9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

The acoustic basis of lexical stress perception

Anders Eriksson[1], *Antti Suni*[2], *Martti Vainio*[2], *Juraj Šimko*[2]

[1]Stockholm University, Sweden, [2]University of Helsinki, Finland
anders.eriksson@ling.su.se,firstname.secondname@helsinki.fi

Abstract

The present study is the first in a series of studies exploring the perception of lexical stress in a number of languages. As stimuli, key words extracted from recordings in Brazilian Portuguese, English, Estonian, French, Italian and Swedish are used. The data represent male and female speakers in all languages and three different speaking styles – spontaneous speech, phrase reading, and wordlist reading. The ultimate goal of the perception studies is to explore the perception of prominence as a function of the acoustic properties of the stimuli and the native language of the listeners. In this paper we compare the prominence scores assigned to syllables by 44 native Swedish speakers with two automatic methods: acoustic feature analysis using acoustic properties of syllables and continuous wavelet transform. Both methods use duration, F_0 and spectral emphasis characteristics of speech signal or a subset thereof.

Our results demonstrate a strong language dependency of the way acoustic characteristics correlate with prominence. Correlations between prominence scores and phonological word stress patterns show that the human raters resolve this language-dependency better than the automatic signal-based methods. Also, the signal feature combinations for which the raters' judgements correlate best with the automatically assigned prominence scores depend on stimulus language to a larger extent that on the signal-based method used.

Index Terms: prominence perception, language-dependence, spectral emphasis, continuous wavelet transform

1. Introduction

The present study is the first in a series of studies exploring the perception of lexical stress in a number of languages. The planned perception studies are the second half of a larger project. In the first part of this project, the goal was to study the acoustics of lexical stress production in a number of typologically different languages. To this end, databases suitable for the purpose were recorded in six languages: Brazilian Portuguese (BPO), English (ENG), Estonian (EST), French (FRE), Italian (ITA) and Swedish (SWE). The data represent male and female speakers in three different speaking styles: spontaneous speech, phrase reading, and wordlist reading. The number of recordings per language varies between 14 (French) and 32 (Italian). Several studies of the acoustics of lexical stress in these languages have been published (e.g. [1, 2, 3, 4]). Identical methods have been used for the study of German [5] and Czech [6].

The goal of the perception studies is to explore the perception of lexical stress as a function of the acoustic properties identified as important correlates to stress level in the production studies

The stimuli used in the perception test are keywords taken from the above-mentioned recordings. The keywords occur in the three different speaking styles spoken by an equal number of male and female speakers. The number of stimulus keywords in the perception test is 72, representing language (6) sex (2) and speaking style (3), by (2) words in each category. The keywords were selected in cooperation with linguists who are native speakers of the languages in question to ensure linguistic representativity.

The test was presented via a web-based interface where raters were asked to judge the prominence of each syllable in the keywords presented one by one in random order with respect to language, sex and speaking style. The technique used was a visual analogue scale in the form of graphical panel of sliders, one slider per syllable that could be adjusted. The keyword to be judged was presented, syllable by syllable, under the sliders. The raters were instructed to adjust the sliders so that the height of the slider corresponded to the perceived relative prominence of each syllable. Slider position was automatically recalculated as a number between 1 and 100 and stored in a database. An examination of the results showed that raters varied somewhat with respect to the degree to which they used the range of possible slider positions (1-100). In the analyses we have therefore tried to minimise the effect of this variation by converting the raw scores to z-values.

In the acoustic analyses of the stimuli, we have used Spectral Emphasis, Duration, F_0-level, F_0-variation and combinations thereof as input in the analyses. In addition to these acoustic features, the prosodic features were analyzed hierarchically utilizing the continuous wavelet transform (CWT). Previously, CWT has been successfully applied for word prominence detection in Finnish [7] as well as English [8]. The current study should provide some evidence of the applicability of the method to syllable level prominence, as well as the degree of language dependency of the method. A similar multi-scale approach utilizing rhythmogram representation [9] has been found suitable for the syllable prominence task [10].

2. Methodology

The perception test was administered via a web page. Before the test, raters had to create an account, using a mail address and fill in a questionnaire asking for their age, sex, regional background, education and proficiency level on a six-point scale (0–5) of the stimulus languages. Next they were directed to a page where the test procedure was carefully explained in the native language of the listener. After that, the test itself followed. The default procedure meant that the program automatically presented the words one by one in random order with respect to language, speaking style, and sex of the speaker. But choosing the words from a word list was also an option. After judging an item and moving on to the next, the judgement by the rater was added to the database and the word was marked as submitted in the wordlist. The raters could listen to a given word as many times as they liked before submitting their answers. If, for some reason, they could not finish the whole test in one session they could log out and come back later to finish

10.21437/SpeechProsody.2018-14

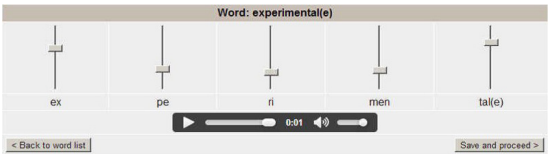

Figure 1: *The response tool used in the experiment.*

the test. All previous results had been stored and by consulting the wordlist, raters could remind themselves what words they had judged and saved.

2.1. Raters

Raters for the test were recruited among students at Stockholm University and friends they in turn recruited. Altogether 48 people performed the test. Results for four of them had to be discarded due to incomplete answers. The remaining 44 raters were 15 male (mean age 24.6 yrs, SD 4.4 yrs) and 29 female (mean age 23.9 yrs, SD 6.9 yrs). The average self-reported proficiency scores of the target languages for the raters were 5, 3.25, 0.02, 0.2, 0.11, 0.7 for SWE, ENG, EST, ITA, BPO and FRE, respectively.

2.2. Response tool

Figure 1 shows the response tool used in the experiment. For each new word the response tool appeared with the sliders positioned in the middle of the range. Raters could listen to the word to be judged as many times as they liked by clicking on the arrow at the bottom. When they felt satisfied that the positions of the sliders corresponded to the perceived relative prominence of the syllables they were instructed to press "Save and proceed". Their responses were then saved in the database and the next word was presented.

The range of the slider corresponded to a scale between 1 and 100. Subsequently, the responses were z-normalized for each individual rater. Finally, these normalized judgements were averaged across all raters.

2.3. Parameters used in the acoustic feature analysis

The acoustic analysis of the stimuli used in this experiment is identical to that used in the production studies. The sound files were transcribed at the segment level using Praat TextGrids. The transcribed files were then used by a script that computed the values of the parameters described below segment by segment. In the present analysis, however, only the acoustic properties of the syllable nuclei have been considered. This is also in accordance with the production studies [1, 2, 3, 4].
Fundamental frequency level is here defined as the F_0 median in the vowel in order to minimize the influence of outliers. The median is measured in semitones relative to 1 Hz.
Duration is measured in ms.

In these analyses we used a simplified version of the Spectral Emphasis.
Spectral Emphasis (dB) = $\text{SPL}_{\text{full}} - \text{SPL}_0$, where
SPL_{full} is the SPL of the full spectrum in a given segment and SPL_0 is the SPL of the low-pass filtered segment using a cutoff frequency of $1.5*F_0$mean at 18 dB/octave (see [11]).

The use of the semitone scale for frequency means that we may expect the variation to be approximately the same for male and female speakers. The semitone scale also reduces skew.

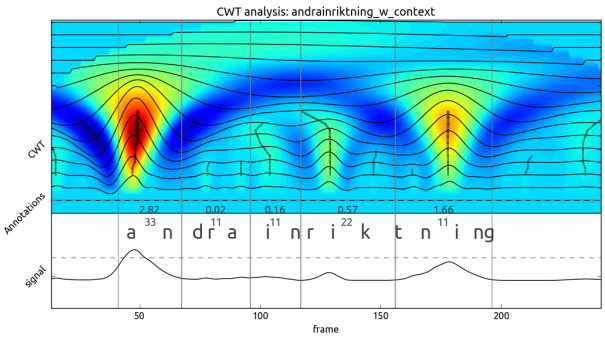

Figure 2: *The wavelet-based estimation of syllable prominence based on the lines of maximum amplitude.*

Using a log scale tends to make the distribution more normal. For this reason, we express duration as $\log_2$(ms). Log-scales are thus used for all parameters.

2.4. Wavelet analysis

Continuous wavelet analysis of the stimuli was performed on the combination of fundamental frequency (F_0), spectral emphasis and duration signals, utilizing a technique developed for word prominence detection described in [8]. In short, F_0 and emphasis signals were extracted and sampled at 200 Hz. Voiceless gaps of the F_0 signals were filled by cubic interpolation. Effect of the gaps was further alleviated by smoothing the emphasis signals. Two duration signals were constructed separately for segments and syllables, respectively, in the following way (similarly as in [8]): the value of each segment/syllable duration was placed in the mid-time point of the unit and zero values were placed at the unit boundaries. Subsequently these points were connected using cubic interpolation to form a smooth duration signal. The two resulting signals, one for the syllables and one for the segments were summed and this sum was used as a duration signal in this work.

The individual signals were then normalized between zero and one and combined by multiplying them (instead of summing as in [8]). Resulting combined signals were subjected to the continuous wavelet transform using a Mexican Hat mother wavelet, with scales a quarter of an octave apart. Lines of maximum amplitude were determined for each syllable from ten scales centered on average syllable length of the stimuli, yielding final prominence estimates (see Figure 2).

3. Results

3.1. Interrater reliability

A requirement if we want to claim that the scores produced by the listeners give us a representative picture of how listeners in general judge prominence (in this case) is that there is reasonably agreement between the judges. There are several methods that may be used to quantify agreement. The one we have chosen here is Cronbach's Alpha. This index ranges from 0 to 1 where a higher index means closer agreement. An often used rule of thumb says that a value above 0.7 for Cronbach's Alpha can be considered "acceptable". As may be seen in Table 1, the reliability scores are in all cases well above the recommended minimum value. A score greater than 0.9 is considered "excellent". Interrater Reliability is alternatively referred to as Internal Consistency. The figures in Table 1 are based on the

z-normalised scores. Using the raw scores produces the same picture with only marginally lower reliability.

Table 1: *Values of Cronbach's Alpha for the languages used in the test.*

BPO	ENG	EST	FRE	ITA	SWE	All
0.903	0.970	0.937	0.848	0.932	0.936	0.935

3.2. Perceptual evaluation, signal-based prominence estimates and phonological stress

In the absence of native judgements for our speech material – except for Swedish words – we use phonological stress as a proxy for prominence patterns. As not all languages investigated here manifest secondary phonological stress, we evaluate the raters' ability to distinguish between unstressed and primary stressed syllables; we expect a consistently higher score for primary stressed syllables compared to the unstressed ones. Although phonological lexical stress in French is not clearly defined, we refer to the last syllable of each French word as stressed one as it is often claimed to be the most "prominent."

We compared the mean rater scores for unstressed with those for the stressed syllables using a t-test. As the second column in Table 2 shows, the raters judged the stressed syllables as, on average, significantly more prominent that the unstressed ones.

A similar comparison for the signal based prominence estimation is summarized in columns 3–9 in Table 2 for both methods and all signal property combinations. For every language there exists at least one signal property combination that assigns significantly higher scores to the stressed than to the unstressed syllables. The appropriate signal properties which capture this distinction, however, vary between languages. For example, while for the English material all tested methods consistently assign a significantly higher average score to the stressed syllables, for Italian this is the case only for the duration (using either method) and duration–emphasis combination for the acoustic feature analysis. Note also that for Brazilian Portuguese some signal-based estimates, in particular those using F_0, did assign significantly *lower* scores to the stressed compared to the unstressed syllables.

3.3. Comparing perceptual evaluation with signal-based prominence estimates

Table 2 suggests that the raters used different cues to detect (phonologically presumed) prominent syllables for different languages, in particular for the Romance ones. Therefore, we directly tested the correspondence between the responses of the raters on the one hand, and the results of the acoustic feature and wavelet analyses on the other. The three types of acoustic properties of the test words – duration, F_0 and emphasis – were used, in all possible combinations. For the acoustic feature analysis, the chosen properties were summed and used as prominence estimates. For the wavelet analysis, the corresponding signals were multiplied prior to CWT-transformation, and the prominence estimates were calculated using the lines of maximum amplitude. These prominence estimates based of the acoustic feature analysis and wavelet technique were compared to the raters' evaluations.

Table 3 summarizes correlations between the signal-based estimates for all possible combinations of signal properties, and the raters' judgements of prominence. Generally, the signal-based estimates correlate positively with the perceptual evaluations by the Swedish listeners. The correlations with signal-based estimates using the best (language-dependent) features range between 0.48 (for Italian) and 0.91 (for English).

As suggested in Table 2, the best signal-based estimates were achieved using different combinations of signal properties for different languages. Note also that the F_0-based estimates for Brazilian Portuguese show negative correlation with the raters' judgments. This is due to the fact that in Brazilian Portuguese there is a peak before the stressed syllable to ensure a low target for the stressed syllable [12]. Although this strategy is not commonly used in Swedish, the Swedish raters identified the stressed syllables in Brazilian Portuguese significantly better than chance. This may be explained by the fact that the stressed vowels in Brazilian Portuguese were characterised by markedly greater duration and also, in most cases by the highest degree of spectral emphasis, properties that also play a significant role in marking stress level in Swedish.

In general, the two signal-based estimation techniques do not dramatically differ in their agreement with the perceptual estimation by the Swedish participants, although the wavelet-based approach performs slightly better in this task. The average correlation across all combination is 0.43 for the acoustic feature method and 0.47 for wavelet-based estimation. For each language individually, the best signal-based estimate is one using the wavelet-based technique.

4. Discussion

The interrater reliability shows that the Swedish raters judged syllable-level prominence level consistently, despite the difficult nature of the task. As far as we can tell from lexical stress judgements, the ratings were also in substantial agreement with phonologically based stress models as they rated primary-stressed syllables as significantly more prominent than the unstressed ones for all languages.

This finding is somewhat surprising as the evaluation of word stress using signal-based techniques shows that stress is signalled differently, i.e., using different signal properties in different languages. For English, individual properties as well as their combinations reliably differentiate stressed and unstressed syllables. For Swedish with its tonal properties, automatic methods using only F_0 fail to assign significantly higher prominence to stressed than to unstressed syllables. Similarly, for Estonian with its rich phonological quantity system, duration correlates rather weakly with stress distinction[1].

For the Romance languages, the correlation between the lexical stress and signal properties (as captured by our methods) is rather weak. For Brazilian Portuguese and Italian, the stress correlates only with duration, and the duration–emphasis combination. For French, the last syllables are longer and have higher F_0 than other syllables. This strong language dependency of stress signalling shows that automatic syllable-level prominence detection should use different features for different languages.

Even though these findings suggest that raters used language-dependent strategies to assign higher prominence to stressed syllables, in general their prominence rating correlated quite well with those made by signal-based techniques. A closer look at Table 3 however shows that the agreement between the raters and the automatic methods is highest for

[1]This is despite the fact that in our Estonian material all syllables were phonologically short.

Table 2: *Significance levels (p-values) of t-test comparing prosody estimates for stressed and unstressed syllables made by the raters and the signal-based techniques: 0.05 > * > 0.01 > ** > 0.001 >***. The significance levels in brackets mark the results where the prominence estimate for the stressed syllables was significantly lower than for the unstressed ones; for the remaining significant differences the estimates were higher for the stressed syllables.*

		dur		F_0		emph		dur&F_0		dur&emph		F_0&emph		All	
	raters	*AFA*	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*
SWE	***	**	**			*	***	***	*	***	***	***	***	***	***
ENG	***	***	***	*	*	***	***	***	***	***	***	***	***	***	***
EST	***		*	***	*		***	***	***		***	***	***	***	***
BPO	**	***	***	(**)	(*)						**	(**)	(***)		
ITA	***	*	*							*					
FRE	**	*	*	**	*			**	**						*

Table 3: *Correlations between raters' prominence estimates and the estimates produced by the two signal-based techniques. The highest correlation for each language are in bold.*

	dur		F_0		emph		dur&F_0		dur&emph		F_0&emph		All	
	AFA	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*	*AFA*	*CWT*
SWE	0.46	0.45	0.26	0.26	0.42	0.59	0.46	0.32	0.52	0.58	0.56	**0.66**	0.62	0.58
ENG	0.65	0.57	0.47	0.41	0.68	0.88	0.79	0.83	0.73	0.91	0.77	0.84	0.84	**0.91**
EST	0.36	0.56	0.65	0.40	0.31	0.51	0.69	**0.81**	0.44	0.67	0.71	0.65	0.75	0.73
BPO	0.44	0.53	-0.41	-0.33	0.29	0.23	-0.02	0.08	0.53	**0.57**	-0.13	-0.13	0.20	0.14
ITA	0.47	**0.48**	0.15	0.25	0.14	0.23	0.41	0.40	0.40	0.43	0.18	0.26	0.37	0.32
FRE	0.40	0.45	0.53	0.56	0.05	0.26	0.60	**0.70**	0.29	0.37	0.45	0.42	0.57	0.59

those signal properties and combinations for which the automatic methods best captured the "objective" contrast between stressed and unstressed syllables. For English, the best correlation is achieved with the combination of all features, for Swedish with the F_0–emphasis combination, for Estonian and French with the duration–F_0 combination, for Italian with the syllable duration alone. For Brazilian Portuguese, the best agreement is achieved with the duration–emphasis combination while the judgements correlated *negatively* with the F_0-based features of the signal, suggesting that the Swedish raters "understood" the F_0-lowering strategy mentioned above.

Interestingly, the correlation between the Swedish raters and the automatic signal-based methods for prominence detection is rather low for the Swedish speech material. The best correlation of 0.66 is achieved for F_0–emphasis properties; this is lower than the "best" correlations for English, Estonian and French. While these best correlations do not themselves correlate with the average proficiency scores reported by the raters, it is possible, that they did not base the prominence judgment in their native tongue purely on signal properties, but are biased by their familiarity with the language in terms of "knowing" which syllables *should* be prominent.

Alternatively, it is possible that our signal-based methods do not fully capture strategies used by human raters. For example, acoustic feature analysis uses only local acoustic features (pertaining to a given syllable) to assign a prominence level to the syllable. The wavelet-based method does capture a wider context to some degree, but also calculates the prominence estimate based on a line of maximum amplitude delimited by syllable boundaries. It is possible that the human rating reflected more global characteristics of words.

We hope that including ratings from subjects with other language backgrounds, as planned, will clarify some of the issues mentioned above.

Our results do not show any big difference between the two signal-based methods in terms of their agreement with human raters. Although the wavelet-based technique performed slightly better than the acoustic feature analysis, the difference is small, and not statistically significant for any of the conditions shown in Table 3.

A potential drawback when using wavelet-based prominence estimation for the given material is in applying this technique for target words recorded separately in the word-list condition. These stimuli are not preceded nor followed by speech material; the wavelet analysis pads the signal by zero F_0 and emphasis values. This padding might distort the prominence estimates. To check this influence, we also calculated correlations between human prominence ratings and signal-based methods for a subset of our material excluding the words recorded within wordlists. For the remaining words cut out from spontaneous speech and phrase reading, the wavelet analysis included an immediate context from the original recordings. Including this context lead to a slight, but not dramatic, increase in correlation between the wavelet-based method and human ratings, with the average correlation increasing from 0.47 to 0.51.

5. Acknowledgements

This work was partly funded by the Academy of Finland DLT project (No. 12933481) and by the Swedish Research Council (VR) under grant 2007-2301.

6. References

[1] A. Eriksson, P. A. Barbosa, and J. Åkesson, "The acoustics of word stress in Swedish: A function of stress level, speaking style and word accent," in *Proc. INTERSPEECH 2013*, Lyon, France, 2013, pp. 778–782.

[2] A. Eriksson and M. Heldner, "The acoustics of word stress in English as a function of stress level and speaking style," in *Proc. INTERSPEECH 2015*, Dresden, Germany, 2015, pp. 41–45.

[3] A. Eriksson, P. M. Bertinetto, M. Heldner, R. Nodari, and G. Lenoci, "The acoustics of lexical stress in Italian as a function of stress level and speaking style," in *Proc. INTERSPEECH 2016*, San Francisco, CA, 2016, pp. 1059–1063.

[4] P. A. Barbosa, A. Eriksson, and J. Åkesson, "On the robustness of some acoustic parameters for signalling word stress across styles in Brazilian Portuguese," in *Proc. INTERSPEECH 2013*, Lyon, France, 2013, pp. 282–286.

[5] J. Behrens, "Die Prosodie des Wortakzentes in Abhängigkeit von Akzentlevel und Sprechstil," BA Thesis, Christian-Albrechts-Universität zu Kiel, 2013.

[6] R. Skarnitzl and A. Eriksson, "The acoustics of word stress in Czech as a function of speaking style," in *Proc. INTERSPEECH 2017*, 2017, pp. 3221–3225.

[7] M. Vainio, A. Suni, and D. Aalto, "Continuous wavelet transform for analysis of speech prosody," *TRASP 2013-Tools and Resources for the Analysys of Speech Prosody, An Interspeech 2013 satellite event, August 30, 2013, Laboratoire Parole et Language, Aix-en-Provence, France, Proceedings*, 2013.

[8] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[9] N. P. M. Todd, "The auditory primal sketch: A multiscale model of rhythmic grouping," *Journal of New Music Research*, vol. 23, no. 1, pp. 25–70, 1994.

[10] B. Ludusan, A. Origlia, and F. Cutugno, "On the use of the rhythmogram for automatic syllabic prominence detection," in *Proc. INTERSPEECH 2011, Florence, Italy, August 27-31, 2011*, 2011, pp. 2413–2416.

[11] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women and children," *Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[12] P. A. Barbosa, Personal communication, 2017.