INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Improving speech synthesis with discourse relations

Adèle Aubin[1], *Alessandra Cervone*[2], *Oliver Watts*[1], *Simon King*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh
[2]Department of Information Engineering and Computer Science, University of Trento
adele.aub@gmail.com, alessandra.cervone@unitn.it, {oliver.watts, Simon.King}@ed.ac.uk

Abstract

This paper explores whether adding Discourse Relation (DR) features improves the naturalness of neural statistical parametric speech synthesis (SPSS) in English. We hypothesize first - in the light of several previous studies - that DRs have a dedicated prosodic encoding. Secondly, we hypothesize that encoding DRs in a speech synthesizer's input will improve the naturalness of its output. In order to test our hypotheses, we prepare a dataset of DR-annotated transcriptions of audiobooks in English. We then perform an acoustic analysis of the corpus which supports our first hypothesis that DRs are acoustically encoded in speech prosody. The analysis reveals significant correlation between specific DR categories and acoustic features, such as F0 and intensity. Then, we use the corpus to train a neural SPSS system in two configurations: a baseline configuration making use only of conventional linguistic features, and an experimental one where these are supplemented with DRs. Augmenting the inputs with DR features improves objective acoustic scores on a test set and leads to significant preference by listeners in a forced choice AB test for naturalness.
Index Terms: speech synthesis, discourse, prosody

1. Introduction

Although there has been considerable progress in improving the quality of synthetic voices over the past ten years, the great majority of text-to-speech (TTS) systems are still not capable of persuading listeners that they are hearing natural speech. Natural-sounding output is important for many applications of TTS, but producing such output is a very complex task since naturalness depends on many aspects that are hard to predict, especially while relying only on the context of a single sentence. Yet most TTS systems still do not consider linguistic context beyond the current sentence. This is particularly limiting for the synthesis of multi-sentence texts, such as audiobooks. The long-range structures in such texts should indeed be reflected in the acoustic properties of the speech in order to sound natural. Taking a wider context of text into account is necessary to enable a better understanding of the semantic context of the utterance. This additional information would facilitate better prediction of sentence prosody.

In this work we focus on one type of feature that provides additional within-sentence information by taking the wider context into account, derived from discourse relations (DRs). According to computational theories of discourse, DRs express the logical structures that connect different parts of a text. DR theory is one of the most successful and widely used theories in computational discourse, with applications in summarization and sentiment analysis among others. Furthermore multiple studies have provided evidence that DRs have dedicated *acoustic* encodings. Some have focused only on the prosody of DRs [1, 2, 3, 4, 5, for example] while others have succeeded in automatically recognizing DRs with some accuracy [6, 7]. One study, in particular, showed that DRs can be useful to improve HMM-based speech synthesis in Mandarin [8]. Previous work thus suggests that DRs could help increase the naturalness of synthesized speech.

In this work we investigate whether adding additional features derived from DRs to the input of a neural network (NN)-based SPSS system improves the naturalness of a synthetic English voice.[1] As there are currently no available corpora mapping DRs to their acoustic realizations, we create a dataset by automatically annotating four audiobooks, read by a native English speaker, with a state-of-the-art discourse parser. Before attempting to use DRs in speech synthesis, we first verify that they have dedicated prosodic encodings in our dataset. This is done by comparing the acoustic correlates of spoken text within each of the five most frequent DRs to comparable spoken text not within a DR. We find the DRs to be prosodically encoded in both F0 and intensity (Section 5). Motivated by these findings, we build synthetic voices using DR information (Section 6) and evaluate them against a baseline voice (Section 7). Our results show that the DR-derived features improve the naturalness of synthetic speech both according to objective acoustic measures (of F0, intensity, duration) and to human judgements.

2. Related work

Discourse Relations (DRs) express how different segments (i.e. elementary discourse units (EDUs)) of a text are logically connected [11]. Although over the years there have been various versions of DR theory, differing for example in the list of DR categories. The two most widely used annotation frameworks are Rhetorical Structure Theory (RST) [12] and that of the Penn Discourse Treebank (PDTB) [13]. Given their popularity in Natural Language Processing applications, over the years the acoustic correlates of DRs have motivated several studies. Some focus on DR prosody [1, 2, 3, 4, 5]; others experimented with automatic classification of DRs [6, 14, 7] or only discourse connectives in speech [15]. In general, prior work supports the hypothesis that DRs do indeed have acoustic correlates.

Encoding discourse structure in TTS systems is still a relatively unexplored field. Recent work has focused on generic paragraph-based features [16, 17]. In this work we propose an approach to encode DR information in neural statistical parametric speech synthesis (SPSS). To the best of our knowledge, the only directly comparable prior work is [8]. Our work differs from in the following aspects, among others: (i) they used Mandarin Chinese while we investigate English (the two being quite different in their prosody since the first is a tone language and the second is not); (ii) they used HMM-based TTS while we explore how to encode discourse features in neural SPSS;

[1]This project was developed across two MSc dissertations ([9], [10]), to which we refer you for further details.

http://dx.doi.org/10.21437/Interspeech.2019-1945

Table 1: *Distribution of the most frequent extracted DRs*

Type of DR	Number of instances	Proportion among DRs (%)	Proportion of utterances containing at least one instance of this DR (%)
elaboration	6,379	43.62	32.53
joint	3,757	25.69	19.24
attribution	2,018	13.80	10.04
background	708	4.84	3.59
contrast	534	3.65	2.81

Table 2: *Definition of the selected DRs with example sentences*

Type of DR	Definition
elaboration	S gives additional information about N *[I went to the shop $_N$][that is next to my house $_S$]*
joint	Multinuclear relation of paired Ns *[I sang $_{N1}$][and I danced $_{N2}$]*
attribution	Statement in N is reported by S *[I thought $_S$][I could do it $_N$]*
background	S gives essential information to understand N *[He ate $_N$][because he was hungry $_S$]*
contrast	Multinuclear relation where Ns are in opposition *[It seems easy, $_{N1}$][but it's not $_{N2}$]*

(iii) they experimented with augmenting TTS only with pause and average duration features while we experiment with other acoustic features; (iv) they relied on manually-annotated DRs while we use a discourse parser.

3. Methodology

We test two hypotheses:

HI: DRs are prosodically encoded.

HII: Using DRs improves the naturalness of neural SPPS.

We validate HI before testing HII: acoustic encoding of DRs must be detectable if an SPSS system is expected to predict it.

3.1. Do discourse relations have acoustic correlates?

If DRs are acoustically encoded, a speaker will produce variation in their prosody in order to convey discourse information, compared to utterance segments where there is no DR. We hypothesize that there will be a significant difference between DR segments and non- DR segments for certain acoustic features. We also suppose that each DR will be encoded differently. Although others [1, 2, 3, 4, 5] have already found that DRs have acoustic correlates, we must verify this on our data.

3.2. Can discourse relations improve TTS?

If DRs are acoustically encoded, discourse information would be an interesting addition to increase the naturalness of an SPSS system. SPSS predicts acoustic parameters from linguistic features, and DRs would add new information to the existing (within-sentence) linguistic feature set.

4. Dataset creation

DRs are linguistic features that hold over spans of words, so they are much less frequent than smaller units such as syllables. We therefore require a corpus large enough to have multiple occurrences of each type of DR. We chose the corpus from the Blizzard Challenge 2012 [18], which includes four audiobooks, read by the same American English male speaker and freely available on LibriVox.org. This dataset contains 27,320 utterances, paired with automatically generated word- and phoneme-labelled alignments and confidence scores indicating how well the labels are likely to match the book sentences. With more than 50 hours of speech, we considered this corpus large enough; [8]'s system yielded improvements using 10 hours of training data. Moreover, the narrative nature of the audiobooks ensures expressive speech with long-range coherence. This means that DR prosody will reflect extra-sentential context, allowing access to the complete discourse structure. Finally, using this corpus makes comparison with the results in [10] possible.

The only drawback of this corpus is that it is not annotated with DRs. We automatically annotated it using a discourse parser. Contrary to [10] who used a PDTB-based parser [19], we selected the RST-based FastNLPParser [20], built on the Stanford CoreNLP toolkit [21]. This parser was used by [7] who found that DRs could be automatically recognized with good accuracy. Moreover, as [8] used RST for DR-augmented TTS and observed a slight preference toward their DR-enriched voice, it confirmed us that RST could be a good framework. FastNLPParser is one of the best RST parsers currently available, with micro-averaged F1 scores of 65.3 for satellite identification, 54.2 for nucleus, 45.1 for relation labelling and 44.2 for full DR identification on the standard Parseval procedure [22]. It is particularly easy to use as it does not require any pre-formatting of the input text.

The text of the books was processed by FastNLPParser in paragraph-sized chunks, which allows the tool to extract DRs across sentence boundaries (as in [10], [8] and [7]). From the discourse structure obtained in this way, we only kept DRs for which the two EDUs were adjacent leaves of the discourse tree, as in [10] and [7]). This was done in order to prevent any interference from nested DRs or extrinsic DRs separating the two EDUs. Since the parsing was done automatically and will therefore contain some error, focusing on adjacent leaves also helps with the reliability of the results, limiting the propagation of mistakes to higher levels of the discourse tree.

We then discarded sentences for which the automatic alignment's confidence score was less than 100%. We also decided to focus on the five most frequent DR types in order to have enough examples to train our SPSS system, shown in Table 1. The five DR types used are explained in Table 2. Our parsed corpus thus contained 19,349 utterances, and more than 31 hours of speech. Some of the utterances did not take part in any DR whereas other utterances contained one or more DR of the following types: attribution (ATT), background (BAC), contrast (CON), elaboration (ELA) and joint (JOIN).

The proportion of DRs that span adjacent sentences varied depending on the type of DR : 9.65% of the ELA relations, 3.00% of the JOI relations, 1.50% of the CON relations, 1.13% of the BAC relations (and roughly 0% of the ATTR relations) were split across two sentences.

5. Acoustic analysis

To test hypothesis HI – that DRs have acoustic correlates – we compared utterance segments labelled with a DR with segments with no assigned DR (NDR). DR segements did not include any information about EDUs; they only indicated time boundaries and their type of DR. We used Praat [23] to extract the follow-

Table 3: *Statistical significance of DRs to predict acoustic features. p-values are reported as '***' for $p < 0.001$, '**' for $p < 0.01$, '*' for $p < .05$ and '.' for $p < 0.1$.*

Relation	F0			INTENSITY		
	Mean	Range	SD	Mean	Range	SD
ATT		**		***	**	**
BAC	*	*		***	**	**
CON		***		***	**	**
ELA	.	***		***	**	**
JOI	***	***	**	***	**	**

ing features for each segment: (1) duration of the segment, (2) minimum F0, (3) maximum F0, (4) average F0, (5) standard deviation (SD) of the F0, (6) minimum intensity, (7) maximum intensity, (8) average intensity, and (9) SD of the intensity. Each group of DRs was compared with a group of NDR segments, for a total of five DR/NDR pairs. Both groups of each pair had the same number of segments. In order to control the influence of duration on the acoustic analysis, we generated random groups of NDRs and picked the group that had the duration distribution that was the most similar to the one of the group of DRs it was to be compared with.

The result for each combination of a DR and an acoustic feature is according to the following hypotheses:
H0 : The DR has no effect on this acoustic feature, compared to the same acoustic feature in an NDR context.
H1 : The DR has an effect on this acoustic feature, compared to the same acoustic feature in an NDR context.

Each pair's means were compared with Welch's t-test. Resulting *p*-values are reported in Table 3. As we can see, intensity in general and F0 range are significantly predicted by DRs. These results are in line with previously mentioned studies which identified an acoustic encoding of DRs.

Hypothesis HI is thus supported, which allows us to proceed to test our second hypothesis, by integrating DR features into an SPSS system in an attempt to improve the naturalness of its synthetic speech.

6. Discourse-augmented TTS

We tested HII by building two neural SPSS voices which could then be compared: a baseline voice (BASE) without DR features and an experimental voice (wDRS) to which DR features were added. DR features aside, the two voices used the same parameters and were built identically.

Linguistic features for the baseline voice were obtained from the text transcription using Festival [24] (e.g. phone identity, phone's neighbours, part-of-speech of the word), which were then force-aligned with the audio. Identical acoustic features were used for both BASE and wDRS; these consisted of mel-cepstral coefficients, band aperiodicities, and F0 on a logarithmic scale, extracted at 5 ms intervals with a modified version of the open-source vocoder WORLD [25] (DRC edition [26]). Duration models and acoustic models to map from linguistic to acoustic features were trained using Merlin, an open-source toolkit for building neural SPSS systems [27]. The training, validation and testing sets used for both voices were identical. These were created so that each DR type (ATT, BAC, CON, ELA, JOI and NDR) is present in the same proportions: 80% of the occurrences of each DR type for the training set, 10% of them for the validation set, and 10% for the testing set. Considering the fact that some utterances contained multiple occurrences of one or several DR, the resulting training set contained 14,893 utterances, the validation one 1,818 and the testing one 1,802. The extracted linguistic features were transformed into vectors of 416 dimensions for BASE, with either binary or continuous numerical features [27]. The toolkit's default hyperparameters were used, except the batch size which was changed to 32 due to memory constraints. Each model was a feed-forward deep NN (DNN) of 6 hidden layers of 1024 tanh units each; training was done with plain stochastic gradient descent with an initial learning rate of 0.02 which was decayed over 25 epochs of training. This DNN architecture and training regime was chosen as it had been used in previous research on naturalness and had yielded good results (e.g.[28]).

The only difference between BASE and wDRS was that DR-type was added to each frame of linguistic features in the case of wDRS. Thus the linguistic feature vectors used by wDRs were 422 instead of 416-dimensional, with one 1-hot encoding of DR type (including NDR).

7. Evaluation

Once the models were trained, audio was generated for a test set, which was then evaluated objectively and subjectively.

7.1. Objective evaluation

For the objective evaluation, we first retrieved various measurements from the validation and test sets to evaluate how close to natural speech the synthesized utterances were. For duration, we computed root-mean-square error (RMSE) and correlation. For the acoustic model, we computed the mel-cepstral distortion (MCD), the distortion of band aperiodicities (BAP), the F0 RMSE, F0 correlation and voiced/unvoiced error (V/UV). We also performed the same acoustic analysis that was previously done to test HI on natural speech (section 5) but this time on the synthetic speech generated by the two voices. This tested whether either of the two voices generate significantly different prosody for DRs compared to NDRs.

7.2. Subjective evaluation

For the subjective evaluation we performed a listening test with 30 English native speakers with no hearing impairment. All participants took the test with Beyerdynamic DT770 headphones in an soundproof booth and were paid for taking part. The listening test comprised 60 forced-choice preference questions. The questions were of two different types: (1) choose the more natural-sounding of two renditions of the same text (2) choose the more natural-sounding of two renditions of the same text, when preceded by a segment of natural speech to serve as contextual prosodic cue.

There were 30 questions of each type: 5 for each of the 6 DR types (including NDR). We chose pairs of utterances that sounded different enough and that only included a single occurrence of a DR (or none for NDRs) so we could isolate each DR. We based our selection on the comparision of the measurements used for the acoustic analysis and on a subjective appreciation of the difference. The questions were put in a random order so the listener would not be bored and would remain attentive, and also to remove any ordering effect. 5 additional questions with a comparison between natural and synthesized speech were included to ensure that participants conducted the task properly and detect cheating. The natural speech utterances were vocoded with WORLD [25] to give them the same signal quality as the synthetic speech.

Table 4: *Comparison of objective duration and acoustic measurements*

	Validation set		Test set	
	BASE	wDRs	BASE	wDRs
Dur RMSE	9.027	9.029	9.281	9.265
Dur corr	0.701	0.701	0.690	0.691
MCD (dB)	5.823	5.713	5.855	5.744
BAP (dB)	0.283	0.279	0.284	0.281
F0 RMSE (Hz)	34.172	33.824	34.714	32.204
F0 corr	0.404	0.435	0.395	0.437
V/UV (%)	7.006	6.958	6.993	6.931

8. Results

The results of the objective evaluation are reported in Table 4. From the acoustic measurements, we can see in Table 4 that, although the results of both voices are quite similar, wDRs usually performs better than BASE. This tendency can be observed on both validation and test set. The results from our second objective evaluation were less clear. Neither of the two voices seem to perform better and to replicate exactly what we had observed during our analysis of real speech. However, the addition of DR features did create variation in the way utterances were synthesized, since the significant acoustic parameters that can be predicted by the type of DR varied depending on the voice.

From the subjective evaluation, the results of both preference tests are presented in Figure 1. Overall, wDRs was significantly preferred over BASE (two proportion z-test, $p < 0.001$). This preference can also be clearly observed for all relation types, except for ATT (two proportion z-test, $p < 0.001$ for all relation types, except ATT ($p = 0.729$)). The addition of a contextual natural speech segment had no significant impact on the preference towards wDRs, except for ELA, where it became non-significant. This can probably be explained by the fact that the ELA utterances (synthesized + natural speech segments) were generally longer, which might have prevented hearing the differences distinctively.

9. Discussion

Our experiments indicate that DRs are acoustically-encoded and that the addition of DRs features creates variation in synthesized utterances. However, our results do not clearly indicate what impact DRs have on the acoustic realization of an utterance. We might question whether we focused on the most relevant prosodic features or if our measurements were influenced by other variables such as duration or position of the sentence in the paragraph as mentioned in [29].

Nevertheless, the listening test clearly indicated that wDRs performed better than BASE. Our results show that the addition of DR features to a TTS system did indeed significantly improve the naturalness of synthesized speech. The objective measurements that compared output from our two voices to natural speech showed that wDRs was better than BASE. Moreover, the listening test also indicated a preference for wDRs over BASE from our participants with regards to the naturalness of both voices. We can therefore affirm that DRs are features that can help make synthesized speech sound more natural.

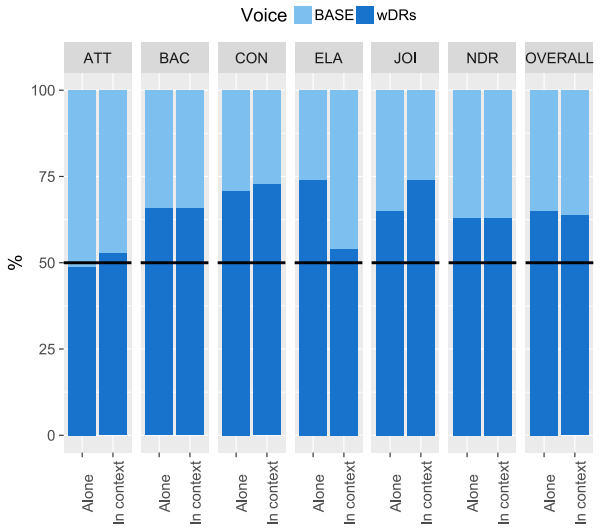

Figure 1: *Comparison of the listening test results depending on the voice, the relation type and the absence or presence of contextual natural speech.*

10. Conclusions

These experiments attempted to improve the naturalness of SPSS with the addition of DR features. Our results showed that DR features are indeed a useful addition to a TTS system as they help create a significantly more natural-sounding voice. We have thus succeeded in combining the DR framework of [8] with a neural SPSS system [10] to produce conclusive results.

However, it must be acknowledged that the automatic parsing of our corpus is certainly not 100% accurate. A large, manually-annotated speech database would be better. It could also be interesting to further investigate which other acoustic parameters are varied by speakers in their acoustic realization of DRs. More fine-grained DR features (e.g. subunits of DRs) may help, given enough data. Our neural TTS system had a very simple architecture which could easily be improved in future.

In conclusion, we have shown that TTS would benefit from further research regarding DRs and other long-span textual features such as co-reference. Current TTS system are missing a general understanding of the relations between utterances.

11. References

[1] J. Hirschberg and D. Litman, "Now let's talk about now: Identifying cue phrases intonationally," in *Proceedings of the 25th annual meeting on Association for Computational Linguistics*. ACL, 1987, pp. 163–171.

[2] J. B. Hirschberg, D. J. Litman, J. B. Pierrehumbert, and G. Ward, "Intonation and the intentional structure of discourse," in *Proceedings of the 10th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI, 1987, pp. 636–639.

[3] R. Herman, "Intonation and discourse structure in english: Phonological and phonetic markers of local and global discourse structure," Ph.D. dissertation, The Ohio State University, 1998.

[4] H. den Ouden, "The prosodic realization of text structure," Ph.D. dissertation, University of Utrecht, 2004.

[5] J. Tylor, "Prosodic correlates of discourse boundaries and hierarchy in discourse production," *Lingua*, vol. 133, pp. 101–126, 2013.

[6] G. Murray, M. Taboada, and S. Renals, "Prosodic correlates of rhetorical relations," in *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*. ACL, 2006, pp. 1–7.

[7] J. Kleinhans, M. Farrús, A. Gravano, J. M. Pérez, C. Lai, and L. Wanner, "Using prosody to classify discourse relations," in *Interspeech 2017*. IEEE, 2017, pp. 3201–3205.

[8] N. Hu, P. Shao, Y. Zu, Z. Wang, W. Huang, and S. Wang, "Discourse prosody and its application to speech synthesis," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[9] A. Aubin, "Improving the naturalness of a statistical parametric speech synthesis system with discourse relations," Master's thesis, University of Edinburgh, 2018.

[10] A. Cervone, "Discourse parsing for statistical parametric TTS," Master's thesis, University of Edinburgh, 2015.

[11] Y. Versley and A. Gastel, "Linguistic tests for discourse relations in the tba-d/z corpus of written german," *Dialogue and Discourse*, vol. 4, no. 2, pp. 142–173, 2013.

[12] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organisation," *Text - Interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 343–381, 1988.

[13] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber, "The penn discourse treebank 2.0." in *LREC*. Citeseer, 2008.

[14] A. Cervone, C. Lai, S. Pareti, and P. Bell, "Towards automatic detection of reported speech in dialogue using prosodic cues," in *Interspeech 2015*. IEEE, 2015, pp. 3061–3065.

[15] G. Riccardi, E. A. Stepanov, and S. A. Chowdhury, "Discourse connective detection in spoken conversations," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6095–6099.

[16] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[17] À. Peiró-Lilja and M. Farrús, "Paragraph prosodic patterns to enhance text-to-speech naturalness," in *Klessa K, Bachan J, Wagner A, Karpiński M, Śledziński D. Proceedings of the 9th International Conference on Speech Prosody; 2018 June 13-16; Poznań, Poland.[Lous Tourils]: ISCA; 2018. p. 512-6.* International Speech Communication Association (ISCA), 2018.

[18] S. King and V. Karasaikos, "The blizzard challenge 2012," in *Proceedings of the Blizzard Challenge 2012*. IEEE, 2012.

[19] Z. Ling, H. T. Ng, and M.-Y. Kan, "A pdtb-styled end-to-end discourse parser," *Natural Language Engineering*, vol. 20, no. 2, pp. 151–184, 2014.

[20] M. Surdeanu, T. Hicks, and M. A. Valenzuela-Escárcega, "Two practical rhetorical structure theory parsers," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: Software Demonstrations (NAACL HLT)*. ACL, 2015, pp. 1–5.

[21] C. D. Manning, M. Surdeanu, J. Bauer, John ad Finkel, S. J. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*. ACL, 2014, pp. 55–60.

[22] M. Morey, P. Muller, and N. Asher, "How much progress have we made on rst discourse parsing? a replication study of recent results on the rst-dt," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. ACL, 2017, pp. 1319–1324.

[23] P. Boersma, "Praat: doing phonetics by computer [Computer program]," *http://www. praat. org/*.

[24] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[25] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99-D, no. 7, pp. 343–381, 2016.

[26] M. Morise, "DRC, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, no. 85, pp. 343–381, 2016.

[27] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proceedings 9th ISCA Speech Synthesis Workship (SSW9)*. IEEE, 2016, pp. 218–223.

[28] S. Ronanki, S. Reddy, B. Bollepalli, and S. King, "Dnn-based speech synthesis for indian languages from ascii text," in *Proceedings of the 9th ISCA Workshop on Speech Synthesis*. IEEE, 2016, pp. 74–79.

[29] M. Farrús, C. Lai, and M. J. D., "Paragraph-based prosodic cues for speech synthesis applications," in *Proceedings of the 8th International Conference on Speech Prosody (SP 2016)*. IEEE, 2016, pp. 1143–1147.