains texts that refer to domestic issues and eating habits; therefore, its vocabulary is likely to be much more domain specific.

Fig. 2 shows the results. It is clear that diphthongs and long vowels are much more likely to occur in a stressed syllable than short vowels. The *rankings* of the vowels in CELEX and Polyphone are very similar; and even the potentially much more domain specific SpeechStyles corpus shows the same trend. The relatively high percentage for the stressed medium-long vowel /y/ [1] (as in the Dutch word *vuur*) in the SpeechStyles corpus is due to its occurrence in several mono- and bi-syllabic words that are specific for the domestic/food domain.

As can be seen in Fig. 2 the prior probabilities can have a considerable effect on the classifications which are computed according to Eq. (3). To avoid any effect of prior probabilities in our

[1] Throughout this paper SAMPA notation is used.

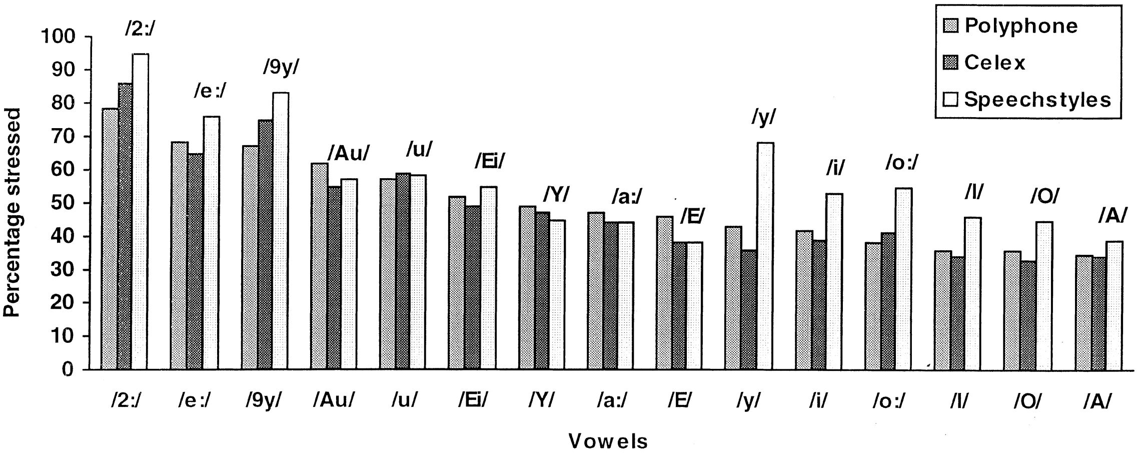

Fig. 2. The a priori probability for each vowel that it is stressed in three different databases.

classification results, we randomly deleted items from the bigger of the paired sets. For example: the test set contained 2865 stressed exemplars of the vowel /a:/ (as in the Dutch word *laan*), and 3165 unstressed ones. For this vowel 300 randomly chosen unstressed samples were removed from the test set.

Distributions of raw features. Fig. 3(a) shows the distributions of the raw feature values for the vowel /9y/ (as in the Dutch word *huis).* It is ob-

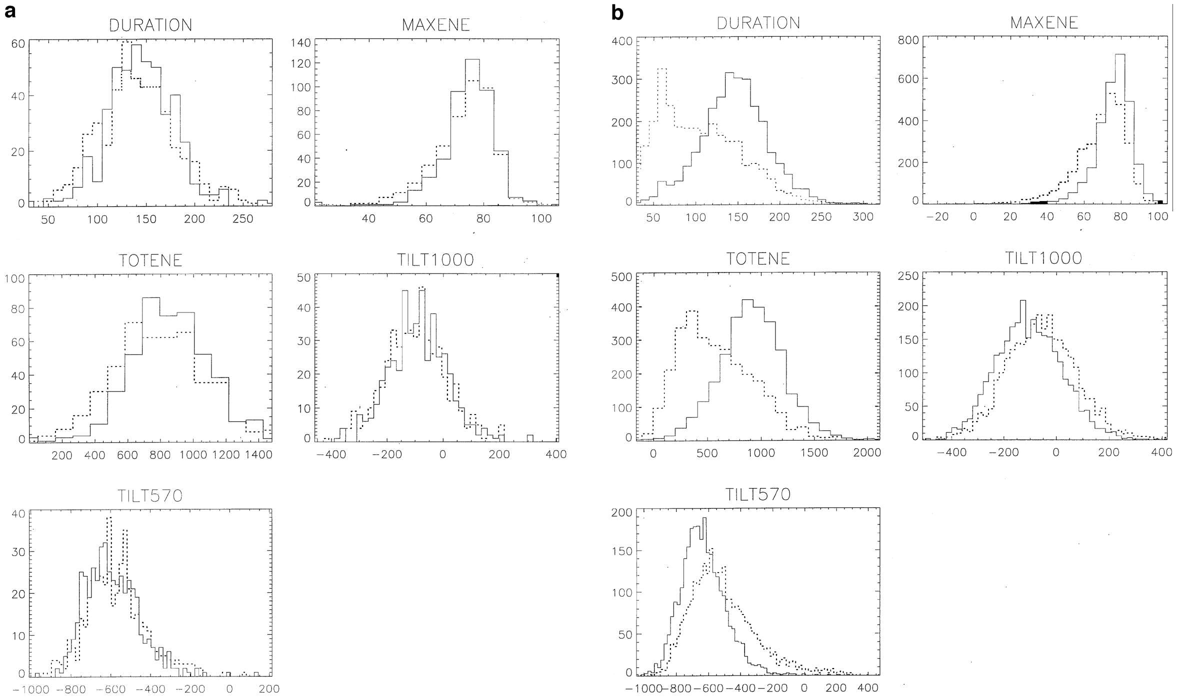

Fig. 3. (a) Distributions for the vowel /9y/. The dotted lines are for unstressed realizations. (b) Distributions for the vowel /a:/. The dotted lines are for unstressed realizations.

vious that the distributions of stressed and unstressed vowels overlap heavily. So, *t*-tests for the means of the distributions yield only significant differences for TOTENE.

However, from Fig. 3(b) it can be seen that for the vowel /a:/ much better separation is possible. In any case, the distributions for the vowel /a:/ seem easier to model by a Gaussian distribution. For this vowel the *t*-tests for all features were significant at the 5% level (with correction for the number of *t*-tests by the Bonferroni procedure). To summarize over all vowels:

- *t*-tests for DURATIONs of individual vowels were significant at the 5% level for all vowels, except /9y/.
- *t*-tests for MAXENEs of individual vowels were significant at the 5% level for all vowels, except /9y/, /Ei/ and /Au/.
- *t*-tests for TOTENEs of individual vowels were significant at the 5% level for all vowels.
- *t*-tests for TILT1000s of individual vowels were significant at the 5% level for all vowels, except /9y/, /Ei/, /Au/, /Y/, /O/, /e:/, /o:/and /i/.
- *t*-tests for TILT570s of individual vowels were significant at the 5% level for all vowels, except /9y/ and /u/.

So TOTENE seems to yield the best separation between stressed and unstressed vowels, and TILT1000 the worst. For the vowel /9y/ separation is hardest. The difference between /9y/ and the other vowels may be due to the special status of this vowel in Dutch constructions like ‘stelde UIT’ (postponed). In fact the word ‘UITstellen’ (to postpone) has lexical stress on UIT, but when it is used in its separated form, the morpheme ‘uit’ is a word which will be considered unstressed, because it is in our list of Dutch function words. For all vowels except /9y/ and /Ei/ the stressed version has a significantly higher MAXENE than the unstressed version. The feature TOTENE is significantly different for all vowels.

Fig. 3(b) clearly shows that even for /a:/ the distributions overlap considerably. Therefore classification of vowels as stressed and unstressed will be very hard on the basis of individual raw features. Perhaps cleverly selected combinations of these features, and proper normalizations for their context might give good classification results.

Correlations between raw features. To prepare the work on combining features to improve classification, we first look at the correlations between the raw features. If two features are highly correlated, little additional classification power may be expected from their combination.

We computed the Pearson-*r* correlations between the raw features for each individual vowel. Table 2 shows these correlations for two vowels. In general TOTENE correlates highly with

Table 2
Correlations between raw features for the vowel /a:/ and the vowel /9y/

	Vowel /a:/				
Feature	DURATION	MAXENE	TOTENE	TILT1000	TILT570
DURATION	1	0.3731	0.9024	−0.2824	−0.4953
MAXENE		1	0.6224	−0.4838	−0.7930
TOTENE			1	−0.4176	−0.6846
TILT1000				1	0.7361
TILT570					1

	Vowel /9y/				
Feature	DURATION	MAXENE	TOTENE	TILT1000	TILT570
DURATION	1	0.2335	0.7725	−0.1559	−0.3744
MAXENE		1	0.5847	−0.5043	−0.7710
TOTENE			1	−0.3291	−0.6196
TILT1000				1	0.7026
TILT570					1

DURATION, while the correlation between MAXENE and DURATION is lower. This pattern is to be expected, since TOTENE combines contributions of MAXENE and DURATION. The relatively high (negative) correlations between MAXENE and TILT570 are also understandable from the way these features are computed. In general we can see that even features which are computed completely independently, like DURATION and TILT570, tend to be more or less correlated, thereby indicating that for instance an increase in duration tends to coincide with a decrease in spectral tilt. Note also that the correlations tend to be higher, in an absolute sense, for /a:/ than for /9y/.

3.2. Classification experiments

In these experiments we trained the Bayesian classifiers with features from the vowels in the training set, and we tested them with the vowels in the test set. Three sets of classification scores were obtained, viz. the proportion correct for unstressed vowels, for stressed vowels and for the full set of vowels. For comparison we also tested the classification performance on the training set. The results of these last tests were, within the confidence intervals, the same as for the test set. This shows that our classifiers showed good generalization behavior from the training set to unseen vowels. In what follows only classification results obtained on the test set are discussed.

General. The best result for classification with a single feature over the individual vowels was 72.6% correct for the vowel /a:/ (see Table 3). This was achieved by normalizing TOTENE for the average energy over the vowels in the utterance. However, this normalization gave no significantly better ($p < 0.05$) classification results than the raw feature TOTENE (70.4%).

Differences between features. The results showed that no one feature or combination of features will always give the best classification performance. But the feature TOTENE, or one of its normalizations is in many cases (10 out of 15) the best.

Differences between vowels. There is a considerable difference (15% points) between the best classifier for the vowel /a:/ (the highest scoring vowel), and the best classifier for the vowel /e:/ (the lowest scoring vowel). This pattern is not surprising if one looks at the distributions of the raw

Table 3
Best classification result for each vowel

Vowel	Description of winning feature	Percentage correct classifications			Confidence interval
		Stressed	Unstressed	Total	
a:	TOTENE, free, 1	76.05	69.14	72.60	71.43–73.77
y	DURATION, calculated, 5	64.48	74.68	69.58	66.83–72.33
o:	TOTENE, free, 1	71.43	66.50	68.97	67.48–70.46
A	TOTENE, free, 6	67.34	66.29	66.81	65.56–68.07
E	TOTENE, free, 7	65.61	67.51	66.56	65.31–67.82
Au	DURATION, free, 6	74.00	54.50	64.25	61.18–67.33
O	TOTENE, free, 7	66.00	60.70	63.35	61.75–64.96
Y	MAXENE, free, 5	79.12	46.73	62.93	60.51–65.35
2:	MAXENE, calculated, 3	68.59	55.78	62.19	57.83–66.55
i	TOTENE, free, 6	61.67	62.61	62.14	60.60–63.69
9y	TOTENE, free, 5	66.17	57.14	61.65	58.54–64.78
9y	TILT570, free, 7	59.03	64.28	61.65	58.54–64.78
I	TOTENE, free, 5	52.03	70.58	61.31	59.86–62.76
u	TOTENE, calculated, 3	60.53	57.74	59.13	56.69–61.59
Ei	TOTENE, free, 5	44.87	72.03	58.45	56.85–60.06
e:	DURATION, calculated, 6	43.38	70.95	57.16	55.54–58.79

Since all ‘best’ results were obtained by normalized features, the second column indicates on which raw feature the normalization was based, whether the normalization was *free* or *calculated*, and the index of the kind of normalization as introduced in Table 1. The /9y/ appears two times because two features had the same score.

features, where the stressed and unstressed distributions show considerable more overlap in /e:/ than in /a:/.

Normalized versus raw features. For 12 of the 15 vowels the best performing raw feature also provided the winning normalized feature. But although for all vowels the best classifier was one which was trained on normalized features, the performance was often not significantly better than that of the best classifier trained on raw features. The largest difference was obtained for the /9y/, where the raw feature TOTENE gave a 54.51% correct classification, while the normalized winner reaches 61.65% correct classification; a significant difference ($p < 0.05$) of 7.14% points. Moreover, the complex normalization for DURATION proposed in (Wightman, 1992) did not perform better than much simpler approaches (a simple normalization for the duration of the previous vowel was significantly better ($p < 0.05$) than the raw DURATION whereas the Wightman-normalization was not).

Differences between normalization-methods. The free normalizations often give good classification results; 11 out of 15 of the winning normalizations are free. Sometimes there are considerable differences in performance between the deterministic normalization and the free normalization: The most extreme example is the vowel /Ei/, where the best classification result (58.5%) is achieved by a free normalization which is much (and also significantly) better than the result for the deterministic normalization (52.49%). The differences between free and deterministic normalization are not always equally large; there are even occasions where the deterministic normalization performs better, but when the differences are large it is always the free normalization which is the winner. This pattern indicates that our calculations for the normalizations were not the optimal way to use the information at hand.

Combining features. The above results were all obtained with classifiers that were trained on a single feature. We also experimented with combinations of features. Several different combinations of energy, duration, and tilt features were used, but only for the vowel /I/ did this yield a classifier which significantly outperformed the best single feature classifiers for this vowel. This classifier was trained on the free normalizations of normalization type 5 for the features DURATION, MAXENE, and TOTENE. It gave a classification result of 64.46% correct. The fact that combinations of features in general did not give better classifiers can be explained by the fact that there is a considerable correlation between some features (DURATION and TOTENE on the one hand, and TILT570 and MAXENE on the other hand), and that one of these correlated pairs sometimes tends to be much better as separator than the other (most extreme for /a:/, where a normalization for TOTENE scores 72.60% correct, and the best normalization for MAXENE 62.28% correct; a difference of more than 10% points).

Since the experiments with combinations of features did not bring us much, we will not discuss these any further, but concentrate on the outcomes of the single feature classifiers.

4. Discussion

The most obvious finding in this study is that our classification scores for stressed and unstressed vowels are far worse than what has been obtained before, e.g. by Ying et al. (1996) and Sluijter and Van Heuven (1996). Yet, before embarking on an attempt to explain this discrepancy, it is interesting to note that the feature TOTENE, which is introduced in this paper, turns out to be the best discriminator for ±stress. A similar feature was found to provide good performance on the same task in (Waibel, 1986). This finding corroborates the intuitive notion that vowels tend to be perceived as stressed when they are longer, louder, or both longer *and* louder than their unstressed counterparts. Because TOTENE captures both variations in loudness and in duration, it is the best candidate for a feature which can be applied to all vowels to discriminate between stressed and unstressed exemplars. For short vowels (which in Dutch cannot vary much in their duration) it will capture the variance in the energy-levels, and for intrinsically louder vowels it will capture the variance in the duration.

As said before, overall the results for correct classification of vowels in read sentences recorded over the telephone as ±lexical stress are somewhat disappointing, especially when they are compared to previous findings for Dutch stress-minimal pairs (Sluijter and Van Heuven, 1996) and results in similar experiments for English (Ying et al., 1996). On the other hand, our results are in line with what one should expect from previous research that attempted to exploit the feature ±stress in automatic speech recognition, where the results were always equivocal at best (Hieronymus et al., 1992; Dumouchel and O'Shaughnessy, 1993; Van Kuijk et al., 1996).

The single most important difference between the studies that showed large, systematic differences between stressed/unstressed vowels and the studies that did not find the same results is the type of speech material on which the experiments were based. Sluijter and Van Heuven (1996) used high-quality, carefully recorded wide band speech. The speakers had to produce the stress-minimal pair KAnon–kaNON, [2] or a reiterant version of those two words, in a carrier sentence which was identical for all items. This kind of experiments can help in understanding which acoustical correlates play a role in the perception or production of a stressed vowel, but the artificial character of the speech material used makes generalization of the findings to 'read' or 'spontaneous' speech difficult. By limiting the measurements and features to differences between exactly the same sounds, in exactly the same left and right phonetic context, in closely controlled different prosodic contexts, the bulk of the variation occurring in more realistic conditions has been removed. This type of research may very well contribute to improved speech synthesis (where the conditions for generation are likely to be known a priori), but it is doubtful whether the type of deterministic knowledge that is obtained from these studies can at all be used in ASR, where the conditions are to be determined, instead of given a priori. It is our goal to study the phenomenon stress in continuous speech. In our speech material the vowels occurred in random, uncontrolled phonetic contexts, and in a wide range of prosodic contexts. This cannot but add tremendously to the amount of variation in any acoustic phonetic feature, and the results of the 'natural' amount of variation are evident: there is such a large degree of overlap between stressed and unstressed vowels that straightforward bottom–up separation becomes very difficult. So, we believe that the results of our study give a better estimate of what stress detection may contribute to automatic speech recognition than the previous studies, which were based on somewhat contrived speech material.

Of course we have to ask ourselves whether the fact that we used telephone speech instead of wide band speech may also have played a role. An important difference is that telephone speech is restricted to 300–3300 Hz. So any information in the lower and higher regions of the spectrum is lost. However Sluijter and Van Heuven (1996) reported that the effect of stress on the spectral tilt is mainly to be found in the higher frequency range (>500 Hz), and that the bands below 500 Hz are hardly affected by stress. The highest frequency band used in their study had a high cut-off frequency of 4000 Hz; there are no indications that the very upper part of the frequency range was decisive for the superior role of spectral tilt in separating stressed/unstressed syllables. So the fact that we used telephone speech should be no problem for spectral tilt. On our measures for duration the telephone channel can have no effect.

The better results reported in (Waibel, 1986) for American English were obtained in a study in which the labeling of vowels as 'stressed' or 'unstressed' was based on perception experiments, and not on abstract linguistic stress like in the present study. In Waibel's material the stressed syllables were labeled (and thus perceived) by humans as stressed, and therefore we can expect that these vowels acoustically stand out by being longer, louder, or having a pitch accent. A second explanation has to do with differences between English and Dutch. In English unstressed vowels

[2] This example is anyway debatable: Although the transliteration of both words is /ka:nOn/, it is perfectly legitimate to reduce the /a:/ in kaNON to an /A/ or even to a schwa, in which case one would be comparing acoustic differences between two different vowels, and not differences between ±stress within one vowel.

tend to be much more reduced than in Dutch, so acoustical differences will be larger for English. An experiment is under way in which we will investigate the performance of a classifier trained and tested on vowels that were perceived as stressed or unstressed by a group of listeners.

It is interesting to discuss our results in the context of the findings of Van Bergem (1993). He studied differences between vowels in C_1VC_2 syllables for equal C_1 and C_2 in six conditions: in stressed or unstressed syllables of a content word and in function words, each occurring in a position that was or was not affected by the presence of a pitch accent. Although Van Bergem concludes that ‘accent’ contributes less to the differences between vowels than stress and word type, it still strikes the eye that he ends up with four conditions that differ significantly. Vowels in the stressed syllable of an accented word differ from stressed vowels in the same syllable when the word does not receive an accent. And unstressed vowels in content words which do carry an accent differ from their unstressed counterparts in non-accented content words and function words. Thus, it appears that ‘accent’ has a non-negligible effect on the acoustic realization of vowels, even if the vowel occurs in unstressed position in the accented word. Again, this finding should help to improve the quality of speech synthesis, where the location of the accents is determined by the linguistic pre-processor. In speech recognition (and speech perception, for that matter) accent positions are not known in advance.

Thus, the most likely explanation for our relatively low classification scores is the large number of phonetic features involved in the realization of the abstract feature ‘lexical stress’, combined with the large number of linguistic and phonetic factors which play a role in the mapping from abstract lexical stress to the acoustic phonetic surface form. This problem is especially apparent in languages like Dutch and German, that easily form nominal and verbal compounds. According to linguistic theory, each compound word has just one syllable that carries lexical stress. But in many cases the syllable(s) that carry lexical stress in the other members of the compound may be actually realized with at least the same amount of phonetic stress as the vowels in function words (or, for that matter, the vowels in content words that happen not to carry a pitch accent). Also, phonological rules that predict stress shifts when two syllables with lexical stress (eligible for pitch accents) are in adjacent positions make one doubt whether the relation between abstract lexical stress and concrete phonetic realization can be untangled to such an extent that effective bottom–up stress detection becomes feasible.

The very large range of variation in all acoustic phonetic features observed in uncontrolled speech, due to a very large number of phonetic and linguistic effects and contexts, is also the best candidate for explaining why normalization of the features did not improve the classification results to the same extent as was found in previous studies on more tightly controlled speech material. Normalization should be more effective if one knows what effects should be ‘normalized away’. In arbitrary sentences, read by a very large number of speakers, there appear to be too many unknown factors that play a role to allow the usual normalizations to be effective. This interpretation is corroborated by the finding that in most cases a ‘free’ normalization outperformed a deterministic use of the normalization context.

An interesting question which arises is of course: What role can lexical stress play in human speech recognition and understanding when stressed vowels are acoustically so very similar to unstressed vowels? After all, it seems that stress will be hard to detect bottom–up from the speech signal, so what use does it have then? We think that lexical stress is used as a potential carrier for sentence accent. In practice, there will be no large acoustic differences between the unstressed and stressed vowels for most words in a sentence, except for those words which are in focus, and should in some way be conspicuous according to the speaker. So the syllable carrying primary lexical stress is that syllable whose acoustic properties (in terms of duration, energy, spectral tilt, and pitch) will change most if the word is accented in the sentence. The other syllables of the word will probably also be affected, but never so much as the stressed syllable. If it should happen that an unstressed syllable gets more conspicuous than the one marked for stress (as can happen when people speak a foreign language), listeners will experience that as a mispronunciation. So there is no rule that the stressed vowel should be ‘stronger’ than its unstressed counterparts in the same word, but there is certainly a rule that – with the exception of the use of contrastive stress – the unstressed vowels may never be stronger than the stressed one. Nevertheless, we also saw that strong (intrinsically longer and louder) vowels tend to be more often stressed than not. So the “choice” of the syllable which gets lexical stress in a word is certainly not random, but governed by phonetic arguments.

Our findings suggest that for Dutch little or nothing is to be gained from the integration of a lexical stress detector in a single pass decoder in an automatic speech recognition system. This does not imply, however, that the feature lexical stress could not play a useful role as an additional knowledge source in a multi-pass decoder, where it could be used to rescore the likelihood of competing solutions, provided of course that suitably trained models are available to capture the acoustic (and perhaps also prosodic) effects of lexical stress in the context of higher level prosodic and syntactic contexts.

5. Conclusions

The main conclusion of this paper is that the acoustic properties of stressed and unstressed vowels in (telephone) speech, based on a linguistic definition of lexical stress, are not very different. Experiments with linear classifiers showed that the best classification result we could obtain is 72.6% correct classification for the vowel /a:/.

Our classification results are low compared to what has been found in previous studies. We believe that this is due to two reasons. First, previous studies have only compared vowels in identical phonetic contexts, produced under carefully controlled reading conditions. This eliminates most of the variation that caused the feature distributions in our experiment to overlap to a very large degree. Second, at least some of the previous studies seem to have confused the notions of stress and accent, so that the claimed differences between stressed and unstressed vowel are in fact often differences between vowels in accented and unaccented words.

There are significant differences in duration, energy and tilt between the unstressed and the stressed variants of most of the vowels in *t*-tests, but the explained variance of each of these features for stressed and unstressed is low. We hoped that normalizations of the raw features for contextual effects, and combinations of the normalized features could yield better classification results, but this was not the case. Normalizations can help a bit, but not much. Combining features does not help much either. This is probably due to the considerable correlations between the raw features.

The best feature turned out to be a new feature which integrates over the energy in a vowel. In this way it represents a combination of Duration and Energy. Spectral tilt was the worst feature in these experiments. So, contrary to the findings of Sluijter and Van Heuven (1996), we find that for most vowels the energy is a better discriminative acoustic correlate of stress than spectral tilt or duration.

The normalized versions of the features yield better classification results than the raw features, but the improvement is smaller than what one might expect. With respect to normalization it seems preferable to feed the classifier with the raw features on which the normalizations are based than computing deterministic normalizations.

Our findings suggest that for Dutch little or nothing is to be gained from the integration of a lexical stress detector in a single pass decoder in an automatic speech recognition system. However, the feature lexical stress could still play a useful role as an additional knowledge source in a multi-pass decoder, where it could be used to rescore the likelihood of competing solutions.

Acknowledgements

The authors would like to thank Dick Van Bergem, Catia Cucchiarini, Henk van den Heuvel, Judith Kessens, Helmer Strik, and Johan de Veth from the A^{2}RT-group of the Department of Language and Speech for their comments on earlier versions of this paper. We also thank the people at the Max-Planck-Institute for Psycholinguistics in Nijmegen for stimulating discussions.

References

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

den Os, E.A., Boogaart, T.I., Boves, L., Klabbers, E., 1995a. The Dutch Polyphone corpus. In: Proc. Eurospeech-95, pp. 825–828.

den Os, E.A., Ellens, M., in 't Veld, C., Boves, L., 1995b. Some figures concerning the transliteration of the Dutch Speech-Styles corpus. In: Proc. XIIIth Internat. Congress on Phonetic Sciences, Vol. 3, pp. 536–539.

Dumouchel, P., O'Shaughnessy, D., 1993. Prosody and continuous speech recognition. In: Proc. Eurospeech-93, pp. 2195–2198.

Hieronymus, J.L., McKelvie, D., McInness, F.R., 1992. Use of acoustic sentence level and lexical stress in HMM speech recognition. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 225–229.

Lehiste, I., Peterson, G.E., 1959. Vowel amplitude and phonemic stress in American English. J. Acoust. Soc. Amer. 31, 428–435.

Pierrehumbert, J., 1994. Prosodic effects on glottal allophones. In: Fujimura, O. (Ed.), Vocal Fold Physiology 8. Singular Press, San Diego, CA.

Radeau, M., Morais, J., 1990. The uniqueness point effect in the shadowing of spoken words. Speech Communication 9, 155–164.

Sluijter, A.M.C., Van Heuven, V.J., 1996. Spectral balance as an acoustical correlate of linguistic stress. J. Acoust. Soc. Amer. 100, 2471–2485.

Van Bergem, D., 1993. Acoustic vowel reduction as a function of sentence accent, word stress, and word class. Speech Communication 12, 1–23.

Van Kuijk, D., van den Heuvel, H., Boves, L., 1996. Using lexical stress in continuous speech recognition for Dutch. In: Proc. Internat. Conf. Spoken Lang. Process., pp. 1736–1739.

Van Wijk, C., Kempen, G., 1980. Funktiewoorden – Een inventarisatie voor het Nederlands. ITL Review of Applied Linguistics 47, 53–68.

Waibel, A., 1986. Recognition of lexical stress in a continuous speech system – A pattern recognition approach. In: Proc. Internat. Conf. Acoust. Speech Signal Process., pp. 2287–2290.

Wang, X., Pols, L., ten Bosch, L., 1996. Analysis of context-dependent segmental duration for automatic speech recognition. In: Proc. Internat. Conf. Spoken Lang. Process., pp. 1181–1184.

Wightman, C.W., 1992. Automatic detection of prosodic constituents for parsing. Dissertation, Boston University.

Ying, G.S., Jamieson, L.H., Chen, R., Mitchell, C.D., Liu, H., 1996. Lexical stress detection on stress-minimal word pairs. In: Proc. Internat. Conf. Spoken Lang. Process., pp. 1612–1615.