INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Controlling prominence realisation in parametric DNN-based speech synthesis

Zofia Malisz[1], *Harald Berthelsen*[2], *Jonas Beskow*[1], *Joakim Gustafson*[1]

[1]Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
[2]STTS – Södermalms talteknologiservice AB, Stockholm, Sweden

[malisz,beskow,jocke]@kth.se, harald@stts.se

Abstract

This work aims to improve text-to-speech synthesis for Wikipedia by advancing and implementing models of prosodic prominence. We propose a new system architecture with explicit prominence modeling and test the first component of the architecture. We automatically extract a phonetic feature related to prominence from the speech signal in the ARCTIC corpus. We then modify the label files and train an experimental TTS system based on the feature using Merlin, a statistical-parametric DNN-based engine. Test sentences with contrastive prominence on the word-level are synthesised and separate listening tests a) evaluating the level of prominence control in generated speech, and b) naturalness, are conducted. Our results show that the prominence feature-enhanced system successfully places prominence on the appropriate words and increases perceived naturalness relative to the baseline.

Index Terms: speech synthesis, prosodic prominence, deep neural networks

1. Introduction

Speakers highlight important stretches in speech by lengthening, pitch excursions and expansion of spectral features. The resulting prosodic prominence patterns are a function of robust linguistic constraints, such as lexical stress and phrase accent, but also of many top-down processes: the effects of discourse, rhythmic context, priming or the information density of neighbouring elements [1, 2].

In synthetic speech, incorrectly placed or missing prominence has a highly negative effect on intelligibility and naturalness and makes listening to long stretches of synthetic speech tiring [3]. The main issue with standard prediction of prosody in TTS is that models are based on the robust features derived from text rather than the speech signal used for training [3]. This leads to a poor representation of the entirety of prosodic variation that cannot be derived from text.

The information on the variation is nonetheless available in the training corpus. Signal-based prosodic features can be extracted using automatic tools that model realised prominence in speech [4, 5, 6] and subsequently added to a TTS system. The development of the speech tagging tools in turn relies on the knowledge on how human listeners - the users of TTS systems - perceive and categorise prominence [1, 7]. Recent advances show that crowdsourcing methods enable to directly access human prominence judgments in a relatively short time [8, 9, 10].

Prominence is often manipulated post-training. In formant synthesis and di-phone synthesis, explicit rules are used to manipulate f0 and duration (but not spectral features) for this purpose. In concatenative speech synthesis, one solution is to pick a specific parameter in the unit selection, e.g. duration, selecting longer segments to realise prominence.

[11] made signal-driven, perceptual-based prominence labels accessible for unit selection in BOSS. Prominence was considered as a unit cost factor. The prominence-configured system was preferred by listeners but did not improve intelligibility relative to baseline. [12] provided unsupervised prediction of prominence from speech on the foot level. They found an optimal number of prominence levels, namely four, based on clustering according to perceptual distinctiveness criteria. The evaluation showed that using prominence as an intermediate representation to compute target pitch contours in concatenative TTS was rated as more natural and expressive than in rule-based approaches. Finally, in a study related to the present experiment, [13] enhanced the training of a statistical-parametric HMM-based system with signal-derived prosodic labels. They used AuToBI [14] to extract features for a female and male voice in the ARCTIC database and showed improved duration and pitch trajectory prediction.

2. Motivation

This work is part of the Wikispeech project [15]. The objective of Wikispeech is to deliver freely available, Wikipedia-optimised text-to-speech through Wikimedia Foundation's server architecture. We aim to improve the realisation of prosody in Wikispeech by including signal-driven prominence features and by enabling parametric control of prominence and emphasis. By making it possible to control prominence realisation, we will be able to model higher level contextual features such as given-new information or the effects of contextual predictability on prominence, such as word surprisal. These are particularly relevant to the nature of Wikipedia texts, e.g., where repetitions of the entry word and "surprising", infrequent terms occur.

In DNN-based synthesis (or in statistical parametric speech synthesis in general) the aim is to re-create the acoustic signal given a set of linguistic features derived from input text. In this process, prominence is implicitly captured and modeled as part of the output acoustics, but it is not annotated in the training data. The linguistic input contains features that are known to correlate with prominence, e.g. at the word level (content word vs. function word) or at the phrase level (e.g. word position in the phrase) but there are no explicit features that correspond to actual *realised* prominence.

There is a good reason for this: it must be possible to drive a text-to-speech synthesiser solely from features that can be derived from text. Requiring explicit prominence annotation is not a feasible approach for a general purpose system. Nevertheless, there are cases (see above) where it is desirable to have explicit control of prominence in a TTS system. The difficulty, however, resides in the fact that prominence is manifested in a large number of acoustic correlates (duration, F0, energy, spectrum) and manipulating it post-hoc is not feasible.

Thus, we need a model that allows for an explicit control of prominence whenever it is desired, for example, if a partic-

http://dx.doi.org/10.21437/Interspeech.2017-1355

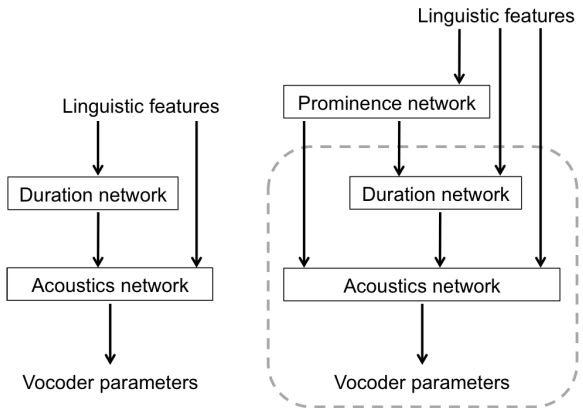

Figure 1: *Left: standard Merlin DNN synthesis architecture. Right: proposed architecture with explicit prominence modelling. The dashed line delimits the components used in the current experiment.*

ular word needs to be emphasised. At the same time, the system should not *require* explicit prominence control to produce a good default synthesis. In other words, it should be possible to override or augment the default prominence at will.

3. Proposed model

We propose to accomplish this by adding an explicit representation of prominence in the DNN synthesis pipeline, to complement the linguistically derived features. The prominence feature is represented at a word- or syllable level, by a continuous value that represents actual realised prominence in the training data (estimated using automatic or manual annotation, see below). The prominence feature augments the vectors of linguistically derived features in the input to the duration and acoustics networks. A separate DNN network (see Fig. 1) predicts the prominence feature from the linguistic input. This means that it will be possible to run the synthesis pipeline using only linguistic features as input, which is one of the requirements, but also to augment the prominence of individual words or syllables by modifying the prominence feature of that word or syllable.

Figure 1 compares the proposed model to the standard model used in the demo voices of the Merlin DNN-based synthesis system. The dashed line delimits the experiments described in the current paper: evaluating the effectiveness of explicitly including prominence as a feature in a DNN synthesis system, but without including the prominence network yet. This means that in the current experiments, explicit prominence values have been provided for each word at synthesis time.

4. Experiment

The present experiment aims to improve prosody realisation in TTS by advancing and implementing models of prominence. Towards this aim, we extract prosodic prominence from speech and explore the feasibility of using explicit prominence information in a statistical parametric speech synthesis system.

Specifically, we first compare the performance of an automatic prominence tagging algorithm [5] with the prominence judgments of human raters - both phonetic experts and non-experts. After this evaluation, we use the automatically-derived feature in an experimental system using Merlin - a parametric DNN-based TTS system. Finally, we conduct separate listening tests, a) comparing the level of prominence control, and b) the naturalness of the generated data. Performance is assessed against the baseline Merlin system; we also compare the experimental levels of prominence control with one another.

5. Methods

5.1. Training corpus

We use the CMU ARCTIC database [16], built for the purposes of speech synthesis, consisting of standardised prompts read by professional voice talents. A characteristic of the ARCTIC corpus is that the utterances do not contain syntactic constructions that might elicit, e.g., contrastive levels of prominence and in general feature subtle prominence variation. This characteristic is an advantage in that automatic and human prominence annotation is more likely to be consistent across the database but might also provide a challenge in discerning prominence categories. For easy comparisons of results with previous work, we decided to apply our methods to this de facto standard TTS training corpus first, before we experiment with more prosodically expressive data.

5.2. Automatic prominence detection

We used the prominence tagger by [5] (PromTagger) to automatically annotate prominence in the SLT ARCTIC data. The PromTagger puts out continuous values for every vocalic nucleus in a sentence, z-score normalised relative to the utterance. These continuous values are then used to provide discrete predictions of whether a nucleus is prominent - the decision is taken on the basis of comparisons with two neighbouring values. The weighting of two acoustic parameters needs to be specified, the force accent (a combination of intensity and duration) and the pitch accent. We ran the PromTagger with the default settings developed for German: force accent weighted = 0.9 and pitch accent weighted = 0.4. There are also two additional parameters referring to the pitch accent alignment relative to the vocalic nucleus, these were also set on the default.

5.3. Human prominence rating

We evaluated the automatic prominence detection and classification by comparison to the ground truth provided by human prominence judgements on the ARCTIC data. Four American English native speakers rated 400 sentences from the ARCTIC database. We chose two ARCTIC voices, one female (SLT) and one male (RMS) for the rating task.

The raters used a three-level prominence scale (not prominent, maybe prominent, prominent) to judge 200 sentences and a four-level scale (+very prominent) to judge the other 200 sentences in two separate sessions. The order in which they used the three- or four-level scale was counterbalanced across raters. Each of the 200 sentences was presented twice, once as read by the female and once by the male voice talent, resulting in 800 stimuli. The order of all stimuli was randomised for each rater. The task was to mark the word or words in a sentence that are ”standing out from the other words”. That is, they were to mark those words that they heard as stronger or more prominent than the others. This instruction avoids the suggestion that they should rely on their linguistic knowledge or expectations regarding lexical stress or rhythmical patterns - but rather encourages them to pay attention to the acoustic variation.

To enable quick and easy prominence ratings we developed a prominence annotating interface: ProMark. The raters were

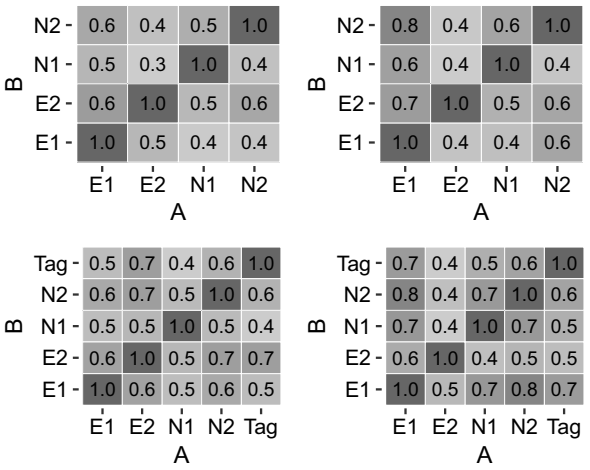

Figure 2: *Cohen's kappa values for the pairwise inter-rater agreement. Top left: ratings on the 3-level prominence scale; Top right: human ratings on the 4-level prominence scale between naive (N1, N2) and expert (E1, E2) raters. Bottom left: ratings on a reduced 0-1 scale between an automatic prominence tagger (Tag) and human raters. Left panel: reduced from 3-level; Right panel: reduced from 4-level.*

forced to choose at least one prominent word in every sentence by pressing a key assigned to the word. The word was then highlighted in shades of green depending on the choice of consecutive levels of prominence (cf. [17]). The raters could listen to the stimulus as many times they wished before rating.

The top panel in Figure 2 shows the inter-rater agreement results based on pairwise Cohen's kappa (κ) for the three-level scale on the left and the four-level scale on the right. Weighted κ (over the diagonal) is more appropriate for ordinal data such as prominence scales, since it accounts for how far from the diagonal each rating is. The mean of the weighted κ (equivalent to the Intra Class Correlation) for the 3-level scale is 0.54 and 0.64 for the 4-scale level. This indicates good agreement among raters overall and a stronger agreement for the 4-scale prominence ratings. The inter-rater agreement values were similar for both ARCTIC voices. We report κ values pooled for both voices in Figure 2. Inspecting the pairwise weighted κ values also suggest some very good agreement between some rater pairs, on the level of 0.8. The naive rater N1 has the lowest overall agreement values.

In the bottom panel of Figure 2, we compare the output of the PromTagger with the raters' annotations for the same subset of ARCTIC sentences. The tagger gives a prominent-non prominent hypothesis (0/1) for each vowel in a word. We mapped these hypotheses onto words, if at least one vowel was tagged as 1, the whole word was marked as prominent. Human ratings coming from the 4-level scale annotation were reduced to a 0/1 oppositions to enable a comparison with the 0/1 tagger output. The PromTagger reaches a good agreement especially against phonetically informed raters, on a par with the other more naive rater against the informed raters (weighted $\kappa = 0.7$).

We conclude that the PromTagger reaches a similar level of accuracy in prominence detection to native speaker human raters in the binary decision. We subsequently used the tool to automatically tag prominence in the rest of the the ca. 1200 utterances in the SLT ARCTIC corpus.

5.4. Implementation in Merlin

We used the HTS-like workflow option [18] to introduce the PromTagger-derived feature to the set of training labels in the experimental Merlin system. The experiments were based on the Merlin example scripts (ARCTIC-SLT full voice).

The PromTagger feature was extracted per syllable nucleus for all utterances in the training corpus, resulting in a real-valued number in the range 0 - 1.2 for each syllable. This feature was multiplied by 100 and rounded to an integer value and transfered to the word level by assigning the maximum syllable prominence in every word as the word prominence. The word prominence feature was then added to the state alignment labels in the training data for the duration and the acoustic models. The question file was amended to include the prominence feature (CQS), leaving all the other default features unchanged. The full Merlin default voice was trained as a baseline and the full prominence-enhanced voice as the experimental system.

Finally, we created a custom synthesis script that accepts input text with a prominence value for every word, that was used to generate the stimuli files used in the listening tests.

5.5. Crowdsourced listening tests

We used six randomly chosen Haskins Syntactic Sentences [19] for the diagnostic and naturalness tests. The HSS include frequent American English words in syntactically correct, meaningless sentences to minimise the effects of contextual cues in intelligibility tests. These sentences are equally suited for testing prominence perception, since local context and frequency influence prosodic prominence in natural speech as well. Additionally, we simplified the sentences into the form NounPhrase1 + Verb + NounPhrase2 + yesterday, e.g., "The leg shut the shore yesterday". We added "yesterday" to reduce final boundary position effects on prominence realisation of the second noun.

After training, each of the six diagnostic sentences was generated using the default baseline, and the experimental, prominence-enhanced voice. Using the experimental voice, we synthesised five levels of relative prominence set over the two nouns: a) 0-200, b) 50-150), c) 100-100, d) 150-50, e) 0-200 for each sentence. The values for all other words were set to zero. The audio files of all the stimuli are available in the digital version of this publication.

We used the Crowdflower platform for the listening tests. The workers saw a randomised list of the prominence-enhanced and baseline sentences and were asked to select "Which word is the stronger one?". We obtained 1080 ratings from 180 raters. Raters were self-reported speakers of American English.

6. Results

6.1. Prominence control

In broad focus, the default nuclear accent placement in English falls on the last argument of the verb, the second noun (N2) [20]. We hypothesise that this will bias the listeners to hear the prominence on N2 when the prominence values are set to be equal between N1 and N2 (condition c)) and in the default Merlin voice. However, we should see a clear effect in the prominence ratings on either N1 or N2 when the ratio of the set prominence values is more than 1, i.e.: settings a), b) and d), e).

Figure 3 shows the proportions of ratings that agreed or disagreed with our hypotheses concerning which noun was prominent. We compared sentences with explicitely set prominence values in conditions a) through e) with the Merlin baseline.

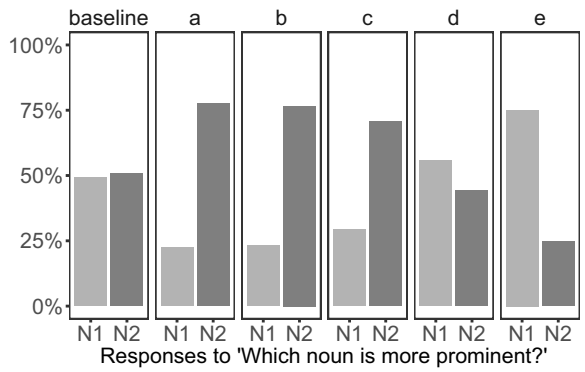

Figure 3: *Crowdsourced responses on whether the first (N1) or the second noun (N2) was rated as prominent. Panels: results for sentences synthesised with the baseline Merlin vs. the experimental system using five relative prominence settings (a)-e)) for N1 and N2 (see Table 1).*

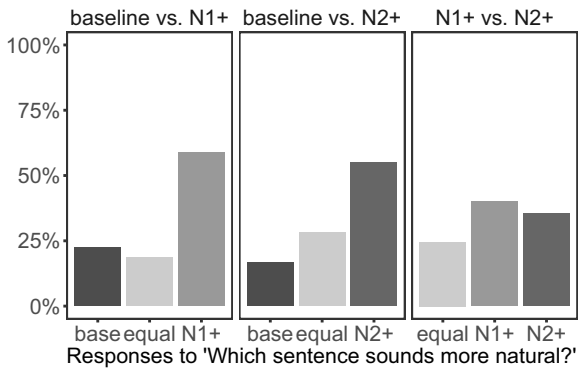

Figure 4: *Crowdsourced naturalness test responses on which sentence (N1+, N2+) was perceived as more natural or equally natural (equal) than the baseline (base). Panels: comparison of different settings and the baseline.*

Table 1: *Generalised mixed model (logit) for the binomial response (correct/not correct word is rated prominent) vs. the baseline default Merlin voice (reference level for Setting).*

Setting					
Condition	N1	N2	Log odds	z-value	p-value
a)	0	200	1.44	6.05	<.001
b)	50	150	1.35	5.66	<.001
c)	100	100	0.95	4.22	<.001
d)	150	50	0.19	0.93	=.35
e)	200	0	1.04	4.62	<.001

Table 1 presents results of a mixed-effects logistic regression estimating the effect of prominence control on whether the response was correct or not, with the Merlin baseline as the reference level. The model included a random intercept for item to account for the variance introduced by the specific sentences to the responses.

In Figure 3 we see that the prominence between the two nouns in the baseline was rated as approx. equal, suggesting that the crowdsourcing workers picked one or the other noun at random. In line with our hypothesis, the 100-100 setting shows a bias towards the second noun, the argument of the verb. Similarly, if prominence is explicitly set on the verb argument (conditions a), b)), there is a strong effect of that condition on the ratings. The logistic model shows that these effects are statistically significant vs. the baseline, with the largest effect sizes in the model. The augmentation of prominence on the first noun in the sentence (conditions d), e)) has a statistically significant effect only in case of condition e) when the prominence is clearly set to override the bias towards N2.

The model also evidenced that the higher the average trust the crowdsourcing platform gave to each participating worker's performance, the probability of a correct response increased (log odds = 5.31, p <.001).

6.2. Naturalness

We conducted a separate naturalness study asking crowdsourced participants to rate "Which sentence sounds more natural?" For each of the six HSS sentences they gave pairwise ratings of the baseline Merlin sentence and the same sentence synthesised using either the b) (N2+) or e) (N1+) prominence control settings. This means they evaluated settings that were found to effectively differentiate the prominence of N1 vs. N2 in the listening test, against the baseline. We obtained 720 judgments from 20 raters.

The results are shown in Figure 4. It is clear that the crowdsourced listeners found the prominence-enhanced synthesis not only mostly equal in quality to the baseline but also frequently as more natural than the baseline. Comparing the N1+ setting to the N2+ setting in the right panel, no preference in naturalness is apparent if one or the other noun is emphasised.

7. Conclusions

We introduced a model for prominence control operationalised as a separate network in DNN-based SPSS. The proposed architecture complements the text-derived features with a signal-derived prominence feature estimated from training data.

In the first steps to providing a prominence feature, we delivered the ground truth from human raters for the ARCTIC database. We observed that a four-level rating scale provides a better agreement between the human raters, similarly to studies in [12, 21]. We also showed that an automatic tagging algorithm agrees well with the native raters.

We also experimentally evaluated part of the proposed architecture by directly including a prominence feature automatically extracted from the training corpus. Listening tests showed we were able to control word prominence using this method in an American English voice built with Merlin.

In a naturalness test, we found that the prominence-enhanced stimuli were substantially better rated than the default voice baseline, confirming observations that current systems suffer from the limited range of prosodic variation [6].

We are currently testing synthetic prominence control on the syllable level. The digital version of this publication includes examples of a sentence excerpted from Wikipedia: one synthesised using a system with syllable-level prominence control and one in a default Merlin rendition.

8. Acknowledgements

This research was funded by The Swedish Post and Telecom Authority (PTS) and by an ICT-TNG postdoctoral grant.

9. References

[1] P. Wagner, A. Origlia, C. Avesani, G. Christodoulides, F. Cutugno, M. D'Imperio, D. Escudero Mancebo, B. Gili Fivela, A. Lacheret, B. Ludusan *et al.*, "Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence," in *International Congress of the Phonetic Sciences*, 2015.

[2] D. Arnold, P. Wagner, and B. Möbius, "The effect of priming on the correlations between prominence ratings and acoustic features," in *Prosodic Prominence: Perceptual and Automatic Identification (Speech Prosody 2010 Workshop)*, 2010.

[3] J. Hirschberg, "Speech synthesis, Prosody," *Encyclopedia of Language and Linguistics,*, pp. 49–55, 2006.

[4] S. Kakouros and O. Räsänen, "3PRO–an unsupervised method for the automatic detection of sentence prominence in speech," *Speech Communication*, vol. 82, pp. 67–84, 2016.

[5] F. Tamburini, C. Bertini, and P. M. Bertinetto, "Prosodic prominence detection in Italian continuous speech using probabilistic graphical models," in *Proceedings of Speech Prosody*, 2014, pp. 285–289.

[6] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, 2016.

[7] A. Rosenberg, E. Cooper, R. Levitan, and J. Hirschberg, "Cross-language prominence detection," in *Speech Prosody*, 2012.

[8] K. Evanini and K. Zechner, "Using crowdsourcing to provide prosodic annotations for non-native speech." in *INTERSPEECH*, 2011, pp. 3069–3072.

[9] M. Hasegawa-Johnson, J. Cole, P. Jyothi, and L. R. Varshney, "Models of dataset size, question design, and cross-language speech perception for speech crowdsourcing applications," *Laboratory Phonology*, vol. 6, no. 3-4, pp. 381–431, 2015.

[10] J. Cole, T. Mahrt, and J. Roy, "Crowd-sourcing prosodic annotation," *Computer Speech & Language*, 2017.

[11] A. Windmann, I. Jauk, F. Tamburini, and P. Wagner, "Prominence-based prosody prediction for unit selection speech synthesis," *Proceedings of Interspeech 2011*, 2011.

[12] M. Mehrabani, T. Mishra, and A. Conkie, "Unsupervised prominence prediction for speech synthesis," *Power*, vol. 2, no. 1.6, pp. 1–3, 2013.

[13] F. Tesser, G. Sommavilla, G. Paci, and P. Cosi, "Experiments with signal-driven symbolic prosody for statistical parametric speech synthesis." in *SSW*, 2013, pp. 183–187.

[14] A. Rosenberg, "AuToBI-a tool for automatic tobi annotation." in *Interspeech*, 2010, pp. 146–149.

[15] J. Andersson, S. Berlin, A. Costa, H. Berthelsen, H. Lindgren, N. Lindberg, J. Beskow, J. Edlund, and J. Gustafson, "WikiSpeech – enabling open source text-to-speech for Wikipedia," in *Proceedings of the 9th ISCA Workshop on Speech Synthesis*, 2016.

[16] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[17] S. Al Moubayed, G. Ananthakrishnan, and L. Enflo, "Automatic prominence classification in Swedish," in *Speech Prosody 2010, Workshop on Prosodic Prominence, Chicago, USA*, 2010.

[18] S. Ronanki, G. E. Henter, Z. Wu, and S. King, "A template-based approach for speech synthesis intonation generation using LSTMs," in *Proc. Interspeech*, San Francisco, USA, September 2016.

[19] P. Nye and J. Gaitenby, "The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences," *Haskins Laboratories Status Report on Speech Research*, vol. 37, no. 38, pp. 169–190, 1974.

[20] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[21] C. Jensen and J. Tøndering, "Choosing a scale for measuring perceived prominence," in *Ninth European Conference on Speech Communication and Technology*, 2005.