10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

PROMIS: a statistical-parametric speech synthesis system with prominence control via a prominence network

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

We implement an architecture with explicit prominence learning via a prominence network in Merlin, a statistical-parametric DNN-based text-to-speech system. We build on our previous results that successfully evaluated the inclusion of an automatically extracted, speech-based prominence feature into the training and its control at synthesis time. In this work, we expand the PROMIS system by implementing the prominence network that predicts prominence values from text. We test the network predictions as well as the effects of a prominence control module based on SSML-like tags. Listening tests for the complete PROMIS system, combining a prominence feature, a prominence network and prominence control, show that it effectively controls prominence in a diagnostic set of target words. The tests also show a minor negative impact on perceived naturalness, relative to baseline, exerted by the two prominence tagging methods implemented in the control module.

Index Terms: speech synthesis, prosodic prominence, expressive speech synthesis, control

1. Introduction

Text-to-speech systems have been greatly improving in terms of quality, defined as intelligibility and naturalness, as well as in terms of expressivity. The field has moved towards rendering longer texts and narratives [1] that require advanced prosodic modeling of higher-level linguistic features and long-distance dependencies such as repeated mentions, information status and information structure. However, the improvements in overall realism have been happening at the expense of parametric control. State-of-the art machine learning-based solutions have sacrificed the pursuit of controllability of low-level (pitch, duration) and, to a lesser extent, high-level concepts (prominence, speaker identity, emotion) as greater gains are to be made in increasing training data size and leveraging new neural architectures to improve quality [2].

Low-level control, particularly, has been on the research back burner in contrast to e.g. the formant synthesis era, when explicit modelling and control of acoustically meaningful features was prevalent. The subsequent paradigm shift, i.e. concatenative signal generation methods proved to possess only limited ability to provide a continuum of acoustic cues in response to input commands, with the exception of MBROLA ([3] based on PSOLA [4]) allowing for control of pitch and duration given a sequence of allophones. In concatenative TTS, high-level concepts, such as prominence, can be explicitly placed on some parts of the input text, for example via prosodic SSML tags. The acoustic correlates of prominence (duration, f0, intensity, spectrum) are then manipulated post-training [5, 6]. But as of yet, such strategies have not yielded realisations of acceptable quality.

As we argue elsewhere [7], state-of-the-art systems have now come to a point where combining superior realism and some levels of control should be possible. Statistical-parametric speech synthesis, as well as neural sequence-to-sequence systems are able to control arbitrary concepts by learning mappings via supervised machine learning. In terms of high-level features, the work in [8, 9, 10, 11] shows that manipulation of factors such gender, age, identity or speaking style [12] in DNN-based systems is possible. Regarding specifically the control of prosody, system improvements in terms of prosodic expressivity would likely make listening to long stretches of synthetic speech, for one, less tiring [13]. It was shown that suboptimal rendering of pitch contours and speech rate have an influence on perception and comprehensibility [14] while attention and memory are affected by both segmental and prosodic quality [15, 16, 17] of synthetic speech.

Finally, in this work, we also aim for societal impact. PROMIS was developed as part of the Wikispeech project [18] in which the general objective is to deliver freely available, Wikipedia-optimised text-to-speech through Wikimedia Foundation's server architecture. A TTS system with improved expressiveness and prominence control would find good use in synthesising Wikipedia: the input are long, domain-homogenous texts in which modeling information beyond the current, short utterance is especially desirable [19]. Additionally, among millions of Wikipedia users, there are those that live in low-literacy regions or suffer from reading problems due to disability. Consequently, adequate control of prosodic prominence to differentiate given vs. new information, surprising and infrequent words, repeated mentions of the entry word etc. would come to a direct benefit of people with reading problems. Ultimately, free and easily available TTS for Wikipedia would offer those users much needed access to information.

2. Approach and architecture

With PROMIS we aim for a model that learns to predict prominence from text and allows to control realised prominence explicitly whenever a particular word or syllable needs to be emphasised. The model does not, however, require explicit prominence control to produce good default synthesis. We use Merlin [20], a statistical-parametric TTS system with a DNN-based learning engine to construct PROMIS. Our concept adds several components to the default Merlin architecture: a continuous prominence feature, a separate prominence network and a prominence control module. The architecture of PROMIS is presented in Fig. 1.

The prominence feature is operationalised at a word- or syllable level by a continuous value that represents actual, realised prominence in the training data, here estimated using automatic annotation. The feature is an explicit representation of promin-

10.21437/SSW.2019-46

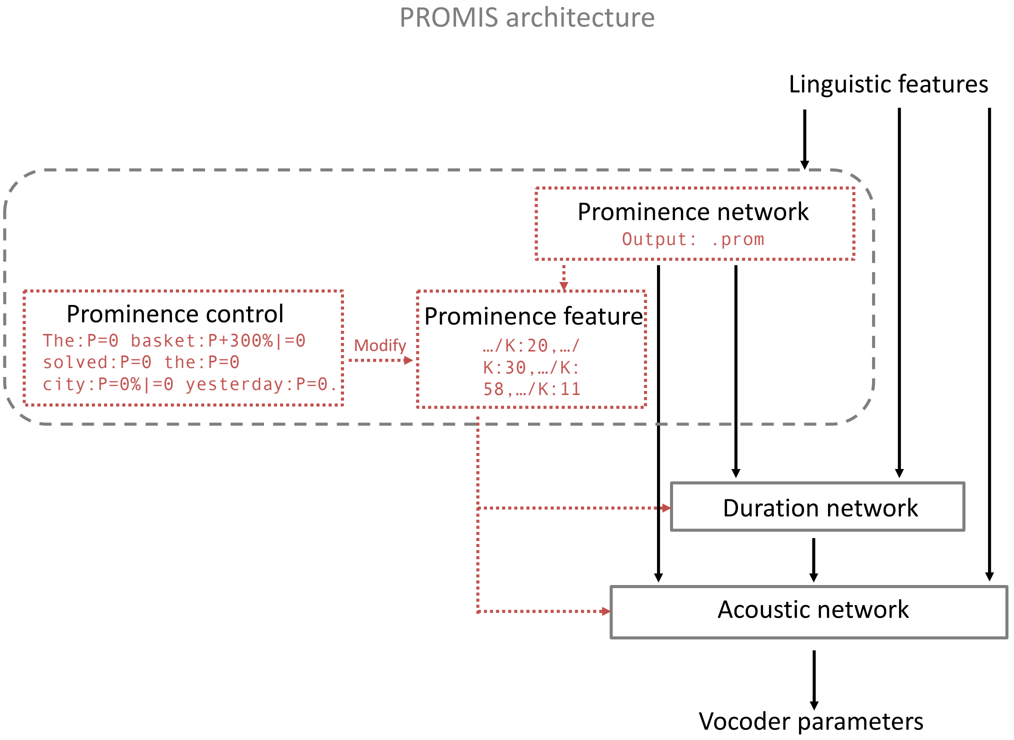

Figure 1: *PROMIS architecture with prominence control scenario (red, dotted elements). The dashed line delimits the components implemented and evaluated in the current experiments.*

ence in the DNN synthesis pipeline (rather than a representation implicitly captured as part of the output acoustics).

The prominence network learns to predict prominence from the linguistic input using the signal-driven prominence values extracted from the training data. The standard, text-based, linguistic feature set, serve as input to the duration and acoustic networks as well as to prominence network. The prominence network is constructed and parametrised analogously to the duration network. The prominence network, via the learned prominence feature, augments the vectors of linguistically derived features in the input to the duration and acoustics networks. The purpose of the independent prominence network is to make it possible to run the synthesis pipeline using only standard linguistic features as input, as well as to augment the prominence of individual words or syllables by modifying the prominence feature of that word or syllable. The modification is enabled by the prominence control module at synthesis time.

The control module uses SSML-like tags and other parameters that enable to explicitly modify the prominence of specific syllables in the output of the prominence network.

In a previous study, we showed that including a signal-driven prominence feature enables the DNN system to synthesise perceivable differences in prominences put on monosyllabic words. The synthesised stimuli with prominence variability were also more frequently rated as more natural, compared to the Merlin baseline. In the present work (dashed, grey line) in Fig. 1, we implement the prominence network, evaluate its predictions and the performance of the control module.

3. Implementation

We train the system on the CMU ARCTIC database [21] using the female voice SLT. The training data consists of 1132 utterances (split into training=1000, validation=66, and test=66 sets).

3.1. Prominence feature

The prominence network learns to predict a prominence feature from the linguistic input given prominence vectors estimated automatically using the PromTagger by [22]. In [8], we evaluated the automatic prominence tagging method [22] against a human gold standard obtained from prominence ratings by native American English speakers. We found that the tagger agrees well with decisions taken by raters (weighted Cohen's κ = 0.7). It reaches an adequate level of accuracy in prominence detection compared to the raters in a binary decision task (is a syllable nucleus prominent or not-prominent?).

The extracted prominence vectors represent realised prominence for each utterance in the training data. The PromTagger puts out continuous values for every vocalic nucleus in a sentence, z-score normalised relative to the utterance, resulting in a real-valued number in the range 0 - 1.2 for each syllable. This feature is multiplied by 100 and rounded to an integer value. For word-based prominence control, this value is transferred to the word level by assigning the maximum syllable prominence in every word as the word prominence. In the present work, we evaluate syllable-based prominence prediction.

3.2. Prominence network

The prominence feature and the standard linguistic features are included in the duration and acoustic network input layers in the default scenario, if manipulation of prominence is desired, the prominence feature is modified as specified in the control module. The dimension of the complete feature input is 417. The duration and prominence feature have output values of 5 each, equivalent to one value for each of the 5 states. The input and output features were normalised as described in [20]. The WORLD vocoder is used to extract acoustic parameters and generate speech waveforms. The rest of the network parameters follow the default Merlin specifications: the DNN models have 6 hidden layers with 1024 units each. The outputs of the acoustic DNN consist of 60 mel-cepstral coefficients, a log F0 value, one-dimensional band aperiodicity coefficients with their Δ, Δ^2 features and a voiced-unvoiced feature extracted at 5 msec frame intervals.

3.3. Control at synthesis

The prominence network, in the prominence control mode, receives tags, modeled on SSML tags from the control module via a custom script. We use two tagging methods here:

- absolute prominence tags (ABS) that assign a specific prominence value to a syllable we wish to control,
- relative value tags (REL) that proportionally augment the prominence value predicted by the network for that syllable.
- symbolic tags such as “strong" or “reduced" that are assigned to particular syllables and implicitly set to absolute values of prominence.

In this study, we evaluate ABS and REL.

4. Evaluation

4.1. Prominence control

First, we test whether the proposed architecture is successful in realising prominence control. To generate stimuli for evaluation, we follow the concept of Haskins Syntactic Sentences (HSS) [23]. The HSS include frequent American English words in syntactically correct, semantically unpredictable sentences (SUS) to minimise the effects of contextual cues in intelligibility tests. SUS is equally suited for testing prominence perception, since local context and frequency effects, that influence prosodic prominence in natural speech, are minimised. We simplify the sentences into the form:

```
NounPhrase1 (N1) + Verb + NounPhrase2 (N2) + yesterday
         'The leg shut the shore yesterday'
```

We add “yesterday" to reduce final boundary position effects on prominence realisation of the second noun.

We use a stimuli list including one-, two- and three-syllable target nouns. To generate sentences with monosyllabic nouns, we randomly pick them from the HSS set. Multisyllabic nouns (two- and three-syllable nouns) were designed by us following the same sentence template. We include the full list of stimuli in the appended Table 4.

The total number of tested sentences was 18, six per each of three tested syllable counts. Each of the 18 sentences was synthesised using a baseline, where prominence control parameters are all set to zero, and using PROMIS with ABS and REL tags as diagnostic sets.

Table 1: *ABS method: prominence tagging values for five relative prominence settings used in this experiment. N1 denotes the first noun in the sentence template, N2, the second noun.*

Prominence setting	**N1**	**N2**
N2++	P=0	P=200
N2+	P=50	P=150
N2=N1	P=100	P=100
N1+	P=150	P=50
N1++	P=200	P=0

Table 2: *REL method: prominence tagging values for five relative prominence settings used in this experiment. N1 denotes the first noun in the sentence template, N2, the second noun.*

Prominence setting	**N1**	**Noun 2**
N2++	P=0%	P+300%
N2+	P+100%	P+200%
N2=N1	P+200%	P+200%
N1+	P+200%	P+100%
N1++	P+300%	P=0%

With the ABS tagging method, we synthesised five levels of prominence set on the stressed syllable of the two target nouns, as exemplified in Table 1. Consequently, the monosyllabic nouns and the five levels of prominence are the same in the ABS set as the ones tested in [8] (we did not test multisyllabic nouns [8]). With the REL tagging method, we synthesise five levels of prominence set on the stressed syllable of the two target nouns, as exemplified in Table 2.

Regardless of the tagging method or syllable count of the target noun, we assign prominence values only to the stressed syllable of the target noun. All other syllables in the noun and in the sentence have the prominence value set to zero, that is, realise the default parameters of the system.

We used the Figure-Eight crowdsourcing platform for listening tests. The workers saw a randomised list of the PROMIS generated sentences including the baseline and were asked to select "Which word is the stronger one?" where the choice options were either N1 or N2. Note that setting N1=N2 in both tagging methods puts equal values of prominence on both nouns, so the hypothesised responses in this setting should be randomly distributed. The listeners evaluated a counterbalanced list of stimuli created both with the ABS and REL method. We obtained 2060 ratings from 56 raters in the test for prominence control. Raters were self-reported speakers of American English.

4.2. Results: prominence control

Figure 2 shows the proportions of ratings that agreed or disagreed with our hypotheses concerning which noun was prominent. The listeners compared sentences generated by PROMIS with five modified prominence value settings, as shown in Tables 1 and 2, to the Merlin baseline.

Table 3 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 rater and for item to account for the variance introduced by the specific sentences to the responses. We also entered an interaction

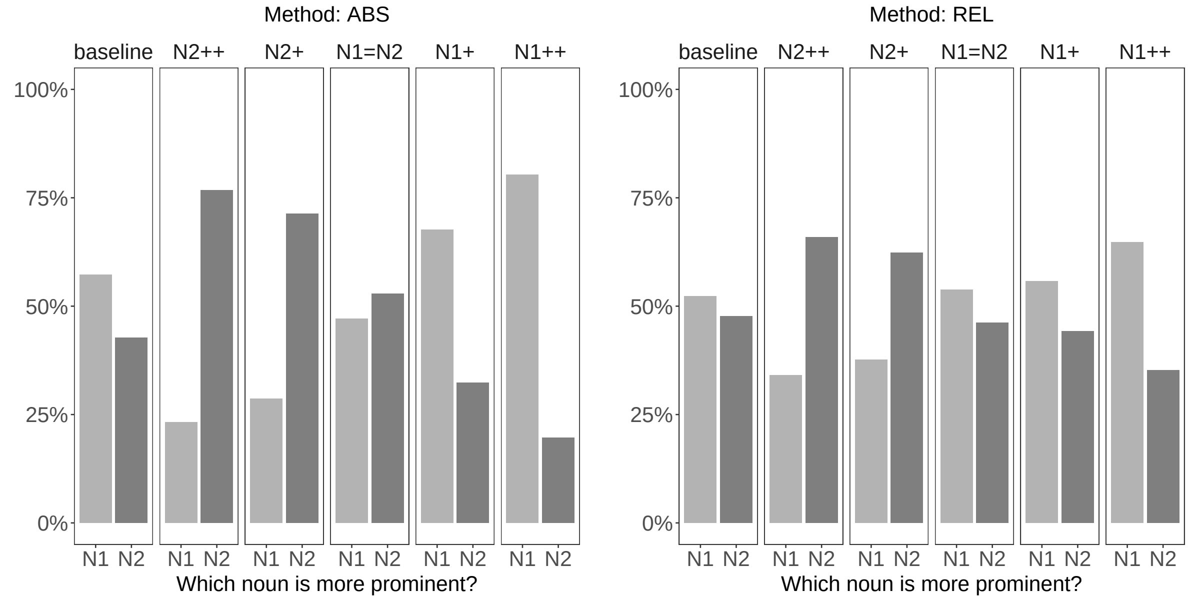

Figure 2: *Crowdsourced responses on whether the first (N1) or the second noun (N2) was rated as prominent, depending on the control module tagging method (left: ABS or right: REL). Panels present results for sentences synthesised with the baseline Merlin vs. the PROMIS system using five relative prominence settings for N1 and N2. See Table 1 (ABS) and Table 2 (REL) for tested tagging values.*

between the tagging method (ABS vs. REL) and prominence setting (N2++, N2+, N2=N1, N1+, N1++) to estimate the potential effect of the tagging method on the ratings of prominence control in each condition.

We find that the listeners correctly rated the words that were controlled with our two tagging methods in all settings. Effect sizes indicate that the most reliable control was achieved in the more expressly controlled prominence settings, condition N2++, and N1++ and other settings follow the expected pattern. We do find, however, that the REL method is statistically significantly less effective in achieving control than the ABS method: the interaction of method and setting was significant (apart from setting N2=N1, where prominence over N1 and N2 is equal) with negative log-odds values in all conditions.

This is to be expected, since in the REL method, the output prominence values that are under parametric control vary and depend on the initial prediction of the model for the default rendition of the sentence. That is if a noun in our sentence did not receive a lot of prominence in the default specification, it will also be less prominent under REL prominence control, compared to the ABS values that strictly override the predictions.

4.3. Naturalness

Subsequently, we conducted a naturalness study asking listeners to rate "Which sentence sounds more natural?" Crowdsourced listeners gave preference ratings for each of the eighteen stimuli choosing between the baseline Merlin sentence and the same sentence synthesised using prominence control. They also compared the naturalness between stimuli with prominence control set on either N1 or N2. The stimuli were presented pairwise and in a counterbalanced order. The prominence control settings submitted to the test were N2+ or N1+ in the ABS stimuli block and N2++ or N1++ in the REL stimuli block. Thus, the listeners evaluated “first best" settings found to reliably differentiate prominence in the control study. We obtained 802 judgments from 21 raters in the ABS block and 965 judgements from 28 raters in the REL block.

Table 3: *Generalised mixed model (logit) for the binomial response (correct/not correct word is rated prominent) vs. the baseline default Merlin voice (the reference level for Setting). Also shown: the negative effect of the REL Method on correctness scores in interaction with Setting, relative to the ABS method (the reference level for Method).*

Setting (× Method):	Log odds	z-value	p-value
N2++	1.56	6.38	<.001
N2+	1.28	5.46	<.001
N2=N1	0.39	1.77	=.07
N1+	1.09	4.73	<.001
N1++	1.83	7.26	<.001
N2++ × REL	-0.83	-2.5	<.05
N2+ × REL	-0.66	-2.0	<.05
N2=N1 × REL	-0.48	-1.5	=0.12
N1+ × REL	-0.76	-2.4	<.05
N1++ × REL	-1.18	-3.5	<.001

4.4. Results: naturalness

The results are shown in Fig. 3. Crowdsourced listeners found the prominence modified synthesis using the ABS method mostly equal in quality to the baseline (43% for N2+, 37% for N1+) and at times more natural than the baseline (24% for N2+ and 26% for N1+). The REL method was less successful in the naturalness test in that the preference for the baseline was even more pronounced than in case of the ABS method: the baseline was picked as more natural at least ca. 50% of the time relative to the stimuli manipulated with REL (57% for N2++ and 49% for N1++).

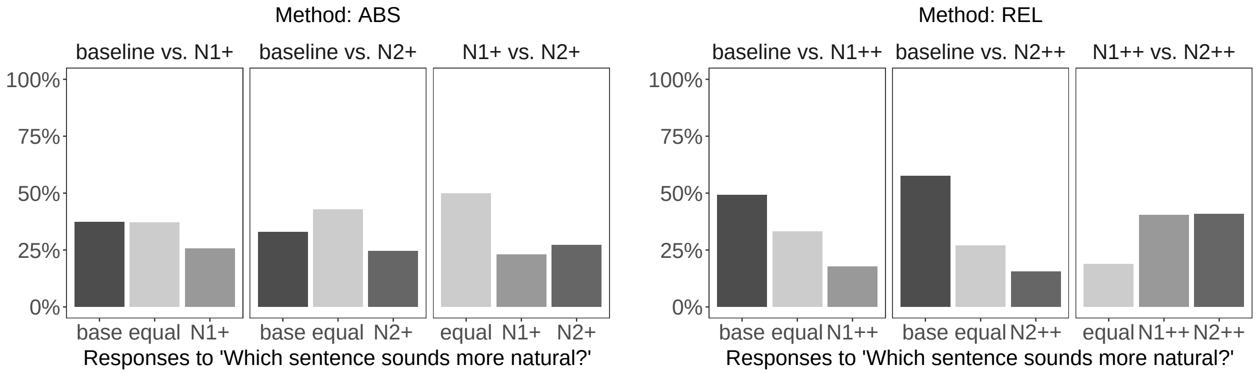

Figure 3: *Crowdsourced naturalness test responses on which sentence prominence-controlled sentence (N1+ or N2+ for method ABS and N1++ or N2++ for method REL) was perceived as more natural or equally natural (equal) than the baseline (base). Panels: comparison of different control settings and the baseline. ABS method (left) vs. the REL method (right).*

In a direct pairwise comparison of the prominence-modified stimuli regarding naturalness, shown in the right panels of Fig. 3 for both control methods, no clear preference in naturalness was apparent depending on whether one or the other noun was emphasised.

5. Discussion and future work

We presented the concept and architecture of PROMIS, a text-to-speech system that integrates several new components into a statistical-parametric pipeline enabling the control of prosodic prominence. The system augments prosodic prominence on selected syllables or words via a separate network that predicts it from text on the basis of a signal-driven representation, a prominence feature. The augmentation is delivered via SSML-like tags provided by a control module.

The combination of the prominence feature, the network and the control module allows for the possibility, but not necessity, to modify prominence values in synthesis. Crowdsourced listening tests showed that this architecture is effective in controlling prominence. Listeners could robustly differentiate between two PROMIS-generated nouns in a sentence frame, where either one noun or the other was prosodically emphasised. We also showed that both tested methods of prominence feature modification, an absolute one, that sets prominence to an explicitly specified value, and a relative one, that augments the predicted value proportionally, offer reliable control. The standard, unmodified Merlin implementation served as a baseline for these tests.

In the future, we would like to compare the predicted output of the prominence network without any modifications to the modified stimuli, as a form of setting a different baseline. We also see the need to experiment with more tagging methods and ways to scale the to-be-modified syllables against the phonetic context of the sentence. The results of the naturalness test, showing minor negative impact of both the absolute and the relative method relative to the baseline, call for more investigation into how the immediate context surrounding the manipulated target words influences the realism ratings of the whole sentence.

Another way to improve the balance between realism and control - that we argued for as a direction generally worth exploring in TTS at the outset of this work - is to integrate the concept of PROMIS into a neural, sequence-to-sequence system such as the Tacotron [24]. These kinds of systems offer synthesis quality evaluated to be statistically indistinguishable from natural speech [7] but feature no extensive controllability of low- or high-level concepts, such as prominence, as of yet.

6. Acknowledgements

Grant support for ZM, JB: Swedish Research Council no. 2017-02861; HB: The Swedish Post and Telecom Authority (PTS); JG: Swedish Research Council no. 2013-4935.

7. Appendix: synthetic stimuli list

Table 4: *Target nouns and carrier sentences. The full carrier sentence template is e.g.:* `The box paid the tree yesterday.` *etc.*

N1	N2	carrier sentence
box	tree	... box paid the tree ...
street	top	... street said the top ...
sea	bag	... sea led the bag ...
wing	shout	... wing met the shout ...
face	judge	... face put the judge ...
rain	drive	... rain sang the drive ...
window	money	... window wore the money ...
desert	office	... desert brought the office ...
message	flavour	... message jumped the flavour ...
basket	city	... basket solved the city ...
business	morning	... business read the morning ...
novel	building	... novel sent the building ...
celery	government	... celery let the government ...
cinema	gravity	... cinema wore the gravity ...
telephone	candidate	... telephone tried the candidate ...
hospital	envelope	... hospital knew the envelope ...
magazine	destiny	... magazine laid the destiny ...
restaurant	discipline	... restaurant shut the discipline ...

8. References

[1] Y. Maeno, T. Nose, T. Kobayashi, T. Koriyama, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, “Prosodic variation enhancement using unsupervised context labeling for hmm-based

expressive speech synthesis," *Speech Communication*, vol. 57, pp. 144–154, 2014.

[2] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: where do the improvements come from?" in *Proc. ICASSP*, 2016, pp. 5505–5509.

[3] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, "The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes," in *Proc. ICSLP*, 1996, pp. 1393–1396.

[4] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5-6, pp. 453–467, 1990.

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

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

[7] Z. Malisz, G. E. Henter, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: a discussion and an evaluation," in *Proc. ICPhS*, 2019.

[8] Z. Malisz, H. Berthelsen, J. Beskow, and J. Gustafson, "Controlling prominence realisation in parametric dnn-based speech synthesis," *Proc. Interspeech 2017*, pp. 1079–1083, 2017.

[9] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *Proc. ICASSP*, 2017, pp. 4905–4909.

[10] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Deep encoder-decoder models for unsupervised learning of controllable speech synthesis," *arXiv preprint arXiv:1807.11470*, 2018.

[11] I. Fukuoka, K. Iwata, and T. Kobayashi, "Prosody control of utterance sequence for information delivering," *Proc. Interspeech 2017*, pp. 774–778, 2017.

[12] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, and et al., "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018, pp. 5180–5189.

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

[14] L. M. Slowiaczek and H. C. Nusbaum, "Effects of speech rate and pitch contour on the perception of synthetic speech," *Human Factors*, vol. 27, no. 6, pp. 701–712, 1985.

[15] C. Delogu, S. Conte, and C. Sementina, "Cognitive factors in the evaluation of synthetic speech," *Speech Communication*, vol. 24, no. 2, pp. 153–168, 1998.

[16] E. Rodero, "Effectiveness, attention, and recall of human and artificial voices in an advertising story. prosody influence and functions of voices," *Computers in Human Behavior*, vol. 77, pp. 336–346, 2017.

[17] M. Wester, O. Watts, and G. E. Henter, "Evaluating comprehension of natural and synthetic conversational speech," in *Proc. Speech Prosody 2016.*, 2016.

[18] 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.

[19] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, p. 006, 2014.

[20] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

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

[22] F. Tamburini and P. Wagner, "On automatic prominence detection for german," in *Proceedings of Interspeech 2007*, Antwerp, Belgium, 2007, pp. 1809–1812.

[23] 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.

[24] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and et al., "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proc. ICML*, 2018, pp. 4693–4702.