Modelling Lexical Stress

Rogier C. van Dalen, Pascal Wiggers, and Leon J. M. Rothkrantz

Man–Machine Interaction
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
r.c.vandalen@ewi.tudelft.nl, p.wiggers@ewi.tudelft.nl,
l.j.m.rothkrantz@ewi.tudelft.nl

Abstract. Human listeners use lexical stress for word segmentation and disambiguation. We look into using lexical stress for speech recognition by examining a Dutch-language corpus. We propose that different spectral features are needed for different phonemes and that, besides vowels, consonants should be taken into account.

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 use the stress pattern to identify words before they have been fully heard as well. When hearing the beginning of a word *octo-*, Dutch listeners will decipher 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 describe how lexical stress can be automatically detected. It will be determined what features correlate most strongly with lexical stress, considering benefits for 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.

Lexical stress has been used to generate a confidence metric [8]. Of those that have actually used lexical stress recognition in a speech recogniser [9–11], only Wang and Seneff [10] have been able to effect 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 [11] 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 [12]. 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) [13]. 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 [10]. 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 Current speech recognition systems have severe problems coping with speech that is pronounced much faster or slower than the speech it is trained on. Phonemes in unstressed syllables are less often realised canonically than those in stressed syllables. Therefore 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:

1. words with the same segmental content and different meanings (e.g. Du. *vóorkomen* 'happen' – *voorkómen* 'prevent');

[1] In both Dutch and English.

2. words of different categories (e.g. En. *récord* – *recórd*);
3. similar words with different stress patterns (e.g. En. *portráy* – *pórtait*).

3.2 Syllables

Lexical stress is specified for syllables as a whole. This poses a problem for speech recognisers, which typically use phonemes as units. Earlier approaches have circumvented this problem by using only vowels for stress detection. When consonants are included as well, their specification must match the vowels' in the same syllable. This can be done by using a consistently stress-marked lexicon: if it contains both /ˈs ˈʌ ˈb d ʒ ɛ k t/ and /s ʌ b ˈd ˈʒ ˈɛ ˈk ˈt/, the recogniser would never hypothesise /s ˈʌ b ˈd ʒ ˈɛ k ˈt/.

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

3.3 Acoustic representation

To integrate recognition in a speech recogniser, stressedness can be modelled one phoneme at a time. We look into acoustic correlates of lexical stress that can be fed into a speech recogniser, for example by including them in the feature vectors.

Fundamental frequency Stress is typically thought to be connected to pitch. However, from linguistic literature [2] and literature on automatic stress recognition [14] it is expected that pitch is not straightforwardly correlated with lexical stress. Its acoustic correlate, the fundamental frequency, can straightforwardly be included in a speech recogniser's feature vector though.

Formants Unstressed phonemes can have more reduced realisations than their stressed counterparts; this is visible in the formant values. Standard MFCCs should be able to capture this difference. Note that MFCCs do not directly model formants, but frequency bands. Separating MFCC-based phone models into stressed and unstressed models, whose formant values are confined to a smaller area, will therefore 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.

Intensity Overall intensity is generally thought to be associated with lexical stress. However, [6] claims that what is often perceived as loudness variation may actually be spectral tilt: speaking effort would be the common cause.

Duration Lexical stress is generally found to be correlated with phoneme duration [6, 7, 12]. However, information about phoneme duration is not available during first-pass recognition. Standard HMMs can encode duration through transition probabilities, but this does not work well in recognition. A number of alternatives have been proposed though [15–18].

Derivatives In [10] it is found that fundamental frequency slope is a better predictor of stress than the raw fundamental frequency. Spectral features are measures for the effort with which phonemes are pronounced. The speaking effort is a continuous measure: it probably increases over the beginning of a stressed syllable and decreases over the end. We therefore expect that derivatives for spectral features also may be correlated with lexical stress, especially for consonants.

4 Experimental set-up

We bootstrapped a speech recogniser, made with HTK [20], from Wiggers' system [21] and did measurements on the Delft DUTAVSC corpus [19]. We used the stress marks from the CELEX lexicon. All phonemes in stressed syllables were marked as stressed, except for function words, which were marked as unstressed. All features were normalised over the whole of one utterance. The intensity measure was included in the feature vectors by HTK. For the energy in spectral bands we used the Linux program *sox* and Praat. [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.

The fundamental frequency was extracted with Praat [22]. Where Praat did not find the fundamental frequency, it was linearly interpolated. This has a number of advantages over using an out-of-range value:

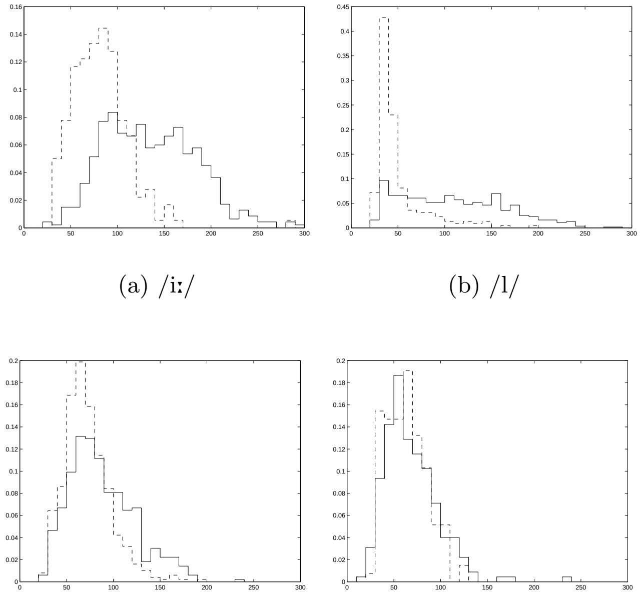
- It formalises the notion of the intonational tune in the linguistic literature [23, 24, 2], where it is pictured as a non-interrupted curve.
- From the linguistic literature, a pitch peak on or near the stressed syllable is expected. Through interpolation, even voiceless phonemes will include pitch information, so that a pitch peak at the onset or the coda of the syllable will be noticed.
- If Praat does not find voicing where there is, linear interpolation provides a reasonable workaround. This increases the algorithm's robustness.
- An out-of-range value, rather than giving the recogniser information about stress, would inject inappropriate information about apparent voicedness.

5 Results

The results were assembled by collecting feature vectors from phones that were segmented with forced alignment. The most discriminating features are candidates for inclusion in a speech recogniser that aims at recognising lexical stress. We did not find any significant results from the fundamental frequency data; duration and spectral features, however, do show much separation.

5.1 Duration

Similarly to [6, 7], we found that duration is in general a good indicator of stressedness. Stressed vowels are quite consistently longer than their unstressed counterparts (see Fig. 1(a)). Not all consonants are, as shown in Fig. 1(d): for stops the duration does not seem to differ at all, probably because stops' complete closure makes it difficult to produce lengthened ones sensibly: only the silence would be longer. Liquids consistently show a large difference, as exemplified by /l/ in Fig. 1(b), while fricatives are in between (Fig. 1(c)).



(a) /iː/ (b) /l/

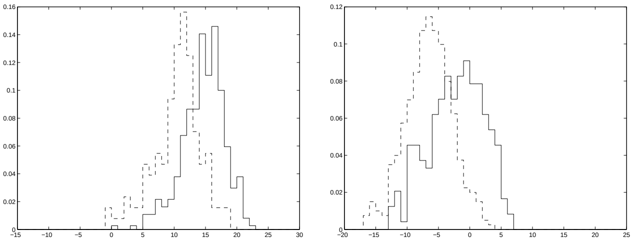
(c) /v/ (d) /b/

Fig. 1. Distributions of durations of stressed (stroked lines) and unstressed (dashed lines) phonemes in ms.

5.2 Spectral tilt

We find that different spectral tilt features apply for different phonemes. For many phonemes stressedness correlates well with some spectral tilt measure. This may be why [6] found clear correlations on a limited set of phonemes, while [7] had troubles finding correlates with a limited set of features (two). Our results show much more difference than the latter found; this may also be due to the telephone speech they used being spectrally impoverished. Figure 2 shows how spectral features correlate with long vowels (as in Fig. 2(a)) and with short vowels (as in Fig. 2(b)).

Most interestingly, the features that work for vowels give similar results for consonants. Figures 3(a) and 3(b) shows how stressed and unstressed consonants



(a) /aː/: 0.5 – 1 kHz

(b) /ɪ/: The difference between 1 – 2 kHz and 2 – 4 kHz

Fig. 2. Distributions of spectral tilt features for stressed (stroked lines) and unstressed (dashed lines) vowels in ms.

differ in terms of spectrum. On the other hand, /n/ (Fig. 3(c)) does not show spectral disparity at all. We suspect two factors play a role here:

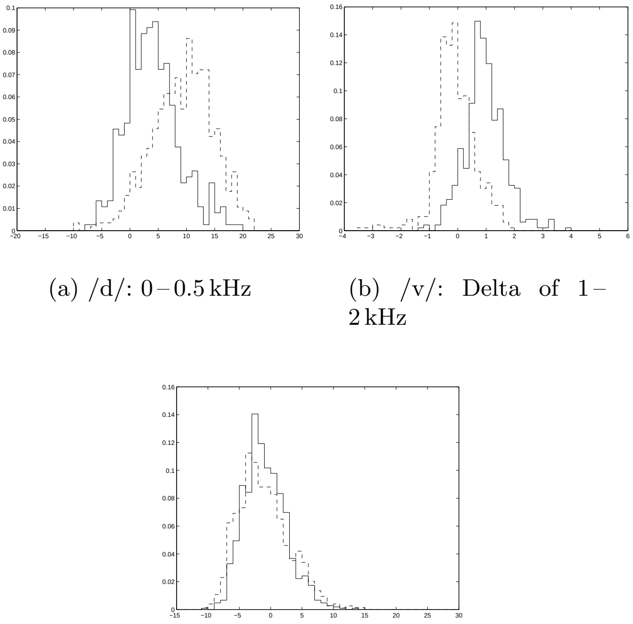
- The effect of speaking effort for fricatives and stops on the spectrum may be greater due to their friction-based realisation.
- From a perception perspective, stressed and unstressed /n/ already differ greatly in duration (similarly to Fig. 1(b)) so the difference in spectrum is not as vital for distinguishing the two.

6 Conclusion

This paper has described the importance and the feasibility of detecting lexical stress in speech. That stress works on the syllable level can be modelled effectively by adding stress marks to the phonemes in the lexical entries of a speech recogniser.

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, intensity) are correlates of stress for consonants. Various spectral tilt features apply to various phonemes. Using the duration of a phoneme while it is being recognised is not well possible with the Viterbi algorithm and standard HMMs. Another algorithm should be used if duration modelling is considered important.

Given the fact that many consonants will participate in the decision whether a syllable is stressed, we hope that implementing lexical stress recognition, even without extensive duration modelling, will improve general recognition performance on three accounts: general phone recognition, word segmentation and word recognition.



(a) /d/: 0 – 0.5 kHz

(b) /v/: Delta of 1 – 2 kHz

(c) /n/: 0.5 – 1 kHz

Fig. 3. Distributions of spectral tilt features for stressed (stroked lines) and unstressed (dashed lines) vowels in ms.

References

1. Ewen, C.J., van der Hulst, H.: The Phonological Structure of Words. Cambridge University Press (2001)
2. Ladd, D.R.: Intonational phonology. Number 79 in Cambridge Studies in Linguistics. Cambridge University Press, Cambridge (1996)
3. Harley, T.: The Psychology of Language: From Data to Theory. Psychology Press, Hove (2001)
4. Thiessen, E.D., Saffran, J.R.: When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. Developmental Psychology **39** (2003) 706–716
5. Cooper, N., Cutler, A., Wales, R.: Constraints of lexical stress on lexical acces in English: Evidence from native and non-native listeners. Language and Speech **45** (2002) 207–228
6. Sluijter, A.: Phonetic Correlates of Stress and Accent. PhD thesis, Leiden University (1995)
7. van Kuijk, D., Boves, L.: Acoustic characteristics of lexical stress in continuous telephone speech. Speech Communication **27** (1999) 95–111
8. Bouwman, A.G.G., Boves, L.: Using information on lexical stress for utterance verification. Proceedings of ITRW on Prosody in ASRU, Red Bank (2001) 29–34

9. van Kuijk, D., van den Heuvel, H., Boves, L.: Using lexical stress in continuous speech recognition for Dutch. Proceedings ICSLP **iv** (1996) 1736–1739
10. Wang, C., Seneff, S.: Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain (2001)
11. van den Heuvel, H., van Kuijk, D., Boves, L.: Modelling lexical stress in continuous speech recognition. Speech Communication **40** (2003) 335–350
12. Greenberg, S., Carvey, H., Hitchcock, L., Chang, S.: Temporal properties of spontaneous speech — a syllable-centric perspective. Journal of Phonetics **31** (2003) 465–485
13. van Son, R.J.J.H., Pols, L.C.W.: An acoustic profile of consonant reduction. Proceedings ICSLP **3** (1996) 1529–1532
14. Xie, H., Andreae, P., Zhang, M., Warren, P.: 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, Australian Computer Society, Inc. (2004) 145–150
15. Wang, X.: Duration modelling in HMM-based speech recognition. PhD thesis, University of Amsterdam (1997)
16. Russell, M.J., Moore, R.K.: Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. Proceedings of ICASSP **10** (1985) 5–8
17. Ramesh, P., Wilpon, J.G.: Modeling state durations in hidden Markov models for automatic speech recognition. Proceedings of ICASSP **1** (1992) 381–384
18. Sitaram, R.N.V., Sreenivas, T.: Incorporating phonetic properties and hidden Markov models for speech recognition. Journal of the Acoustical Society of America **102** (1997) 1149–1158
19. Wojdeł, J.C.: Automatic Lipreading in the Dutch Language. PhD thesis, Delft University of Technology, Delft (2003)
20. Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book (for HTK version 3.2.1) (2002)
21. Wiggers, P., Wojdel, J.C., Rothkrantz, L.J.: Development of a speech recognizer for the Dutch language. Proceedings of 7th annual scientific conference on web technology, new media, communications and telematics theory, methods, tools and applications (EUROMEDIA) (2002) 133–138
22. Boersma, P.: PRAAT, a system for doing phonetics by computer. Glot International **5** (2001) 341–345
23. Bolinger, D.: Intonation and its Parts. Edward Arnold, London (1986)
24. Bolinger, D.: Intonation and its Uses. PhD thesis, Stanford University (1989)