IDIAP RESEARCH REPORT

PROBABILISTIC AMPLITUDE DEMODULATION FEATURES IN SPEECH SYNTHESIS FOR IMPROVING PROSODY

Alexandros Lazaridis[a] Milos Cernak
Philip N. Garner

Idiap-RR-12-2016

APRIL 2016

[a]Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Probabilistic Amplitude Demodulation features in Speech Synthesis for Improving Prosody

Alexandros Lazaridis, Milos Cernak and Philip N. Garner
Idiap Research Institute, Martigny, Switzerland
{alaza,milos.cernak,phil.garner}@idiap.ch

April 4, 2016

Amplitude demodulation (AM) is a signal decomposition technique by which a signal can be decomposed to a product of two signals, i.e, a quickly varying carrier and a slowly varying modulator. In this work, the probabilistic amplitude demodulation (PAD) features are used to improve prosody in speech synthesis. The PAD is applied iteratively for generating syllable and stress amplitude modulations in a cascade manner. The PAD features are used as a secondary input scheme along with the standard text-based input features in statistical parametric speech synthesis. Specifically, deep neural network (DNN)-based speech synthesis is used to evaluate the importance of these features. Objective evaluation has shown that the proposed system using the PAD features has improved mainly prosody modelling; it outperforms the baseline system by approximately 5% in terms of relative reduction in root mean square error (RMSE) of the fundamental frequency (F0). The significance of this improvement is validated by subjective evaluation of the overall speech quality, achieving 38.6% over 19.5% preference score in respect to the baseline system, in an ABX test.

Index Terms: Probabilistic amplitude demodulation, speech synthesis, deep neural networks, speech prosody

1 Introduction

In human-to-human communication, through speech, the speaker conveys information on different levels i.e., linguistic (e.g. phonetic and linguistic information), paralinguistic (e.g. speaking style or emotions of the speaker) and extralinguistic levels (e.g. socio-geographical background of the speaker). Prosody is related to all of these levels and varies depending on the message that is desired to be conveyed to the listener [10]. In acoustic terms, prosody is mainly composed by three aspects, i.e., the fundamental frequency, duration of phonetic units and intensity [3, 4]. Since the properties of prosodic features are units of speech larger than segments, prosody is related not only to segmental level information, but also to the suprasegmental one. Consequently, the correlation of segmental and suprasegmental information levels becomes very important in prosody modelling. Robust modelling of prosody is essential since very often changing prosody could even change the underlying meaning of the message [13]. This makes it very important not only for text-to-speech (TTS) synthesis systems and related applications but also for broader applications such as speech-to-speech translation (S2ST), where prosody becomes one of the essential information that needs to be analysed (in the source language), transferred to the target language and synthesized.

A speech signal conveys information on different time-scales. Traditionally, sequential speech processing suggests the segmental and suprasegmental time-scales be used for different models of our interest, such as for the acoustic and prosodic modelling. Different time-scales have often been treated

independently in the past. However, we can hypothesise that they are related, and that this relation is important also for prosody modelling.

Over the last decades, an increasing interest can be observed in the literature, concerning the spectro-temporal structure of the speech signal and its correlation to the phonological structure of language and speech perception [2, 8, 16]. In research related to children with impaired phonological development, in several languages [21, 6, 5], reduced sensitivity to the amplitude demodulation structure of acoustic signals was observed across languages, relating the extraction of information about phonological structure, to the energy patterns of the amplitude envelope. Nonetheless, it remains unclear which modulations (time-scales), are the most important relating acoustic information with phonological. Investigating this issue, Leong and Goswami [15], studied how acoustic spectro-temporal structure is related to the linguistic phonological structure of speech, using amplitude demodulation in three time-scales, i.e, prosodic stress, syllable and onset-rime unit (phonemes) levels.

In this work an attempt is made to provide us an insight into the prosody hierarchy. For this reason we have selected the probabilistic amplitude demodulation (PAD) approach [23]. The PAD method is noise robust and allows the algorithm to be steered using a-priori knowledge of modulation time-scales, i.e., the user can specify the prosodic tiers — stress, syllables, and utterance — to be analysed. And as an analytic model, it is assumed to be language independent. The PAD method can be used iteratively to get progressively slower prosodic tiers.

A novel speech synthesis with enhanced prosody (SSEP) system is presented in this paper. An attempt is made to investigate the importance of PAD features used as additional input feature scheme in DNN-based speech synthesis. Two level amplitude demodulation is performed in this work. A first demodulation is performed with a syllable-based modulation where an average syllable duration in samples is used as parameter. The resulting syllable envelope is used as input signal for progressively slower demodulation at the stress level, to generate a stress envelope. The motivation behind this attempt is to manage to capture the relation between segmental and suprasegmental levels, using the PAD technique. We hypothesize that the PAD features are able to capture this correlation and are going to be beneficial in speech synthesis.

The remainder of the paper is organized as follows. In Section 2, the proposed SSEP scheme is presented. The experimental protocol is described in Section 3. In Section 4, the objective and subjective evaluation results are presented. Finally the conclusions are given in Section 5.

2 Speech synthesis with enhanced prosody

In this section, the probabilistic amplitude demodulation (PAD) scheme along with the DNN-based speech synthesis framework are described. The combination of these two schemes lead to the proposed speech synthesis with enhanced prosody (SSEP) scheme.

2.1 Probabilistic amplitude demodulation

The probabilistic amplitude demodulation (PAD) models the speech signal y_t as:

$$y_t = c_t.m_t \tag{1}$$

where c_t and m_t are a carrier and modulator components, respectively. The modulator is represented as a non-linear function

$$m_t = m(x_t) = \sigma_m \log(1 + \exp(x_t)) \tag{2}$$

of the transformed-modulator signal x_t, with the amplitude σ_m, drawn from a stationary Gaussian process. The covariance function of x_t represents the typical time-scale of variations of m_t, and importantly, it can be controlled manually using a-priori user knowledge. The carrier is modelled as a Gaussian process which is uncorrelated in time.

There are many solutions for solving Eq. (1). The PAD method describes a Bayesian inference given the data for extracting the amplitude modulation structure. More specifically, posterior probability of

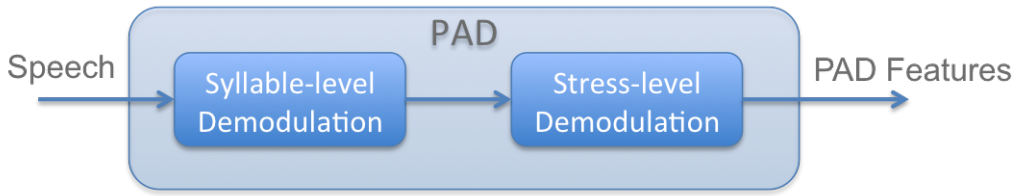

Figure 1: *Two-level Probabilistic amplitude demodulation scheme; syllable-level and stress-level demodulations.*

all the possible modulators and carriers given the data is:

$$p(c_1^T, m_1^T | y_1^T, \theta) = \frac{p(y_1^T, c_1^T, m_1^T | \theta)}{p(y_1^T | \theta)}, \tag{3}$$

where $p(y_1^T, c_1^T, m_1^T | \theta)$ is the joint probability of the signal, carrier and modulator, T is the number of frames of the processed speech signal, and θ corresponds to the model parameters. The most probable modulator and carrier are obtained by the *maximum a posteriory* (MAP) inference as:

$$\hat{c}_1^T, \hat{m}_1^T = \underset{c_1^T, m_1^T}{\operatorname{argmax}} \; p(c_1^T, m_1^T | y_1^T, \theta), \tag{4}$$

using a gradient-based method that is used to search for the optimal solution. To allow the demodulation to be user steerable, i.e., perform the demodulation using a specific time-scale, the parameters of the model θ can be obtained by the MAP inference as:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; p(\theta | y_1^T) = \underset{\theta}{\operatorname{argmax}} \; p(y_1^T | \theta) p(\theta), \tag{5}$$

where the prior over parameters $p(\theta)$ is set by the user. The maximum-likelihood estimate is recovered when the prior is uniform, i.e., $p(\theta) \propto c$.

To reveal different time-scale information present in the speech signal, we used the PAD process to decompose the signal into a cascade of modulators and a carrier [22]. The time-scales of the modulators are considered as the prior constants $p(\theta)$, creating a concept of steered demodulation. A first demodulation is performed with a syllable-based modulation where an average syllable duration in samples is used as the parameter prior $p_{syll}(\theta)$. The obtained syllable envelope σ_{syll} is used as input signal for progressively slower demodulation at the stress level, using a different prior $p_{stress}(\theta)$, to generate a stress envelope σ_{stress} (see Figure 1). The general purpose values for the speech signal demodulation could be 5Hz for the first decomposition with the syllable frequency, and an average between the half and one third of the syllable frequency for the stress modulation frequency. For example, considering 16kHz sampled data, the values could be $p_{syll}(\theta) = 3200$ samples and $p_{stress}(\theta) = 8000$ samples. The better prior estimate of the syllabic rate, the more accurate the obtained cascaded demodulation.

In addition, the PAD method is able to deal with noisy data, as it explicitly incorporates additive uncorrelated Gaussian noise around the product of $c_t.m_t$.

In Figure 2 the PAD syllable- and stress-level modulations are shown for the utterance *it's generally a frog or a worm*.

2.2 Speech synthesis with enhanced prosody scheme

In this subsection, initially, the DNN-based speech synthesis framework, following the framework of [25, 14], and constitutes the baseline system in our experiments, see Section 3, is described and consequently the proposed SSEP scheme is presented.

A DNN is a feed-forward artificial neural network with multiple hidden layers between the input and output layer, creating a mapping function between the input (i.e. linguistic features) vector and the output (i.e. acoustic features) vector. In the training phase, the input text is processed and

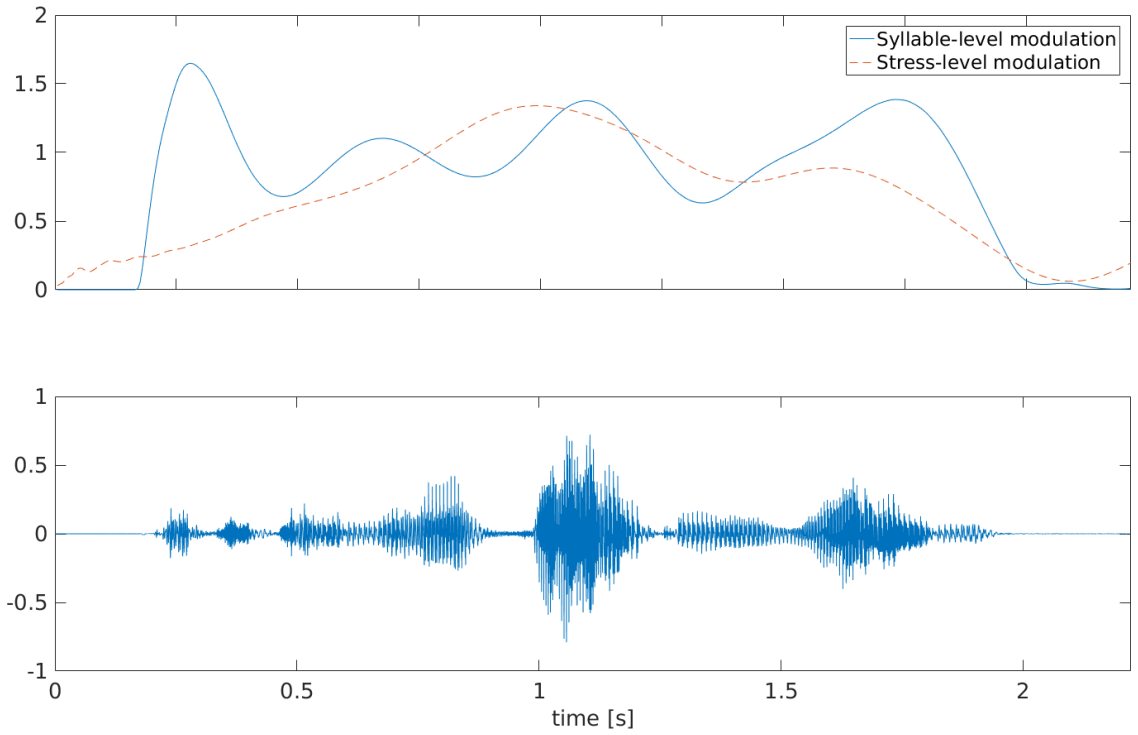

Figure 2: *The PAD syllable- and stress-level modulations of the utterance "it's generally a frog or a worm".*

transformed into labels, which contain linguistic features in an appropriate format for training the DNNs, i.e., containing binary and numerical features. Back-propagation is used for training the DNN using the input and output data.

The text corresponding to each audio file has to be converted into a sequence of labels suitable for DNN training. A conventional and freely available TTS front-end was used for this [1]. The text is turned into a sequence of labels (text-based labels), which contain segmental information and rich contextual parameters such as lexical stress and relative position within syllables, phrases or sentences. The standard "*full*" labels generated by the scripts, i.e. quinphone segmental information, and a large number of categorical, numeric, or binary linguistic and prosodic information, was used [26]. These labels were aligned with the speech signal through a phone-based forced alignment procedure, using the Kaldi toolkit [17]. The models for the alignment were trained on the training plus development sets, and state-level labels force-aligned to acoustic frame boundaries were generated for the training, development and evaluation sets.

Concerning the output features, the STRAIGHT [26] vocoder was used for the acoustic analysis and feature extraction, essentially using the default settings from the EMIME [24] scripts: 25ms frame window, 5ms frame shift, STRAIGHT Mel-cepstral analysis with 40 coefficients, single F0 value, and 21 coefficients for band aperiodic energy, extracted by the STRAIGHT vocoder. For each acoustic feature, derivatives of first and second order are added. The overall acoustic vector dimension is 186.

A slightly modified version of the Kaldi toolkit for the DNN training was used. An automatic procedure was used to convert the labels into numeric values: the categorical data (such as segmental information) was turned into arrays of binary values, while the numerical and binary data was preserved.

Since training requires a frame-level mapping between input labels and acoustic features, the segment-based labels have to be sampled so that we have an input label per acoustic frame. The DNN system was trained using the state position within the phone as categorical data, plus using two position features, i.e. numeric values corresponding to the frame position within the current state, and to the frame position within the current segment, plus the standard "*full*" labels (i.e. a total of 403 input features). Furthermore it should be noted that the input (label) data was normalized globally so that each component had values between 0.01 and 0.99. The output (acoustic) data was further normalized for each component to be of zero mean and unit variance; the output activation function

was a sigmoid.

Unlike other approaches (such as Zen [25] or Qian [18]), we did not remove silent frames from the training. The training procedure was standard: we used a stochastic gradient descent based on back propagation. The minimisation criterion was the Mean Square Error (MSE). The training was run on the *training* set, and we used the *development* set for cross-validation.

In the synthesis phase, the input text is processed by the same front-end as in the training phase, creating the input vectors and the trained DNN is used in a forward-propagation manner for mapping them to output vectors. The aligned label files from the evaluation set were used for synthesis. Synthesis was performed doing a forward pass through the network, followed by acoustic trajectory smoothing [9], through applying the "mlpg" tool from SPTK [11] and global variance computed on each acoustic component. This was followed by resynthesis using the STRAIGHT vocoder.

In Figure 3, the proposed SSEP scheme is shown. As can be seen, the DNN-based speech synthesis (baseline) system is combined with the PAD scheme.

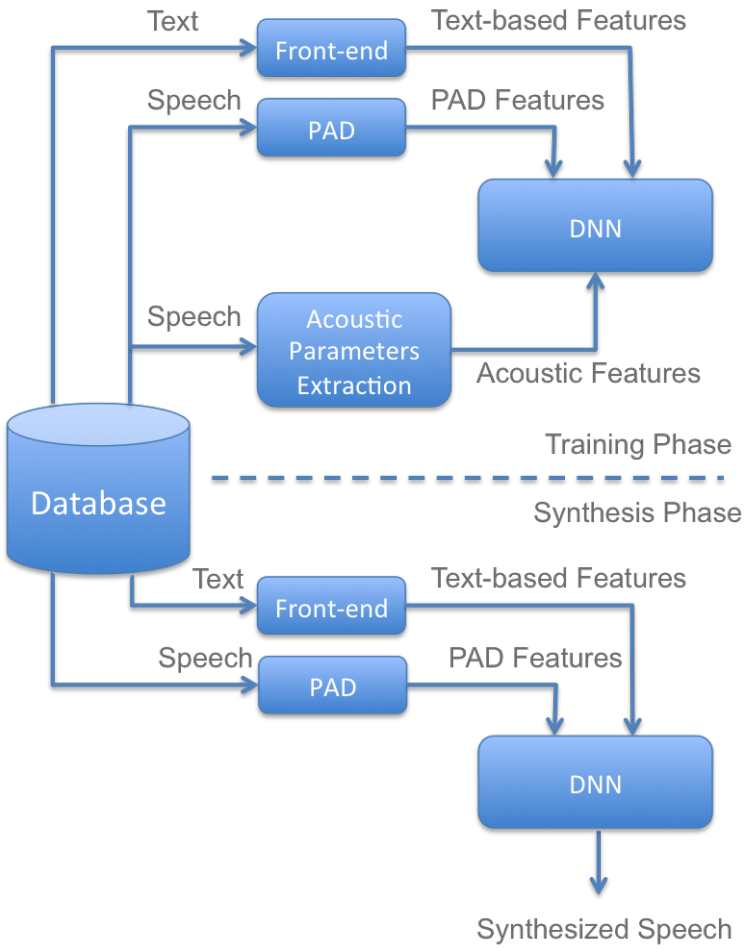

Figure 3: *Speech synthesis with enhanced prosody scheme.*

During the training phase, in parallel with the baseline scheme, the PAD scheme is used to extract the PAD features. These features are combined, on frame-level, with the text-based features and used as the input features for the DNN. The output features remain the same as in the baseline system described above. During the synthesis phase, both the text-based and the PAD features are extracted in the same way as in the training phase.

Since in a real scenario, during the synthesis phase, the speech signal is not available, in order to extract the PAD features, these features need to be predicted from text. Alternatively, this scheme could be used in a S2ST scenario. In this case the PAD features would be extracted from the source speaker in the source language, be transformed/adapted to the target speaker and language and consequently be used in the proposed speech synthesis scheme.

3 Experiments

3.1 Database

For the experiments the blizzard-challenge-2008 [19, 20] database was used. The speaker is known as "Roger" and is a native Uk English male speaker. The database consists of 15 hours of data, comprising around 9.6k utterances. For our experiments a subset of the database was used, composed by the "carroll", "arctic" and the three news sets (i.e., "theherald 1,2,3"). The total number of utterances of this subset was approximately 4.8k corresponding to 7.5 hours of speech. This subset was split in a training set of 4273 utterances, a development set of 335 utterances and an evaluation set of 158 utterances. The sampling frequency of the audio is 16kHz.

3.2 DNN-based speech synthesis setup

The DNNs were built implementing various combinations of the number of hidden layers (i.e. from 4 to 6 hidden layers), and nodes (i.e. 1000 and 2000 nodes) in each layer. Each layer comprised an affine component followed by a sigmoid activation function. Based on the development set, the best performance in respect to mel-cepstral distortion (MCD) [12] and root mean square error (RMSE) of the F0 was achieved by the DNN system composed of 4 hidden layers and 2000 units per layer.

3.3 PAD features setup

For the extraction of the PAD features, a frame window of 25ms and a frame shift of 5ms were used. The default (not calculated based on the specific speaker) syllable frequencies of 5Hz was selected. The two PAD features were combined with the frame-level text-based input features as described in Section 2.2.

4 Results

To validate our hypothesis, that these features will be beneficial, especially in prosody modelling, both objective and subjective evaluation was performed.

4.1 Objective evaluation

The MCD between original and synthesized samples is used as an objective metric to compare the two systems. Higher MCD values indicate lower speech quality of the synthesized speech samples. Additionally for evaluating the two systems in respect to prosody modelling, the RMSE of F0 was calculated for each system. These results can be seen in Table 1.

Table 1: *MCD in dB and RMSE of F0 in Hz for the baseline and proposed systems on the evaluation set.*

System	MCD (dB)	F0 (Hz)
Baseline	3.938	19.096
Proposed	3.912	18.208

As can be seen from the results, the reduction in MCD of the proposed SSEP system over the baseline one is very small, i.e, approximately 0.7% relative improvement. Nonetheless, the reduction of RMSE of F0 of the SSEP system over the baseline one is approximately 4.7%, showing a small but clear relative improvement in respect to prosody modelling. The results are statistically significant ($p < 0.05$).

It should be pointed out that the PAD features were extracted using the "default" syllable frequency of 5Hz and not a specific one based on the speaker. This means that there is potential for improve

the PAD features extraction procedure, by using a syllable rate estimation before the PAD extraction. This is expected to give more accurate PAD features and consequently better performance of the SSEP system. Furthermore, since the database used in these experiments consists of read speech, where prosody variations are constraint due to the strict speaking style, it is expected that the importance of the PAD features, when more expressive or emotional speech (e.g. audiobooks) is used, will be substantially bigger.

4.2 Subjective evaluation

In order to identify whether the improvement in the reduction of the RMSE of F0 would be perceivable by listeners, a subjective evaluation ABX test was performed.

We employed a 3-point scale ABX subjective evaluation listening test [7], suitable for comparing two different systems. In this test, listeners were presented with pairs of samples produced by two systems (A and B) and for each pair they were indicating their preference for A, B, or *both samples sound the same* (X). The material for the test consisted of 15 pairs of sentences such that one member of the pair was generated using the baseline DNN speech synthesis (system A) and the other member was generated using the proposed SSEP system (system B). Random utterances from the evaluation set were used. 27 listeners (native and non-native English) participated in the ABX test. The subjects were presented with pairs of sentences in a random order with no indication of which system they were represented with. They were asked to listen to these pairs of sentences (as many times as they wanted), and choose between them in terms of their overall quality. Additionally, the option X, i.e. *both samples sound the same*, was available if they had no preference for either of them.

As can be seen in Figure 4, the SSEP system clearly outperforms the baseline one, achieving double preference score, i.e., 38.6% over 19.5% respectively. In addition the *both samples sound the same* ("Equal") choice achieved a 41.9%.

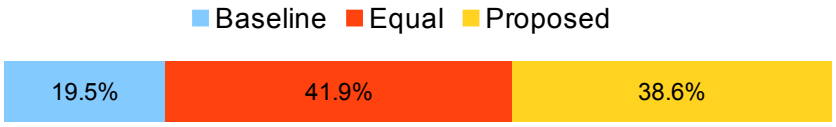

Figure 4: *Subjective evaluation ABX test results (in %) of the baseline and proposed systems.*

Furthermore, it should be pointed out that, according to the feedback from many of the listeners, bigger differences in prosody between the audio pairs was perceived, when the variations in prosody were bigger. This confirms our hypothesis, that the contribution of PAD features, when using more expressive and emotional speech, will be bigger.

5 Conclusions

A novel speech synthesis with enhanced prosody scheme was proposed. The probabilistic amplitude demodulation technique was used as additional, to the standard text-based input features, scheme in deep neural network speech synthesis. Both the objective and subjective evaluation showed improvement in F0 modelling of the speech synthesis with enhanced prosody system compared to the baseline one. The proposed system using PAD features achieved approximately 5% relative reduction in the root mean square error of F0 in respect to the baseline system without using the PAD features. The improvement in F0 modelling was validated by subjective ABX listening test, where the proposed system achieved a preference score of 38.6% over 19.5% of the baseline.

As future work, lower time-scales (i.e. higher frequency) demodulations will be investigated, e.g. phoneme-level. Furthermore, the authors are interested in investigating ways to predict these features from text for evaluating whether these features could be beneficial also in text-to-speech synthesis. Finally, using this technique in speech-to-speech translation, transferring these features from the source speaker (in the source language), to the target speaker (in another language), is another very interesting path which will be investigated.

6 Acknowledgements

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS), and under SP2: the SCOPES Project on Speech Prosody.

References

[1] Alan Black, Paul Taylor, and Richard Caley. The festival speech synthesis system: System documentation (1.3.1). Technical Report HCRC/TR-83, Human Communication Research Centre, December 1998.

[2] R. Drullman, J. M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95(2):1053–1064, February 1994.

[3] Thierry Dutoit. *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, 1997.

[4] S. Furui. *Digital Speech Processing: Synthesis, and Recognition, Second Edition,*. Signal Processing and Communications. Taylor & Francis, 2000.

[5] Usha Goswami. A temporal sampling framework for developmental dyslexia. *Trends in cognitive sciences*, 15(1):3–10, 2011.

[6] Usha Goswami, H.-L. Sharon Wang, Alicia Cruz, Tim Fosker, Natasha Mead, and Martina Huss. Language-universal sensory deficits in developmental dyslexia: English, Spanish, and Chinese. *J. Cognitive Neuroscience*, 23(2):325–337, 2011.

[7] Volodya Grancharov and W. . Bastiaan Kleijn. Speech Quality Assessment. In Jacob Benesty, Sondhi, and YitengArden Huang, editors, *Springer Handbook of Speech Processing*, pages 83–100. Springer Berlin Heidelberg, 2008.

[8] Steven Greenberg, Hannah Carvey, Leah Hitchcock, and Shuangyu Chang. Temporal properties of spontaneous speecha syllable-centric perspective. *Journal of Phonetics*, 31(34):465 – 485, 2003. Temporal Integration in the Perception of Speech.

[9] HTS. HMM-based speech synthesis system version 2.1. 2010.

[10] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. Spoken language processing: A guide to theory, algorithm, and system development. 2001.

[11] Satoshi Imai and Takao Kobayashi. Speech signal processing toolkit (SPTK) version 3.7, 2013.

[12] R. F. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proc. of ICASSP*, volume 1, pages 125–128 vol.1. IEEE, May 1993.

[13] John Laver. *Principles of Phonetics*. Cambridge University Press, Cambridge, 1994.

[14] Alexandros Lazaridis, Blaise Potard, and Philip N. Garner. DNN-based speech synthesis: Importance of input features and training data. In N. Fakotakis A. Ronzhin, R. Potapova, editor, *International Conference on Speech and Computer, SPECOM 2015*, Lecture Notes in Computer Science, pages 193–200. Springer Berlin Heidelberg, 2015.

[15] Victoria Leong and Usha Goswami. Acoustic-emergent phonology in the amplitude envelope of child-directed speech. *PLoS ONE*, 10:1–37, 12 2015.

[16] Victoria Leong, Michael A. Stone, Richard E. Turner, and Usha Goswami. A role for amplitude modulation phase relationships in speech rhythm perception. *J. Acoust. Soc. Am.*, 136(1):366–381, July 2014.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The Kaldi speech recognition toolkit. In *Proc. of ASRU*, 2011.

[18] Yao Qian, Yuchen Fan, Wenping Hu, and F.K. Soong. On the training aspects of deep neural network (DNN) for parametric tts synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3829–3833, 2014.

[19] Volker Strom, Robert Clark, and Simon King. Expressive prosody for unit-selection speech synthesis. In *Proc. Interspeech*, Pittsburgh, 2006.

[20] Volker Strom, Ani Nenkova, Robert Clark, Yolanda Vazquez-Alvarez, Jason Brenier, Simon King, and Dan Jurafsky. Modelling prominence and emphasis improves unit-selection synthesis. In *Proc. Interspeech 2007*, Antwerp, Belgium, August 2007.

[21] Zsuzsanna Surányi, Valéria Csépe, Ulla Richardson, Jennifer M Thomson, Ferenc Honbolygó, and Usha Goswami. Sensitivity to rhythmic parameters in dyslexic children: A comparison of Hungarian and English. *Reading and Writing*, 22(1):41–56, 2009.

[22] R. E. Turner and M. Sahani. Probabilistic amplitude demodulation. In *Proc. of Independent Component Analysis and Signal Separation*, pages 544–551, 2007.

[23] R. E. Turner and M. Sahani. Demodulation as Probabilistic Inference. *IEEE Trans. on Audio, Speech, and Language Processing*, 19(8):2398–2411, November 2011.

[24] Mirjam Wester, John Dines, Matthew Gibson, Hui Liang, Yi-Jian Wu, Lakshmi Saheer, Simon King, Keiichiro Oura, Philip N. Garner, William Byrne, Yong Guan, Teemu Hirsimäki, Reima Karhila, Mikko Kurimo, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, and Junichi Yamagishi. Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *SSW7*, pages 192–197, 2010.

[25] Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using Deep Neural Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7962–7966, 2013.

[26] Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.