10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Lexical stress perception as a function of acoustic properties and the native language of the listener

Anders Eriksson[1], *Rosalba Nodari*[2], *Juraj Šimko*[3], *Antti Suni*[3], *Martti Vainio*[3]

[1]Stockholm University, Sweden, [2]Scuola Normale Superiore di Pisa, Italy, [3]University of Helsinki, Finland

anders.eriksson@ling.su.se,rosalba.nodari@gmail.com,firstname.secondname@helsinki.fi

Abstract

The study is part of a series investigating production and perception of lexical stress in a number of languages including Brazilian Portuguese, English, Estonian, French, Italian and Swedish. The production database contains data representing male and female speakers in the above languages in three speaking styles – spontaneous speech, phrase reading, and wordlist reading. Keywords from these recordings, representing male and female speakers and all speaking styles are used. The participants' task is to judge the relative syllable prominences of the keywords presented one by one. In a previous study, subjects were native Swedish speakers. In the present study subjects are native speakers of Italian.

In the analyses, perception results are correlated with acoustic variables shown to be important in the production studies. From the previous perception study we know that acoustic syllable prominence affects perceived syllable prominence. But there is also a possibility that listeners' perception may be biased by expectations based on the listeners' native language. The main result is that there are great similarities between the Swedish and Italian listeners in the way acoustic prominence affects perceived prominence, but we are also able to demonstrate a case of native language bias.

Index Terms: Models of speech perception, Acoustic and articulatory cues in speech perception, Perception of prosody

1. Introduction

The present study is part of a larger study investigating the production and perception of lexical stress. During the first phase, the production studies, we have examined the acoustics of lexical stress production in a number of typologically different languages – Brazilian Portuguese (BPO), English (ENG), Estonian (EST), French (FRE), Italian (ITA) and Swedish (SWE). The production database contains recordings by male and female speakers using three different speaking styles: spontaneous speech, phrase reading, and wordlist reading. The number of recordings per language varies between 14 (French) and 32 (Italian). Results for these languages have been described in a number of published studies (e.g. [1, 2, 3, 4]). Identical methods have been used for the study of German [5] and Czech [6]. From these studies, we have learnt how a number of fundamental prosodic parameters – Duration, f_0-level, f_0-variation, and Spectral Emphasis – are used to signal prosodic prominence.

In the next phase of this research effort, the goal is to examine to what extent these parameters may explain the perception of prosodic prominence. It seems like a reasonable assumption that acoustic properties which have been shown to be strongly correlated with the production of prominence variation should also be significantly involved in the perception of prominence variation. This has indeed also been confirmed in a previous study [7]. In that study we also compared the use of the parameters mentioned above and combinations thereof with analyses utilizing the continuous wavelet transform (CWT). Previously, CWT has been successfully applied for word prominence detection in Finnish [8] as well as English [9]. This will be described in more detail in the Methods section.

The stimuli in the perception studies are keywords taken from the above-mentioned production 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.

In the previous study [7] mentioned above, the subjects were native speakers of Swedish. It makes some sense, however, to suggest that prominence perception may be influenced not only by acoustic prominence but also by linguistic expectations based on familiarity with the languages used in the stimuli. A study [10] using Swedish stimuli presented to Swedish native speakers and English listeners with no knowledge of Swedish suggested that the language background of the listener may introduce a rather substantial bias. To approach the question of listener language bias in a broader context, we presented the test we had used for Swedish listeners [7] to Italian listeners as a possible way of identifying differences that may be attributed to listener language bias by comparing the answers from the two listener groups. The test was presented via a web-based interface where raters were asked to judge the prominence of each syllable in the keywords presented in random order with respect to language, sex and speaking style.

2. Methodology

The technique used was of a visual analogue type in the form of a graphical panel of sliders, one slider per syllable that could be adjusted to match the perceived syllable prominences. The test was designed in such a way that it was possible to leave an incomplete test and come back later to finish it. For that purpose, the first step was to create an account using a mail address as the identifier. In the next step, the raters had to fill in a questionnaire asking for their age, sex, regional background, education and self-estimated proficiency level on a six-point scale (0–5) of the stimulus languages. After that, the test itself could begin. The default presentation option was presenting the keywords in random order, but choosing the words from a word list was also an option. This option was created to make it possible to return to a given word and reconsider a rating should the rater wish to do so. The raters could listen to a given word as many times

10.21437/SpeechProsody.2020-92

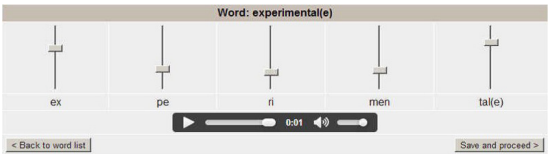

Figure 1: *The response tool used in the experiment.*

as they liked before submitting the answer. By consulting the wordlist, raters could remind themselves what words they had judged and saved.

2.1. Raters

Collection of perception data was administered from the department of linguistics at *Scuola Normale* in Pisa. Raters were recruited among students and staff but we also cast a much wider net involving colleagues and acquaintances outside the department. Not all who registered completed the test. Some did in fact not submit a single answer. We therefore decided to only consider the 45 participants who had submitted answers to all 72 items. From those 45, a further 4 participants were discarded for skipping over a suspiciously large number (19 to 34) of responses. The remaining 41 raters were 15 male (mean age 32 yrs, SD 8.9 yrs) and 26 female (mean age 40 rs, SD 11.8 yrs). The average self-reported proficiency scores of the target languages for the raters were 0, 2.6, 0, 5, 0.2 and 1.1 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. When the raters felt satisfied that the positions of the sliders corresponded to the perceived relative syllable prominence they were instructed to press "Save and proceed". Their responses were then saved in a database and the next word was presented.

Slider positions were stored in the database as values between 1 and 100. For the final analyses, these values were z-normalized for each individual rater and the normalized judgements averaged across all raters.

2.3. Parameters used in the acoustic feature analysis (AFA)

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 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:

$$\textit{Spectral Emphasis}\ (\text{dB}) = \text{SPL}_{\text{full}} - \text{SPL}_0,$$

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 $1.5\times\text{mean}(f_0)$ at 18 dB/octave (see [11]).

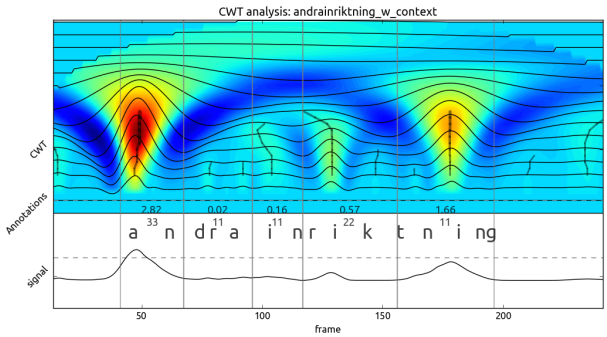

Figure 2: *The wavelet-based estimation of syllable prominence based on the lines of maximum amplitude.*

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 skewness. 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 (CWT)

In order to investigate if the perception results are affected by factors not captured by the aggregate values of the raw acoustic data, we also analyzed the acoustic parameter trajectories hierarchically with continuous wavelet analysis (CWT). This type of analysis has proven useful as a way of modelling prosody in many cases. Specifically, in detection of word prominence in English, CWT has provided marked improvements compared to raw signals [9]. We therefore wanted to see if these results could be generalized to capturing perceived prominence also at the syllable level in target languages with widely varying phonological structures. CWT was performed on combinations of fundamental frequency (f_0), spectral emphasis and duration signals, utilizing a technique developed for word prominence detection described in [9], adapted to syllable level prominence as in [7]. Syllable prominence was quantified as lines of maximum amplitude across wavelet scales, depicted in Fig. 2.

3. Results

3.1. Interrater reliability

Table 1: *Values of Cronbach's Alpha for both Italian and Swedish raters and the languages used in the test.*

	BPO	ENG	EST	FRE	ITA	SWE	All
ITA	0.92	0.96	0.92	0.90	0.93	0.89	0.92
SWE	0.90	0.97	0.94	0.85	0.93	0.94	0.94
Inter	0.90	0.97	0.91	0.82	0.90	0.88	0.91

A perception experiment is in principle a way of measuring something using human judgement as the instrument. Like other types of instruments it has to be calibrated for reliability. There are several statistical methods for doing that, and the method we have chosen for the present experiment is Cronbach's Alpha. The method returns values from 0 to 1, where 0 means no reliability and 1 perfect reliability. An often used rule of thumb says that a value above 0.7 can be considered an "acceptable" degree of reliability. As may be seen in Table 1, the reliability scores are in all cases well above the recommended minimum. A score greater than 0.9 is considered "excellent".

Table 2: *Pearson correlations between prominence judgements by Italian and Swedish raters and the corresponding estimates produced by the two signal based techniques, based on data from 11396 judgements by Swedish raters and 10612 by Italian raters.*

		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	ITA	0.53	0.58	0.11	0.18	0.60	0.71	0.38	0.30	0.69	0.73	0.59	0.60	0.68	0.57
	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	ITA	0.70	0.65	0.34	0.44	0.72	0.82	0.74	0.78	0.78	0.87	0.70	0.81	0.82	0.85
	SWE	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	ITA	0.32	0.51	0.41	0.33	0.38	0.40	0.48	0.67	0.48	0.55	0.58	0.48	0.63	0.56
	SWE	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	ITA	0.58	0.69	-0.47	-0.35	-0.03	-0.03	0.05	-0.04	0.39	0.57	-0.46	-0.49	0.01	-0.06
	SWE	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	ITA	0.33	0.32	-0.04	0.09	0.14	0.21	0.18	0.19	0.30	0.31	0.06	0.22	0.20	0.17
	SWE	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	ITA	0.52	0.51	0.53	0.56	-0.18	0.00	0.66	0.71	0.16	0.09	0.27	0.21	0.47	0.33
	SWE	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

The second and third rows in Table 1 show Alpha values for the z-normalised perception scores for Italian and Swedish [7] raters. It is immediately apparent from the figures in Table 1 that the Interrater reliability is at the same level for both groups.

The last row in Table 1 shows the results on interrater agreement test between the two rater groups using Cronbachs Alpha applied to the z-normalised scores. If we compare the results with the internal agreement in the two groups, we may observe that the agreement levels, both per stimulus language and total are at the same levels. Also, a Univariate ANOVA, using Speaker Language and Listener Language as independent variables and z-normalised judgements as the dependent variable, shows no significant differences.

Based on such observations we may say that the two groups seem to have interpreted the task the same way and also landed in almost identical judgements.

3.2. Correlating acoustics and perception

Correlations as a function of signal-bases parameters were analysed in quite some detail in our previous perception study [7] and for details we refer to that study. A major motivation for repeating the study with listeners with a different native language was, however, to gain insights into the possibility of a perceptual bias caused by listener language. We will therefore here give priority to listener group comparisons.

Table 2 shows a summary of Pearson correlations between signal-based parameters and perceptual estimates of the Italian as well as Swedish raters. Our first observation is that the results presented in the table are almost identical between the rater groups. Correlation values differ minimally between the two groups and, more importantly, pick out the dominant correlations between signal based properties and perception in very similar ways.

The results, however, also suggest some noteworthy effects of stimulus language. Only duration based correlations are significant for Italian. Duration plays major role also for Brazilian Portuguese, but here we also find significant negative correlations when f_0 is involved. We may explain both findings by referring to our production studies ([3, 4]). In both languages duration plays a dominant role. In Brazilian Portuguese, there is usually a f_0 peak before the stressed syllable to ensure a low tone in the stressed syllable. The duration correlation, 0.33, for Italian by the Italian participants is somewhat puzzling given that duration is the main correlate in production. For the Italian listeners the correlation for phonological stress is 0.81 (see below), while the correlation for duration is only 0.33, suggesting that they have been able to keep the two apart which speaks in favour of a possible linguistic bias over acoustics. In this case the difference is particularly great, but there are similar trends in other cases. The role of phonological stress is further evaluated in the following section.

CWT does not provide a marked advantage over raw acoustic data for duration and f_0. Neither of these features are likely to exhibit word-internal hierarchical structure, so equal performance with raw values is expected. On the other hand, spectral emphasis with more temporal variation does benefit from CWT, yielding better correlations with listeners' prominence judgements than raw values.

3.3. Comparing perception with phonological stress

Table 3: *Comparisons of correlations between raters' prominence judgements and stress with the highest correlations between the judgements and signal-based prominence estimates (CWT-based unless specified). Significance levels (Bonferroni corrected):* $0.05 >^{*} > 0.01 >^{**} > 0.001 >^{***}$).

stimuli	raters	stress		best	
SWE	ITA	0.68	-	0.73	dur&emph
	SWE	0.91	$>^{*}$	0.66	f_0&emph
ENG	ITA	0.93	-	0.87	dur&emph
	SWE	0.91	-	0.91	dur&emph
EST	ITA	0.32	$<^{**}$	0.67	dur&f_0
	SWE	0.63	-	0.81	dur&f_0
BPO	ITA	0.82	-	0.69	dur
	SWE	0.56	-	0.57	dur&emph
ITA	ITA	0.81	$>^{***}$	0.33	dur (AFA)
	SWE	0.61	-	0.48	dur
FRE	ITA	0.83	-	0.71	dur&f_0
	SWE	0.47	-	0.70	dur&f_0

The third column in Table 3 lists correlations between the average rater estimates (for Swedish and Italian subjects) with phonologically defined stress for the languages operationalised as: Primary stress = 3, Secondary = 2, Unstressed = 1. To compare the (dependent) correlations for the two rater groups rating the same language, we used Hotelling-Williams test (R package psych). The test compares the judgements-stress correlations

for the Italian raters with the judgements-stress correlations for the Swedish raters, given the correlation between the judgments of the two rater groups.

The differences between rater groups are significant ($p < 0.001$) except for English ($p = 0.26$). Interestingly, the correlations between the ratings and stress are higher for Swedish subjects rating prominence for Swedish and Estonian material, and for Italian raters on Romance languages.

To evaluate to what extent the raters based their judgments on their potential knowledge of language phonology rather than signal properties, we compared the stress-rating correlations with the highest correlations between the judgments and signal-based estimates achieved for each material–rater group combination (given the correlations between the stress and signal-based estimates). Interestingly, as shown in Table 3, the correlation with stress is significantly higher than the best correlation with signal-based method only for the subject ratings for their own mother tongue. (For Estonian, the subject ratings correlate significantly better for the duration–f_0 CWT estimate than with phonological stress for Italian raters.)

4. Discussion and Conclusions

As we have seen in Section 3.2, Italian and Swedish listeners arrive at nearly identical prominence judgements. We illustrate this graphically by Fig. 3 showing a reconstruction of the slider position for one of the keywords based on listener group means. We could have chosen any one of the other keywords, they all show the same picture with minor variations.

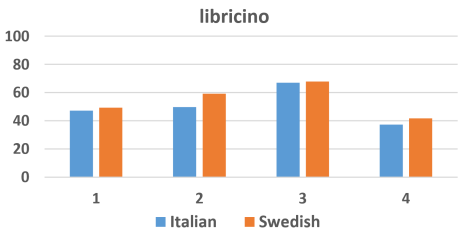

Figure 3: *A representative example of average prominence ratings for the two listener groups. Y-axis indicates prominence and X-axis syllables.*

The prosodic judgements are thus the same in both groups, at least in the sense that the observed variation is not statistically significant. At one level one might suggest that this result is trivial. The method used is typical of a psychoacoustic experiment, including neutral instructions not suggesting any particular interpretation of "prominence". In the best of worlds then, there should be no language bias. But our data suggest that there is in the sense that Swedish and Italian subjects often seem to reach their decisions in somewhat different ways. For Italian stimuli for example, the Italian listeners probably recruit their native language intuitions and give that some priority over the acoustics. The Swedish listeners, who do not know any Italian, rely more on acoustics, in this case primarily duration. A similar bias exists for Swedish listeners and Swedish stimuli.

To sum up, we have noted that the subjects in the two experiments (the previous Swedish study and the Italian study described here) behave in very similar ways. The way they score syllable prominence does not differ significantly, at last not in a global perspective. But at the same time, we have been able to identify a case of listener language bias in the scoring of native language stimuli and in the way acoustic information is used. How can we understand these results in a common framework?

The method used in the two studies is classical psychoacoustics using a technique called cross-modality matching. In a flawless psychoacoustics design, language should play no role at all, it should all be a result of the properties of the perceptual system of the listener and there is no reason to expect that to be different in Italian compared to Swedish listeners. In such a scenario, we assume that the listeners rely on some holistic interpretation of prominence that need not be further specified. In a similarly flawless acoustic model based on all and only the relevant factors involved in predicting perceived syllable prominence, we would also expect the two groups to perform the same. But if neither model is perfect, we end up somewhere in between, and that opens a window for bias and that bias should be most marked for stimuli in the listener's native language.

This is where the phonological stress model comes in. Stress models are perceptually grounded, albeit representing idealisations. We have seen that the phonological models correlate better with perception scores than the acoustic models in 4 out of 6 of the languages. The difference is not great, but it is there. If we assume that the phonological representation comes close to a psychoacoustic representation, this assumption generates some predictions that we may check against our results.

The first prediction is, of course, that Italian and Swedish listeners will rate prominence in very similar ways, which is indeed what we have seen. But this should not be without exception. It should be limited to languages where stress is contrastive, which is Swedish, English, Brazilian Portuguese and Italian. It should not apply to Estonian where stress is fixed which means that the various phonetic cues to prominence may be recruited for different purposes making stress identification via acoustics confusing. And this is precisely what we may observe. Two observations stand out against these predictions; French in both groups and Swedish as judged by Italian listeners. French does not have lexical stress at all, but final syllables are nevertheless often more prominent both acoustically and perceptually. We have used this observation by marking final syllables in French as stressed and Italian listeners seem to agree judged by the correlation with our invented stress marking. For Swedish listeners, acoustic variables are better correlated. So the results are ambiguous and, more importantly, marking stress in French has no support in phonological theory.

Swedish, finally, is a typical language with lexical stress so why are Italian listeners less accurate in hearing this? Well, even if we here base our predictions on phonological stress, acoustics is by no means irrelevant. In addition to contrastive stress, Swedish has also a two-way contrastive tone based word accent which means that two words which have identical stress patterns can be distinctive with respect to accent and this influences the acoustic properties of syllables as well as was shown in our production study [1]. Swedish listeners can, of course, easily keep the two types apart while the Italian listeners have to base their judgments on the combined acoustic effect of stress and accent; an assumption in perfect agreement with the results. We are not claiming that the above account is THE solution but we do think that the above correlations are interesting enough to merit further studies involving listeners with other native languages to see if the observations made here may be generalised to other listener languages.

5. Acknowledgements

Data collection for this work was supported by a grant from Magnus Bergvalls Stiftelse (2017-02435).

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] A. Eriksson, A. Suni, M. Vainio, and J. Šimko, "The acoustic basis of lexical stress perception," in *Proceedings of the 9th International Conference on Speech Prosody 2018*, Poznan, Poland, 2018.

[8] 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.

[9] 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.

[10] A. Eriksson, E. Grabe, and H. Traunmüller, "Perception of syllable prominence by listeners with and without competence in the tested language," in *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, 2002, pp. 275–278.

[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.