# ACOUSTIC CHARACTERISTICS OF LEXICAL STRESS IN CONTINUOUS SPEECH

*David van Kuijk* *Louis Boves*

Department of Language and Speech, Nijmegen University, Nijmegen, The Netherlands

## ABSTRACT

In this paper we investigate acoustic differences between vowels in syllables that do or don't carry lexical stress. The speech material on which the investigation is based differs from the type of material used in previous research: we used phonetically rich sentences from the Dutch POLYPHONE corpus.

We shortly discuss the definition of the linguistic feature 'lexical stress' and its possible impact on the phonetic realization. We then proceed to explain the experiments that were carried out and the presentation of the results.

Although most of the Duration, Energy and Spectral Tilt features that we used in the investigation show statistically significant differences for the population means for stressed and unstressed vowels, it also appears that the distributions overlap to such an extent that automatic detection of stressed and unstressed syllables yields accuracy scores of not much more han 65%. It is argued that this is due to the large variety in the ways in which the abstract linguistic feature 'lexical stress' is realized in the acoustic speech signal.

## 1. INTRODUCTION

In CART-like procedures to determine the optimal set of sub-word units many phonetic features have been tried. One potentially important feature, viz. lexical stress, has not received much attention. Yet [1] and [2], [3] contain suggestions that 'stress' might be useful for improving speech recognition performance.

Our previous experiments have shown that blind use of lexical stress distinctions does not significantly improve recognition performance [4]. One way to explain that finding is to assume that the input features (cepstrum coefficients and log-energy) are not appropriate to distinguish stressed vowels from unstressed ones. Specifically, in our decoder duration is not explicitly modeled, while [5], [6] showed that duration is (one of) the most important indicators for lexical stress in Dutch. Dutch being a language that contains phonologically and phonetically long and short vowels requires vowel-dependent duration distributions. Integration of such distributions in a conventional HMM decoder is relatively straightforward.

In addition to duration, energy and spectral tilt have also been observed to be related to lexical stress. However, these features might require clever normalizations to make them useful in ASR.

In the present study we investigate to what extent it is possible to distinguish vowels in lexically stressed syllables from the corresponding vowels in unstressed syllables in continuous speech. In order to do so, we study several different techniques to normalize the raw acoustic feature values.

## 2. BACKGROUND

In this section the theoretical and operational background of research into the use of stress as a knowledge source in ASR is summarized.

### 2.1. Stress versus Accent

In order for a feature like 'stress' to be useful in ASR it must be possible to make unambiguous decisions about its presence or absence in the speech signal. According to modern linguistic theory every word in the (mental, theoretical linguistic) lexicon carries lexical stress. Stress is assumed to be a feature of a syllable, but for practical purposes 'stress' can be attributed to the vowel. Although this approach has been very fruitful in linguistic theorizing, it may not be optimally suited for use in ASR: in the lexicon the theoretical feature 'stress' is not only attributed to the most prominent syllable in words like *tele'phone*, where the stressed syllable is likely to be acoustically different from the surrounding unstressed syllables, but also to short function words like *'the, 'a, 'is*, etc., that are normally pronounced much like the unstressed syllables in *tele'phone*.

'Lexical stress', as a theoretical feature attached to all words in the lexicon, must be distinguished from 'accent'. Accented syllables are only a subset of the stressed syllables in an utterance, and they are (almost) always marked by conspicuous acoustic characteristics, with pitch as the most important one. While lexical stress is predictable (because it is marked in the lexicon) accent is not, because the location of the actual accents realized by a person speaking an utterance depend on the syntactic, semantic and pragmatic structure of the utterance, as well as on its discourse context.

In summary: 'accented' syllables are characterized by relatively conspicuous acoustic attributes, but their location in the speech signal cannot be predicted from just a verbatim transcription. This makes it difficult to use 'accent' in automatic training procedures. 'Stress', on the other hand, is easy to include in automatic training, but its relation to

acoustic characteristics may be so vague and ill-defined that it may turn out impossible to use for its own reasons.

### 2.2. Research on stress detection

Up to now, investigations of the effects of lexical stress on the acoustic properties of vowels have mainly been based on isolated words with contrastive stress (like *'comment* versus *com'ment*) [7]. While [5] and [6] investigated the acoustic properties of stressed and unstressed vowels, [1] and [8] have attempted automatic classifications of stressed and unstressed vowels for American English. [1] achieved correct classifications of up to about 88% on a relatively small database.

In [8] syllables in bi-syllabic stress-minimal word pairs (e.g. *'combine-com'bine* almost perfect classification results were obtained on the speech of a single male speaker for the words embedded in the carrier sentence *Please say .... again*. 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, 5% classification error was obtained with a three-feature classifier (two different normalizations of Duration plus a normailized Energy). Adding three more speakers did not increase the error rate.

These experiments help in understanding which acoustical correlates play a role in the perception and production of stressed vowels, but the artificial character of the speech material used makes generalization of the findings to fluent speech questionable. It is our goal to investigate the phenomenon 'stress' in continuous (though still read) speech. Also, we attempt automatic classification of the vowels as stressed or unstressed, again based on selected acoustic parameters.

## 3. METHOD

For the analyses we used 5000 phonetically rich sentences from the POLYPHONE corpus ([9]) as training material. and 5000 other phonetically rich sentences from this corpus as testing material. Both sets contained utterances of about 550 male and 550 female speakers. The speakers are selected from all Dutch provinces.

We trained our off-the-shelf HMM-recognizer with the 5000 training sentences, and then carried out a forced segmentation on both sets. 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.

### 3.1. 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 feature values were derived from the results of the forced segmentation, in the ways explained below.

#### *3.1.1. Duration*

*Duration* is straightforwardly determined by the forced segmentation. Measurement accuracy is determined by the 10-ms frames. Durations have a minimum of 30 ms, because in our recognizer a minimum of 3 states per acoustic model must be visited.

#### *3.1.2. Energy*

*Energy* is represented in two ways: the maximum energy (MAXENE), and the total energy (TOTENE). MAXENE is defined as the energy of the frame with the greatest energy value. TOTENE is an integration over all frames within that vowel. So TOTENE depends on the duration of the vowel in addition to its maximum energy.

#### *3.1.3. Spectral Tilt*

The spectral tilt of a vowel was determined by comparing the energy in the lower spectral bands with that in the higher spectral bands. The spectral tilt is computed in the frame with the highest energy in the vowel (so the same frame from which the MAXENE is taken). Spectral tilt was expressed in four ways: TILT1000 was computed by subtracting the log- energy in the spectral band from 300 Hz to 1 kHz and the log-energy in the band between 1 and 4 kHz. TILT570 was computed by subtraction of the log-energy in the band between 300 and 570 Hz and the log-energy in the band from 570 Hz to 4 kHz. TILT350 was computed by comparing the log-energy below 350 Hz with the energy in the remaining part of the spectrum. TILT3525 was computed by comparing the energy in the band below 350 Hz with the energy in the spectral band from 2500 to 2900 Hz.

### 3.2. Normalizations

The raw features defined above are known to be highly context dependent. Therefore, we defined several transformations of the raw features, that are all meant to remove the context dependency. No normalization was attempted for the spectral tilt measures. First, it is not obvious how spectral tilt may depend on contexts. Second, the different ways of defining spectral tilt may already account for some normalization. Last but not least, the spectral mean subtraction which is part of the pre-processing also implies a crude normalization of the spectral tilt measures.

#### *3.2.1. Duration*

The duration of a vowel is known to be influenced by many factors, like intrinsic duration, speaking rate, lexical stress, position in word, and word class of the word it is in ([11]). In the light of the claims made by [10] it seems to be possible to normalize for speaking rate. In addition to the simple normalization of vowel duration by the average phoneme duration in the utterance we also performed a more complex normalization proposed by [12]

#### *3.2.2. Energy*

The energy of a vowel is also known to be dependent on factors like the degree op openness, 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).The overall effort with which an utterance is produced can vary within a broad range. We defined several normalization procedures, by comparing the energy of a vowel to the energy of its left neighbor, to its right neighbor, to the average energy of all vowels to its left

in the utterance and to the average energy of all vowels in the utterance.

## 4. STATISTICS

We have performed a large number of tests to investigate the extent to which individual features or combinations of features can help to distinguish stressed and unstressed vowels. First, we have carried out $t$-tests, to check whether the sample means for stressed and unstressed vowels differ. This simple test was limited to individual features.

Next, we have attempted to automatically separate stressed and unstressen vowels by means of linear classifiers, based on individual features as well as on a large number of sets comprising two or three features. Classification experiments were carried out for individual vowels and for the complete set of vowels, without taking vowel identity into account.

#### *4.0.3. The classifier*

For our experiments we use a simple Bayesian classifier as used in [1] and [8]. We assume that the features can be jointly modeled by a N- dimensional normal distribution for each of the two classes (stressed and unstressed vowels). Since in this study we are not really interested in the priors we can leave them out the equations, and then our classifier reduces to a maximum likelihood classifier.

## 5. RESULTS

In this section we summarize the results of our experiments.

### 5.1. $t$-tests

For the unnormalized features we first determined the $t$ values. The results of the $t$-tests showed that for all vowels except / y/ the stressed version has a significantly longer duration than the unstressed version ($p = .05$), and that for all vowels except /o /, / y/, and / i/ the stressed version has a significantly higher MAXENE than the unstressed version. The feature TOTENE is significantly different for all vowels. For spectral tilt the majority of the tests also showed significant differences between stressed and unstressed vowels, with the exception of :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| TILT1000 | 2: | 9y | Au | | Y | e: | i | u | |
| TILT570 | 2: | 9y | | I | Y | e: | i | u | y |
| TILT350 | 2: | 9y | Au | I | | e: | i | | y |
| TILT3525 | 2: | 9y | Au | I | Y | e: | | | y |

In general the variance in the raw features accounted for by 'stress' is low. One might have expected to see a closer relation between stress and the normalized versions of the features, but that is not what we found. Specifically, the complex normalization for Duration proposed in [12] did not perform beter than much simpler approaches.

### 5.2. Classification experiments

We attempted automatic classification of stressed and unstressed tokens of all individual vowels, but in this paper we will concentrate on the problem of classifying stressed and unstressed vowels in general, so without knowing the vowel identity, because these data are most interesting for the problem of stress detection. Moreover, perhaps with the exception of the vowel /a/, classification experiments for individual vowels did not yield results that are much better than the scores we obtained for the complete vowel set. In this paper we will only report the most interesting results.

#### *5.2.1. Training on the complete set*

In the first set of classification experiments the classifiers were trained on all vowels in the training set. Classification was attempted on the test set. Three sets of classification scores were obtained, viz. the proportion correct for unstressed vowels, for stressed vowel and for the full set of vowels. The test set contained about twice as many unstressed vowels than stressed ones.

It is interesting to note that some features appeared to be more powerful predictors of either stressed or unstressed vowels than for the other category. MAXENE succeeded in classifying 71% of the stressed vowels and only 47% of the unstressed vowels correctly. Duration, on the other hand, obtained 48% correct for stressed vowels and 72% correct for unstressed vowels. The classifier that performed best on the test set, viz. the combination of TOTENE and TOTENE normalized by comparing it to the value of the left hand neighbor vowel, reached 67% correct for stressed, 65% correct for unstressed and 66% correct for the total vowel set.

#### *5.2.2. Training on 'clear examples'*

Classification scores of approximately 65% correct for arbitrary words in read sentences are much lower than what has been reported in the literature for word pairs like *'comment – com'ment*, even when they are embedded in natural sentences. One of the differences between our material and minimal pairs distinguished by lexical stress might be that the phonetic realization of lexical stress in stress-minimal pairs is much more salient than it is in arbitrary words. Therefore, we decided to run an additional series of experiments, in which we trained the classifier on the 10% most salient stressed vowels and the 10% most obviously unstressed vowels. To find the two extreme sets, we computed

$$log(p(x|S)) - log(p(x|U))$$

for each stressed vowel, and

$$log(p(x|U)) - log(p(x|S))$$

for each unstressed vowel. In other words, for each vowel and for each parameter (set) we computed the difference between the likelihood that this vowel was stressed/unstressed given the original classifier, trained on the complete material. This simple procedure orders the stressed vowels on the 'stress' continuum. The resulting classifiers were then tested on the complete test set of 5000 sentences.

The best result (68% overall, with 43% for stressed and 81% for unstressed vowels) was obtained for TOTENE normalized with respect to the average TOTENE of all vowels in the sentence. TOTENE alone obtains a score of 67.5% correct, but now the discrepancy between the performance on stressed (22% correct) and unstressed (91% correct) is

even bigger. However, normalization of TOTENE by the average value of the preceding vowels in the utterance performs as well as normalization by all vowels (67% overall correct); moreover, its performance is much less biased towards unstressed vowels (52% correct for stressed and 75% correct for unstressed vowels).

## 6. DISCUSSION

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 stress-minimal pairs. However, 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.

The most likely explanation for our relatively low classification scores is the wide range of the phonetic features underlying the realization of the abstract feature 'lexical stress'. 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 the main 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 stress detection becomes feasible.

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. 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.

## 7. CONCLUSION

The main conclusion of this paper is that the acoustic properties of stressed and unstressed vowels, based on a linguistic definition of lexical stress, are not very different. Classification may become a little better if the classifier knows which vowel it is, but even then the best result is 72% correct classification (for the vowel /a/).

There are significant differences in duration and energy between the unstressed and the stressed variants of most of the vowels on $t$-tests, but often not for spectral tilt. So: contrary to the findings of [7] we find that the energy of a vowel for most vowels is a better discriminative acoustic correlate of stress than spectral tilt. In our data TOTENE (a combination of energy and duration) normalized with respect to the preceding syllables in the same sentence seems to be the best feature for automatic stress detection.

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 might appear preferable to feed the classifier with the raw features on which the normalizations are based than computing determninistic normalization.

Using TOTENE (a combination of Duration and Energy) as a feature it was found that combination of features did not improve classification performance considerably.

## REFERENCES


[1] A. Waibel, 'Recognition of lexical stress in a continuous speech system - A patern recognition approach', in Proc. ICASSP-86, pp. 2287- 2290.

[2] P. Dumouchel & D. O'Shaughnessy, 'Prosody and continuous speech recognition', in Proc. Eurospeech-93, pp. 2195-2198.

[3] J.L. Hieronymus, D. McKelvie & F.R. McInness, 'Use of acoustic sentence level and lexical stress in HMM peech recognition', in Proc. ICASSP-82, pp. **I** 225-229.

[4] D. van Kuijk, H. van den Heuverl & L. Bobes, 'Using Lexical stress in continuous speech recognition for Dutch', in Proc. ICSLP-96, pp. 1736-1739.

[5] D. van Bergem, 'Acoustic vowel reduction as a function of sentence accent, word stress and word class on the quality of vowels', Speech Communication, **12**, pp. 1-23, 1993.

[6] A.M.C. Sluijter, 'Phonetic correlates of stress and accent', The Hague (1995), Academic Graphics.

[7] A.M.C. Sluijter & V.J. van Heuven, 'Spectral balance as an acoustical correlate of linguistic stress', J. Acoust. Soc. Am., **100** (4), pp. 2471-2485, 1996.

[8] G.S. Ying, L.H. Jamieson, R. Chen & C.D. Michell, 'Lexical stress detection on stress-minimal word pairs', in Proc. ICSLP-96, pg. 1612- 1615.

[9] E. A. den Os, T. I. Boogaart, L. Boves & E. Klabbers, 'The Dutch Polyphone corpus', in Proc. Eurospeech-95, pp. 825-828, 1995.

[10] J.P. Verhasselt & J.-P. Martens, 'A Fast and Reliable Rate of Speech Detector', in Proc. ICSLP-96, pp. 2258-2261.

[11] X. Wang, L. ten Bosch & L. Pols, 'Integration of context- dependent durational knowledge into HMM-based speech recognition', in Proc. ICSLP- 96, pp. 1073-1076.

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