Lexical Stress in Continuous Speech Recognition

Rogier C. van Dalen, Pascal Wiggers, Léon J. M. Rothkrantz

Faculty of Electrical Engineering, Mathematics, and Computer Science
Delft University of Technology, Delft, The Netherlands
R.C.vanDalen@tudelft.nl

10.21437/Interspeech.2006-597

Abstract

Human listeners use lexical stress for word segmentation and disambiguation. We look into using lexical stress for large-vocabulary speech recognition for the Dutch language. It appears that beside vowels, consonants should be taken into account. By introducing stressed phonemes, and features for spectral bands and the fundamental frequency, we reduce the word error rate by 2.6 %.
Index Terms: speech recognition, lexical stress, Dutch.

1. Introduction

Prosody is an important part of the spoken message structure. The foundation of prosody of many languages is laid by *lexical stress* [1]. Higher prosodic levels attach to the words at stressed syllables [2]. Lexical stress is used by listeners to identify words. Though the orthography does not normally encode stress, English has minimal pairs like *súbject – subjéct*, *trústy – trustée*, and *désert – dessért*. Pairs like *thírty – thirtéen* or *digréss – tígress* differ very little except in the stress pattern.

Even though in English and Dutch stress is not on a fixed syllable of the word, all morphologically simplex words of Germanic origin and many others do start with a stressed syllable. Listeners use this for segmentation of speech into words [3]. English-hearing children appear to associate stressed syllables with word onsets at the age of seven months already [4].

Dutch listeners also use the stress pattern to identify words before they have been fully heard. After hearing the beginning of a word *octo-*, Dutch listeners will have deciphered whether it is *octó-* or *ócto-* and reconstruct *octóber* or *óctopus* [5].

Garden-variety speech recognisers do not use lexical stress, useful though it may be. This paper will look into how lexical stress can be automatically detected, and how it can be used to improve speech recognition.

2. Related Work

There has been research on the acoustic correlates of lexical stress. Sluijter [6] in fundamental linguistic research on the acoustic properties of stress minimal pairs demonstrated that lexical stress in English and Dutch is signalled mostly through duration, formant frequencies, intensity, and *spectral tilt*. The latter is a feature that denotes the energy in high frequency bands relative to the energy in low frequency bands. Van Kuijk [7] examined the acoustic properties of a larger corpus of Dutch telephone speech and found similar results: a combination of duration and spectral tilt was the best predictor for lexical stress.

Of those that have used lexical stress in a speech recogniser [8, 9, 10], only [9] has a performance gain. This is probably what the other authors are after as well; but how this is to be done is not discussed. Van den Heuvel [10] hopes "distinguishing stressed and unstressed vowel models may have a general impact on recognition results."

Notably, none of the previous approaches has taken into account the well-observed influence that stress has on consonants: stressed and unstressed consonants are realised differently [1] and stressed consonants have a longer duration [11]. Consonants are influenced by speaking style in the same ways as vowels are: duration, spectral tilt and formant frequencies (for consonants with a formant structure) [12]. This suggests similar effects can be found for lexical stress on consonants. The closest thing to a rationale for not regarding consonants in automatic lexical stress recognition is the claim that consonants do not carry lexical stress in [9]. This claim is not further motivated, and it will be demonstrated to be incorrect.

3. Model

3.1. Objectives

Since humans use lexical stress in processing speech, modelling it could help speech recogniser performance. We expect the following advantages from using lexical stress.

Phone model accuracy Separating phone models into stressed and unstressed versions may increase predictive strength of the models, improving recognition. For example, unstressed vowels tend to become /ə/[1]. Because the range /ə/–/aː/ is split into into /ə/–/aː/–/ˈaː/, the phone models may become more accurate.

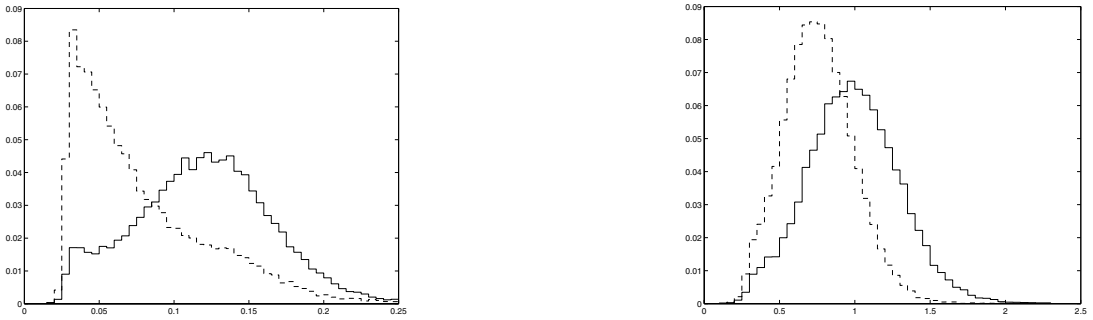
Word segmentation English hearers, when presented with a faint recording "conduct ascends uphill", will reconstruct words starting at stressed syllables, for example, "the doctor sends a pill" [3]. Humans use stress for segmentation; a speech recogniser could use this strategy too.

Word recognition Lexical stress signals differences between words with the same segmental content and different meanings (e.g. Du. *vóorkomen* 'happen' – *voorkómen* 'prevent'), words of different categories (e.g. En. *récord – recórd*), and similar words with different stress patterns (e.g. En. *portráy – pórtrait*).

3.2. Syllables

Lexical stress is specified for syllables as a whole: consonants' specifications for stress must match the vowels' in the same syllable. This can be done by using a consistently stress-marked lexi-

[1] In both Dutch and English.



(a) /aː/ (b) /v/

Figure 1: *Distributions of the duration of stressed (stroked lines) and unstressed (dashed lines) phones (in s).*

con: if it contains both /ˈsˈʌˈbdʒɛkt/ and /sʌbˈdʒˈɛkˈt/, the recogniser would never hypothesise /sˈʌbˈdʒˈɛkˈt/.

In the linguistic literature a difference is made between realisations in the coda and in the onset. For example, English /t/ is pronounced as [t^h] in *táil*, but as [t] in *rétail* and *líght* [1]: /t/ is only aspirated in the onset of a stressed syllable.

(1) (after [1]) relates different levels of linguistic prominence. This article only discusses word stress, which is admittedly a simplification: foot stress and phrasal stress are not taken into account, let alone intonation.

				H				tone	
				|					
(	·			×			)	phrasal stress	
(×	)	(·		×			)	word stress	(1)
(×	·)	(×	·)	(×	·	·)		foot stress	
σ	σ	σ	σ	σ	σ	σ	σ		
ə	lɛŋ	θɪ	ɒ	rə	tɔː	rɪ	əʊ		

3.3. Acoustic representation

We look into acoustic correlates of lexical stress that can be used in a speech recogniser, for example by including them in the feature vectors.

Fundamental frequency From linguistic literature [2] and literature on automatic stress recognition [13] it is expected that pitch is not straightforwardly correlated with lexical stress. The fundamental frequency can be included in a speech recogniser's feature vector though.

Formants Unstressed phonemes can have more reduced realisations than their stressed counterparts. Separating MFCC-based phone models into stressed and unstressed models, whose formant values are confined to smaller areas, will increase MFCCs' ability to recognise the phonemes.

Spectrum The energy in a number of frequency bands can be extracted from the waveform to yield information about the spectral tilt.

Duration Lexical stress is generally found to be correlated with phoneme duration [6, 7, 11]. However, information about phoneme duration is not available during first-pass recognition with Viterbi. Standard HMMs can encode duration through transition probabilities, but this does not work well in recognition. A number of alternatives have been proposed though [14, 15].

Derivatives In [9] it is found that fundamental frequency slope is a better predictor of stress than the raw fundamental frequency. We expect that derivatives for spectral features and the fundamental frequency may be correlated with lexical stress. To capture this correlation, we divide up consonants into those in the onset, and those in the coda.

4. Experimental set-up

We train a speech recogniser on the CGN corpus [16]. 772 recordings are selected for their degree of preparation (only "scripted" recordings are used for consistency so that we do not need different acoustic models for different speaking styles). These 54 842 recordings with a phrase each comprise 775 034 words and almost 53 hours of material. Training data makes up 80 % of the recordings, the test set 10 %, and the evaluation set another 10 %. Every speaker occurs in one set only.

We use the syllable divisions and stress marks from the CELEX lexicon. All phonemes in stressed syllables are marked as stressed, except in function words. All features are normalised over one utterance. HTK includes intensity and MFCC data in the feature vectors. For the energy in spectral bands we use the Linux program *sox* and Praat [17]. [6] chooses spectral bands so that the formants least influence the results; we use the same bands: 0 – 0.5 kHz, 0.5 – 1 kHz, 1 – 2 kHz, and 2 – 4 kHz. We include differences between the spectral bands to capture spectral tilt.

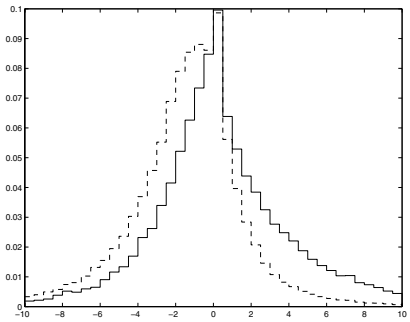
The fundamental frequency is extracted with Praat. Where Praat does not find the fundamental frequency, it is linearly interpolated.

Due to such factors as the overall speaking rate, phone duration is not easily measured objectively. As an approximation of the relative length of a phone by a certain speaker in a specific utterance, it is normalised. Define p_j as a phone from the corpus. p_j is the realisation of the phoneme $i = r(p_j)$; n_i is the number of realisations of the phoneme i. $d(p_j)$ is the actual duration of p_j. U_k is one utterance. The expected duration for a phoneme i is defined as

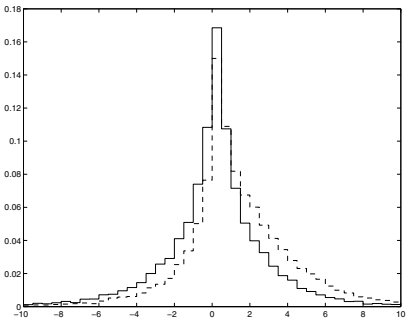
$$\mu_i = \frac{\sum_{p_j : r(p_j)=i} d(p_j)}{n_i}. \tag{2}$$

The normalised duration of p_j is

$$d'(p_j) = d(p_j) \cdot \alpha_k, \quad p_j \in U_k \tag{3}$$



(a) /d/ in the onset



(b) /k/ in the coda

Figure 2: *Distributions of ΔF_0 for stressed (stroked lines) and unstressed (dashed lines) consonants (in Hz).*

where α_k is the speaking rate of utterance U_k, which is defined as

$$\alpha_k = \frac{\sum_{p_j \in U_k} \mu_{r(p_j)}}{\sum_{p_j \in U_k} d(p_j)}. \quad (4)$$

5. Results

Phoneme features are collected by forced-aligning phones with a trained recogniser. Durations are found using the phoneme model boundaries. The feature vectors contain an energy feature, 12 MFCC features, Sluijter's spectral band features, and the fundamental frequency. We collect the averages over the feature vectors for the middle states of HMMs. We perform t-tests on features for stressed and unstressed variants of phonemes. All test are done on significance level 0.01. Most features appear to be significant for many phonemes; we select those that are significant for the most phonemes.

Similarly to [6, 7], we find that duration in general is a good indicator of stressedness. Stressed vowels are consistently longer than their unstressed counterparts (see Fig. 1). This is significant for 92 % of phonemes.

Interestingly enough, and unlike we found in an earlier experiment on a much smaller corpus [18], the fundamental frequency behaves much like we would expect. As Fig. 2 shows, it tends to increase more in onsets of stressed syllables than in onsets of unstressed syllables. This feature is significant for 92 % of consonants.

For many phonemes stressedness correlates well with spectral tilt measures. Sluijter's spectral band B1 (0 – 0.5 kHz) is significant for 89 % of phonemes. First-order derivatives for spectral bands, for example, ΔB4 (2 – 4 kHz) with 93 %, are mostly significant. Difference between spectral bands, which would seem to implement Sluijter's spectral tilt [6] features, do not perform as well. Most interestingly, the features that work for vowels give similar results for consonants.

5.1. Recognition rate

The speech recogniser we use to test our hypothesis — using lexical stress will improve speech recognition rate — is not exactly state of the art. It uses context-independent models for lack of time to train it; for 4.5 % of words no stress-marked transcription is available so they are replaced by *xxx*. However, the purpose of the baseline recogniser is not to recognise speech well, but to be compared to the stress-enabled speech recogniser. Both recognisers differentiate between consonants in onsets and codas.

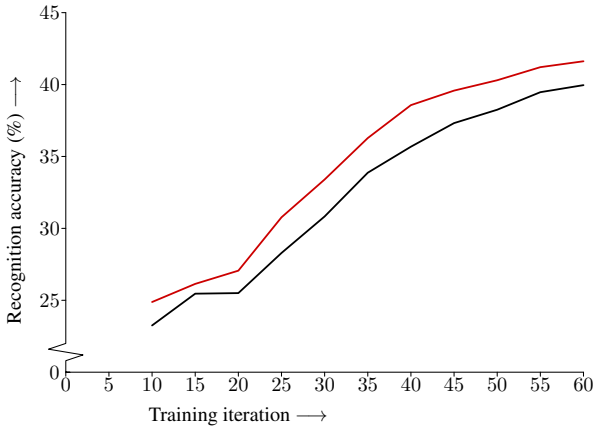


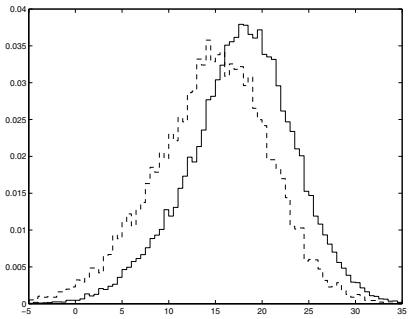
Figure 4: *Recognition rates on the evaluation set while training.*

Fig. 4 shows that during training the stress-enabled speech recogniser consistently performs better than the baseline. Word error rates after 60 training iterations are 56.72 % and 55.27 %; this is a relative improvement of 2.6 %.

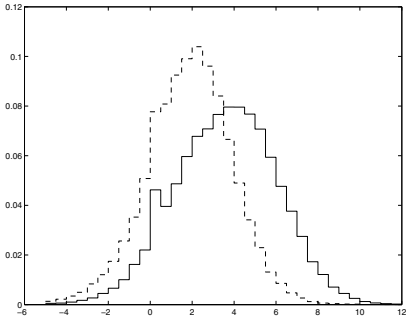
6. Conclusion

This paper has described the importance and the feasibility of using lexical stress in a speech recogniser. That stress works on the syllable level can be effectively modelled by adding stress marks to the phonemes in the speech recogniser lexicon. Lexical stress has been demonstrated to influence acoustically not only vowels, but also consonants. The same features that are canonically associated with stressed vowels (duration, spectral tilt, and intensity) are correlated with stressed consonants. The Viterbi algorithm and standard HMMs cannot use the duration of a phone that is being recognised.

We hoped to improve recognition performance on three accounts: phone recognition, word segmentation and word recognition. We have created a proof-of-concept implementation. The acoustic model does not use phone duration; the feature vectors include all features we could think of; all consonants and vowels,



(a) Energy in B1 (0 – 0.5 kHz) for /iː/.



(b) Derivative of energy in B4 (2 – 4 kHz) for /d/.

Figure 3: *Distributions of spectral band features for (stroked lines) and unstressed (dashed lines) phones.*

without exception, are split into stressed and unstressed variants. The cruel linguistic model — word stress is copied straight from the lexicon — is a poor man's prosody model compared to (1). Still, we have been able to reduce the word error rate by 2.6 %. This is a strong indication that modelling prosody with the right tools is vital for good speech recognition.

7. References

[1] Colin J. Ewen and Harry van der Hulst, *The Phonological Structure of Words*, Cambridge University Press, 2001.

[2] D. Robert Ladd, *Intonational phonology*, Number 79 in Cambridge Studies in Linguistics. Cambridge University Press, Cambridge, 1996.

[3] Trevor Harley, *The Psychology of Language: From Data to Theory*, Psychology Press, Hove, 2001.

[4] Erik D. Thiessen and Jenny R. Saffran, "When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants," *Developmental Psychology*, vol. 39, no. 4, pp. 706–716, 2003.

[5] Nicole Cooper, Anne Cutler, and Roger Wales, "Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners," *Language and Speech*, vol. 45, no. 3, pp. 207–228, 2002.

[6] Agaath Sluijter, *Phonetic Correlates of Stress and Accent*, Ph.D. thesis, Leiden University, 1995.

[7] David van Kuijk and Loe Boves, "Acoustic characteristics of lexical stress in continuous telephone speech," *Speech Communication*, vol. 27, pp. 95–111, 1999.

[8] D. van Kuijk, H. van den Heuvel, and L. Boves, "Using lexical stress in continuous speech recognition for Dutch," *Proceedings* ICSLP, vol. IV, pp. 1736–1739, 1996.

[9] Chao Wang and Stephanie Seneff, "Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain," 2001.

[10] Henk van den Heuvel, David van Kuijk, and Lou Boves, "Modelling lexical stress in continuous speech recognition," *Speech Communication*, vol. 40, pp. 335–350, 2003.

[11] Steven Greenberg, Hannah Carvey, Leah Hitchcock, and Shawn Chang, "Temporal properties of spontaneous speech — a syllable-centric perspective," *Journal of Phonetics*, vol. 31, no. 3–4, pp. 465–485, 2003.

[12] R. J. J. H. van Son and L. C. W. Pols, "An acoustic profile of consonant reduction," *Proceedings* ICSLP, vol. 3, pp. 1529–1532, 1996.

[13] Huayang Xie, Peter Andreae, Mengjie Zhang, and Paul Warren, "Detecting stress in spoken English using decision trees and support vector machines," in *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation*. 2004, pp. 145–150, Australian Computer Society, Inc.

[14] Xue Wang, *Duration modelling in* HMM*-based speech recognition*, Ph.D. thesis, University of Amsterdam, 1997.

[15] Ramachandrula N. V. Sitaram and Thippur Sreenivas, "Incorporating phonetic properties and hidden Markov models for speech recognition," *Journal of the Acoustical Society of America*, vol. 102, no. 2, pp. 1149–1158, 1997.

[16] N. Oostdijk, "The Spoken Dutch Corpus: Overview and first evaluation," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2000, vol. II, pp. 887–894.

[17] Paul Boersma, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341–345, 2001.

[18] Rogier C. van Dalen, Pascal Wiggers, and Leon J. M. Rothkrantz, "Modelling lexical stress," in *Lecture Notes in Artificial Intelligence*, vol. 3658, pp. 211–218. Springer, 2005.