Investigating Spectral Amplitude Modulation Phase Hierarchy Features in Speech Synthesis

Alexandros Lazaridis, Milos Cernak, Pierre-Edouard Honnet and Philip N. Garner

Idiap Research Institute, Martigny, Switzerland

{alaza,milos.cernak,pierre-edouard.honnet,phil.garner}@idiap.ch

Abstract

In our recent work, a novel speech synthesis with enhanced prosody (SSEP) system using probabilistic amplitude demodulation (PAD) features was introduced. These features were used to improve prosody in speech synthesis. The PAD was applied iteratively for generating syllable and stress amplitude modulations in a cascade manner. The PAD features were used as a secondary input scheme along with the standard text-based input features in deep neural network (DNN) speech synthesis. Objective and subjective evaluation validated the improvement of the quality of the synthesized speech.

In this paper, a spectral amplitude modulation phase hierarchy (S-AMPH) technique is used in a similar to the PAD speech synthesis scheme, way. Instead of the two modulations used in PAD case, three modulations, i.e., stress-, syllable- and phoneme-level ones (2, 5 and 20 Hz respectively) are implemented with the S-AMPH model. The objective evaluation has shown that the proposed system using the S-AMPH features improved synthetic speech quality in respect to the system using the PAD features; in terms of relative reduction in mel-cepstral distortion (MCD) by approximately 9% and in terms of relative reduction in root mean square error (RMSE) of the fundamental frequency (F0) by approximately 25%. Multi-task training is also investigated in this work, giving no statistically significant improvements.

Index Terms: spectral amplitude modulation phase hierarchy, 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 phonological 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 [1]. In acoustic terms, prosody is mainly composed by three aspects, i.e., the fundamental frequency, duration of phonetic units and intensity [2, 3]. 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 [4]. 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 a part 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 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 [5, 6, 7]. In research related to children with impaired phonological development, in several languages [8, 9, 10], reduced sensitivity to the amplitude demodulation structure of acoustic signals was observed across languages. This led to the conclusion of the existence of correlation between the extraction of information about phonological structure and the energy patterns of the amplitude envelope. Nonetheless, it remains unclear which modulations (time-scales) are the most important relating acoustic with phonological information. Investigating this issue, Leong and Goswami [11] 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 our recent work [12], the probabilistic amplitude demodulation (PAD) approach [13] was used in a novel speech synthesis with enhanced prosody (SSEP) system. An attempt was made to investigate the importance of PAD features used as additional input feature scheme in DNN-based speech synthesis. 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. Furthermore, as an analytic model, it is assumed to be language independent. The PAD method can be used iteratively to get progressively slower prosodic tiers. In our case, two level amplitude demodulation was performed. A first demodulation was performed with a syllable-level modulation where an average syllable duration in samples was used as parameter. The resulting syllable envelope was used as input signal for progressively slower demodulation at the stress level, to generate a stress envelope. Our hypothesis, that the PAD features would be able to capture this correlation and would be beneficial in speech synthesis, was validated [12].

In this work the PAD scheme is replaced by the spectral amplitude modulation phase hierarchy (S-AMPH) [11] approach for improving speech synthesis. An attempt is made to inves-

tigate the importance of S-AMPH features used as additional input feature scheme in DNN-based speech synthesis. Three level amplitude demodulation is performed in this work. The stress-level demodulation, to generate a stress envelope (2 Hz amplitude modulation). The syllable-level modulation where an average syllable duration in samples is used as parameter (5 Hz amplitude modulation). Finally, phoneme-level demodulation is performed (20 Hz amplitude modulation). The motivation behind this attempt is to capture the relation between segmental and suprasegmental levels, using the S-AMPH technique. We hypothesize that the S-AMPH features are able to capture this correlation and are going to be beneficial in speech synthesis. Furthermore, the additional phoneme-level information is expected to play a significant role.

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

2. Spectral amplitude modulation phase hierarchy

2.1. Spectral amplitude modulation phase hierarchy model

The boundaries for a parsimonious spectral filterbank are identified using the principal component analysis (PCA) procedure [14]. This dimensionality reduction in the frequency domain spanning 100-7250 Hz resulted into the top 5 components contributing the highest amount of variance individually, and cumulatively accounted for 65% of the total variance. The spectral bands were then identified from the rectified component loading patterns, resulting into the filterbank edges of 100, 300, 700, 1750, 3900 and 7250 Hz. Thus, 5 spectral bands were identified in the spectral dimensionality reduction process.

A similar statistical approach was used to identify modulation rate bands [14]. The speech samples were first spectrally-filtered into 5 spectral bands. The Hilbert envelope was then obtained for each spectral band, and this envelope was further filtered into 24 logarithmically-spaced between 0.9-40 Hz modulation rate channels to give a high-dimensional 5 (spectral band) x 24 (modulation rate) channel representation for each speech sample. The aim of the PCA procedure was to reduce this large number of 24 modulation channels into a smaller number of non-redundant modulation rate bands. A descriptive analysis suggested that the entire modulation rate spectrum may be usefully divided into 3 regions: a narrow syllabic rate band at about 4 Hz, a band of slower modulations below 4 Hz that could correspond to the prosodic stress patterns, and a band of faster modulations above 4 Hz. Thus, only the top 3 principal components of the modulation rate PCA procedure, accounting cumulatively for 60-80% of the total variance, are used for the S-AMPH features. The identification of 3 major modulation rate bands or modulation time-scales fits well with theoretical proposals regarding the typical time-scales of 3 major phonological units in speech: stress pattern (about 2 Hz), syllables (about 5 Hz) and onset-rimes/phonemes (about 20 Hz) [11].

In this work, the Matlab implementation of the S-AMPH feature extraction taken from S6 Appendix of [11] is used. Figure 1 shows a scheme of the feature extraction process. In Figure 2 the S-AMPH stress-, syllable- and phoneme-level modulations are shown for the utterance *it's generally a frog or a worm*.

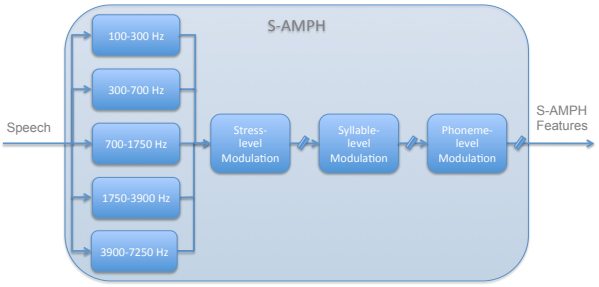

Figure 1: *Three-level spectral amplitude modulation phase hierarchy scheme; stress-level, syllable-level and phoneme-level modulations.*

2.2. Speech synthesis scheme

In this subsection, initially, the DNN-based speech synthesis framework, which follows the framework of [15, 16], and constitutes the baseline system in our experiments (see Section 3) is described and consequently the proposed speech synthesis scheme is presented.

2.2.1. DNN-based speech synthesis framework

A DNN is a feed-forward artificial neural network with multiple hidden layers between the input and output layers, 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 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 [17]. 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 [18]. These labels were aligned with the speech signal through a phone-based forced alignment procedure, using the Kaldi toolkit [19]. 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 [20] vocoder was used for the acoustic analysis and feature extraction, essentially using the default settings from the EMIME [21] 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,

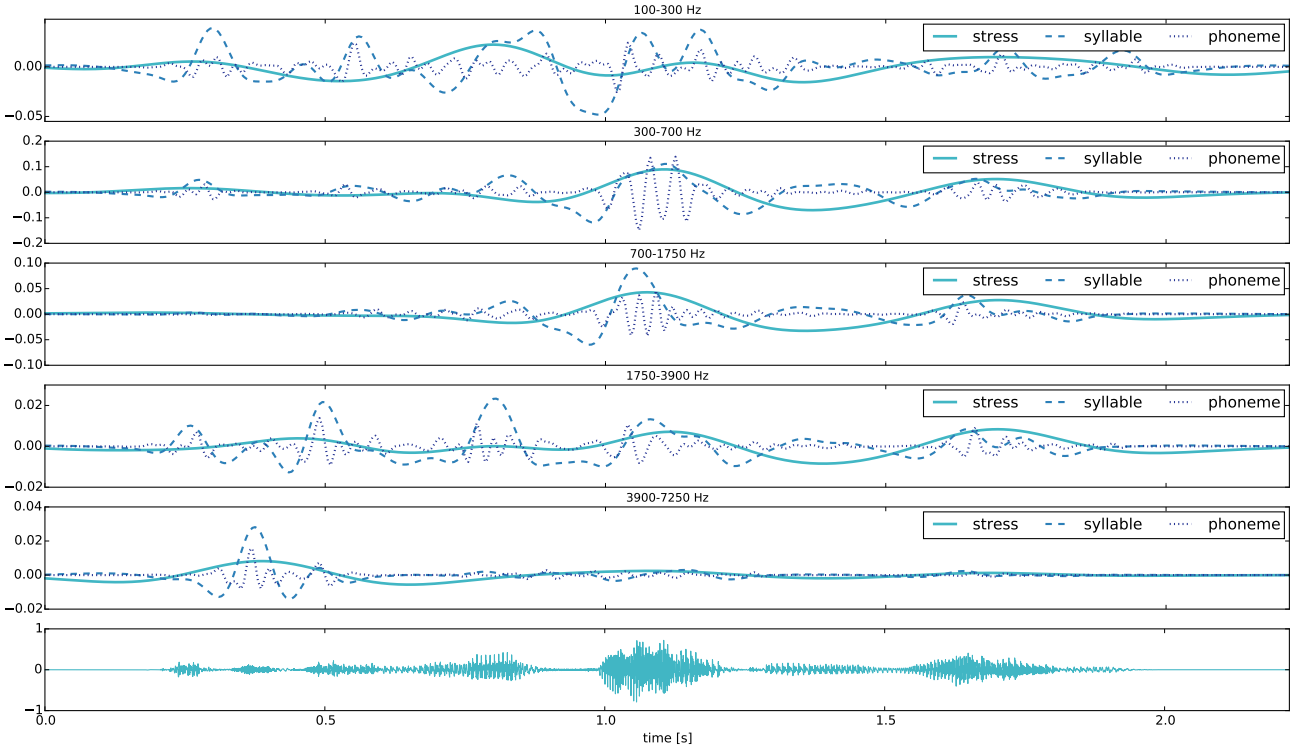

Figure 2: *The S-AMPH stress-, syllable- and phoneme-level modulations for the five spectral bands of the utterance "it's generally a frog or a worm".*

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 those of Zen et al. [15] or Qian et al. [22]), 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 [23], through applying the "mlpg" tool from SPTK [24] and global variance computed on each acoustic component. This was followed by resynthesis using the STRAIGHT vocoder.

2.2.2. S-AMPH speech synthesis framework

In Figure 3, the proposed speech synthesis with S-AMPH feature scheme is shown.

During the training phase, in parallel with the baseline scheme, the S-AMPH scheme is used to extract the S-AMPH 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 S-AMPH 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 S-AMPH features, these features need to be predicted from text. Alternatively, this scheme could be used in a S2ST scenario. In this case the S-AMPH 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. System

3.1. Database

For the experiments the blizzard-challenge-2008 [25, 26] 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, corresponding to approximately 9.6k utterances. For our experiments a subset of the database was used, composed of 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.

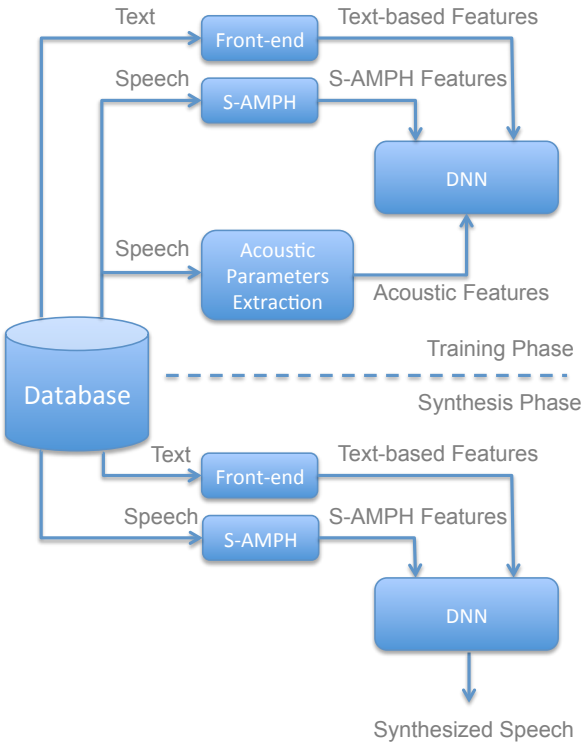

Figure 3: *Speech synthesis scheme.*

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 16 kHz.

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) [27] 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 25 ms and a frame shift of 5 ms were used. The default (not calculated based on the specific speaker) syllable frequency of 5 Hz was selected. The two PAD features were combined with the frame-level text-based input features as described in [12]. Furthermore, 16 neighbouring frames (8 previous and 8 next), were used for PAD features. This parameter was not used in our previous work [12], and it further improves the SSEP system performance.

3.4. S-AMPH features setup

Five-spectral band filtered signal with Hilbert envelopes from each spectral band overlaid was used in S-AMPH model:

- 100-300 Hz
- 300-700 Hz
- 700-1750 Hz
- 1750-3900 Hz
- 3900-7250 Hz

Three modulation rate bands (Stress, Syllable & Phoneme) are extracted from each of the envelopes in the 5 spectral bands:

- Stress-level: 2 Hz
- Syllable-level: 5 Hz
- Phoneme-level: 20 Hz

Furthermore, 10 neighbouring frames (5 previous and 5 next), were used for S-AMPH features.

4. Experiments

To validate our hypothesis, that the S-AMPH features will be beneficial, and further improve the quality of synthetic speech in respect to the baseline and to the SSEP system, 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 three systems. Higher MCD values indicate lower speech quality of the synthesized speech samples. Additionally for evaluating the three systems in respect to prosody modelling, the RMSE of F0 was calculated for each system. These results are presented in Table 1.

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

System	# of neighbouring frames	MCD (dB)	F0 (Hz)
Baseline	0	3.938	19.096
SSEP	0	3.912	18.208
SSEP	16	3.872	17.546
S-AMPH	0	3.744	15.298
S-AMPH	10	3.569	13.602

As can be seen from the results, the reduction in MCD of the SSEP (using neighbouring frames) system over the baseline one is very small, i.e, approximately 1.7% relative improvement. Nonetheless, the reduction of RMSE of F0 of the SSEP (using neighbouring frames) system over the baseline one is approximately 8.1%, showing a small but clear relative improvement in respect to prosody modelling. The results are statistically significant ($p < 0.05$). Furthermore, the speech synthesis system based on S-AMPH features (without neighbouring frames) system is clearly outperforming the SSEP system by 8.8% and 25.3% relative improvement in MCD and RMSE of F0 respectively.

Finally, an attempt was also made to use these features in a multi-task training scheme in the DNN-based speech synthesis scheme. Multi-task training has been recently used in speech synthesis [28, 29, 30], for improving the quality of synthetic speech; not always achieving significant improvement. In our case, when multi-task training was used in either case, i.e., using PAD or S-AMPH features, the improvement shown in the objective evaluation measurements was not significant. Further investigation is needed.

4.2. Subjective evaluation

To further validate our hypothesis and evaluate whether the improvement shown in the objective measurements is perceivable by humans, a subjective evaluation ABX test was performed. The ABX test was performed only between the baseline and the SSEP system without using neighbouring frames. These two systems were selected since the SSEP system showed the smallest improvement with respect to the baseline system.

We employed a 3-point scale ABX subjective evaluation listening test [31], 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%.

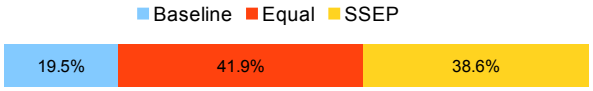

Figure 4: *Subjective evaluation ABX test results (in %) of the baseline and SSEP systems.*

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 and S-AMPH features, when using more expressive and emotional speech, will be bigger.

5. Conclusions and future work

The spectral amplitude modulation phase hierarchy (S-AMPH) technique was used in this paper for improving speech synthesis. The hypothesis that the information which exists in different time-scales of a speech signal and the correlation among these time-scales, would be captured by the S-AMPH features and learned by the DNNs for improving synthetic speech, was validated. The evaluation showed improvement in synthetic speech quality; in terms of relative reduction in mel-cepstral distortion (MCD) by approximately 9% and in terms of relative reduction in root mean square error (RMSE) of the fundamental frequency (F0) by approximately 25%. Multi-task training was also investigated in this work, giving no significant improvements.

It should be pointed out that, since the database used in these experiments consists of read speech, where prosody variations are constrained due to the strict speaking style, it is expected that the importance of both the PAD and the S-AMPH features, when more expressive or emotional speech (e.g. audiobooks) is used, will be substantially bigger.

As future work we intend to also subjectively evaluate the new proposed system using S-AMPH features and neighbouring frames, which has shown the highest performance. Nonetheless, due to the large reduction in the errors in respect to all the other systems, it is expected that the same trend will be seen in this subjective test.

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.

7. References

[1] X. Huang, A. Acero, and H.-W. Hon, "Spoken language processing: A guide to theory, algorithm, and system development," 2001.

[2] T. Dutoit, *An Introduction to Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers, 1997.

[3] S. Furui, *Digital Speech Processing: Synthesis, and Recognition, Second Edition,*, ser. Signal Processing and Communications. Taylor & Francis, 2000. [Online]. Available: https://books.google.ch/books?id=X6mZGqZmcbgC

[4] J. Laver, *Principles of Phonetics*. Cambridge: Cambridge University Press, 1994.

[5] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception." *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/8132899

[6] S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, "Temporal properties of spontaneous speecha syllable-centric perspective," *Journal of Phonetics*, vol. 31, no. 34, pp. 465 – 485, 2003, temporal Integration in the Perception of Speech.

[7] V. Leong, M. A. Stone, R. E. Turner, and U. Goswami, "A role for amplitude modulation phase relationships in speech rhythm perception." *J. Acoust. Soc. Am.*, vol. 136, no. 1, pp. 366–381, Jul. 2014.

[8] Z. Surányi, V. Csépe, U. Richardson, J. M. Thomson, F. Honbolygó, and U. Goswami, "Sensitivity to rhythmic parameters in dyslexic children: A comparison of Hungarian and English," *Reading and Writing*, vol. 22, no. 1, pp. 41–56, 2009.

[9] U. Goswami, H.-L. S. Wang, A. Cruz, T. Fosker, N. Mead, and M. Huss, "Language-universal sensory deficits in developmental dyslexia: English, Spanish, and Chinese." *J. Cognitive Neuroscience*, vol. 23, no. 2, pp. 325–337, 2011.

[10] U. Goswami, "A temporal sampling framework for developmental dyslexia," *Trends in cognitive sciences*, vol. 15, no. 1, pp. 3–10, 2011.

[11] V. Leong and U. Goswami, "Acoustic-emergent phonology in the amplitude envelope of child-directed speech," *PLoS ONE*, vol. 10, pp. 1–37, 12 2015.

[12] A. Lazaridis, M. Cernak, and P. N. Garner, "Probabilistic amplitude demodulation features in speech synthesis for improving prosody," Idiap, Idiap-RR Idiap-RR-12-2016, 4 2016.

[13] R. E. Turner and M. Sahani, "Demodulation as Probabilistic Inference," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2398–2411, Nov. 2011.

[14] V. Leong, "Prosodic rhythm in the speech amplitude envelope: Amplitude modulation phase hierarchies (AMPHs) and AMPH models," Ph.D. dissertation, University of Cambridge, Cambridge, Sep. 2012.

[15] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using Deep Neural Networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7962–7966.

[16] A. Lazaridis, B. Potard, and P. N. Garner, "DNN-based speech synthesis: Importance of input features and training data," in *International Conference on Speech and Computer, SPECOM 2015*, ser. Lecture Notes in Computer Science, N. F. A. Ronzhin, R. Potapova, Ed. Springer Berlin Heidelberg, 2015, pp. 193–200.

[17] A. Black, P. Taylor, and R. Caley, "The festival speech synthesis system: System documentation (1.3.1)," Human Communication Research Centre, Technical Report HCRC/TR-83, December 1998.

[18] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[19] 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.

[20] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[21] M. Wester, J. Dines, M. Gibson, H. Liang, Y.-J. Wu, L. Saheer, S. King, K. Oura, P. N. Garner, W. Byrne, Y. Guan, T. Hirsimäki, R. Karhila, M. Kurimo, M. Shannon, S. Shiota, J. Tian, K. Tokuda, and J. Yamagishi, "Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project," in *SSW7*, 2010, pp. 192–197.

[22] Y. Qian, Y. Fan, W. Hu, and F. 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*, 2014, pp. 3829–3833.

[23] HTS, "HMM-based speech synthesis system version 2.1," 2010. [Online]. Available: http://hts.sp.nitech.ac.jp

[24] S. Imai and T. Kobayashi, "Speech signal processing toolkit (SPTK) version 3.7," 2013.

[25] V. Strom, R. Clark, and S. King, "Expressive prosody for unit-selection speech synthesis," in *Proc. Interspeech*, Pittsburgh, 2006.

[26] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Proc. Interspeech 2007*, Antwerp, Belgium, Aug. 2007.

[27] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. of ICASSP*, vol. 1. IEEE, May 1993, pp. 125–128 vol.1.

[28] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[29] Q. Hu, Z. Wu, K. Richmond, J. Yamagishi, Y. Stylianou, and R. Maia, "Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 854–858.

[30] M. Ribeiro, O. Watts, J. Yamagishi, and R. Clark, "Wavelet-based decomposition of F0 as a secondary task for DNN-based speech synthesis with multi-task learning," in *The 41st International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016)*, 2016.

[31] V. Grancharov and W. B. Kleijn, "Speech Quality Assessment," in *Springer Handbook of Speech Processing*, J. Benesty, Sondhi, and Y. Huang, Eds. Springer Berlin Heidelberg, 2008, pp. 83–100.