Effects of F0 Estimation Algorithms on Ultrasound-Based Silent Speech Interfaces

Pengyu Dai
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
pengyudaii@gmail.com

Mohammed Salah Al-Radhi
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
malradhi@tmit.bme.hu

Tamás Gábor Csapó
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
csapot@tmit.bme.hu

***Abstract*—This paper shows recent Silent Speech Interface (SSI) progress that translates tongue motions into audible speech. In our previous work and also in the current study, the prediction of fundamental frequency (F0) from Ultra-Sound Tongue Images (UTI) was achieved using articulatory-to-acoustic mapping methods based on deep learning. Here we investigated several traditional discontinuous speech-based F0 estimation algorithms for the target of UTI-based SSI system. Besides, the vocoder parameters (F0, Maximum Voiced Frequency and Mel-Generalized Cepstrum) are predicted using deep neural networks, with UTI as input. We found that those discontinuous F0 algorithms are predicted with a lower error during the articulatory-to-acoustic mapping experiments. They result in slightly more natural synthesized speech than the baseline continuous F0 algorithm. Moreover, experimental results confirmed that discontinuous algorithms (e.g. Yin) are closest to original speech in objective metrics and subjective listening test.**

Keywords—Silent Speech Interface, Articulatory-To-Acoustic Mapping, Fundamental Frequency

I. INTRODUCTION

During the past few years, there has been a significant interest in articulatory-to-acoustic conversion, which is often referred to as "Silent Speech Interface" (SSI) [1]. This has the main idea of recording the soundless articulatory movement and automatically generating speech from the movement information without the subject producing any sound. Such an SSI system can be beneficial for the speaking impaired (e.g. after laryngectomy). For scenarios where regular speech is not feasible, information should be transmitted from the speaker (e.g. extremely noisy environments; military applications). For this automatic conversion task, typically ultrasound tongue imaging (UTI) [2, 3, 4, 5, 6], permanent magnetic articulography (PMA) [7], electromagnetic articulography (EMA) [8], electromyography (EMG) [9] or multimodal approaches [10] are employed.

State-of-the-art SSI systems use the 'direct synthesis' principle, where the speech signal is generated directly from the articulatory data, using vocoders [3]. However, most of these approaches focus on predicting just the spectral features of the vocoder (e.g. Mel-Generalized Cepstrum, MGC). The reason for this is that while there is a direct relation between tongue movement and the spectral content of speech, the F0 value depends on the vocal fold vibration, which has no direct connection with the movement of the tongue and face or the opening of the lips [11]. However, there is some evidence that tongue shapes differ in voiced and unvoiced sounds; for example, the vibration of the vocal folds may slow down during consonant articulation [13]. Along with other factors, these changes correlate with the specific articulatory configuration of the obstruents; that is, the volume of space between the glottis and the obstacle [14]. Despite these facts, most authors studying SSI systems take the unpredictability of F0 for granted and use the original F0, a constant F0 or white noise as excitation.

A few studies attempted to predict the voicing feature and the F0 curve using articulatory data as input. Nakamura et al. utilized EMG data, and they divided the problem into two steps. First, they used support vector machines (SVM) for voiced/unvoiced (V/U) discrimination, and in the second step, they applied a Gaussian mixture model (GMM) for generating the F0 values. According to their results, EMG-to-F0 estimation achieved a correlation of 0.5, while the V/U decision accuracy was 84% [9]. Hueber et al. experimented with predicting the V/U parameter and the spectral features of a vocoder, using ultrasound and lip video as input articulatory data. They applied a feed-forward deep neural network (DNN) for the V/U prediction and attained an accuracy score of 82%, which is very similar to Nakamura et al. [3]. Another two studies experimented with EMA-to-F0 prediction. Liu et al. compared DNN, RNN and LSTM neural networks to predict the V/U flag and voicing. They found that the strategy of cascaded prediction, namely using the predicted spectral features like auxiliary input, increases the accuracy of excitation feature prediction [15].

Zhao et al. found that the velocity and acceleration of EMA movements are effective in articulatory-to-F0 prediction and that LSTMs perform better than DNNs in this task. However, although their objective F0 prediction scores were promising, they did not evaluate their system in subjective human listening tests [16].

Although there has been some research on articulatory-to-F0 prediction, only two deep learning experiments for estimating the F0 curve from ultrasound tongue images alone are proposed [17, 18]. We presented our results for DNN-based F0 estimation from ultra-sound images [18]. In contrast with others who worked with EMG signals, our input articulatory representation contains no information directly related to vocal fold vibration. We applied a 2-stage DNN-based approach where one machine learning model seeks to estimate the voicing feature, while another one aims to predict the F0 value for voiced frames. During the evaluation (synthesis) step, the outputs of the two DNNs are merged. It was achieved by taking the output value of the F0 predictor network where the voicing network decided in favor of voicing and returning a constant value for frames judged to be

978-1-6654-2786-9/21/$31.00 ©2021 IEEE

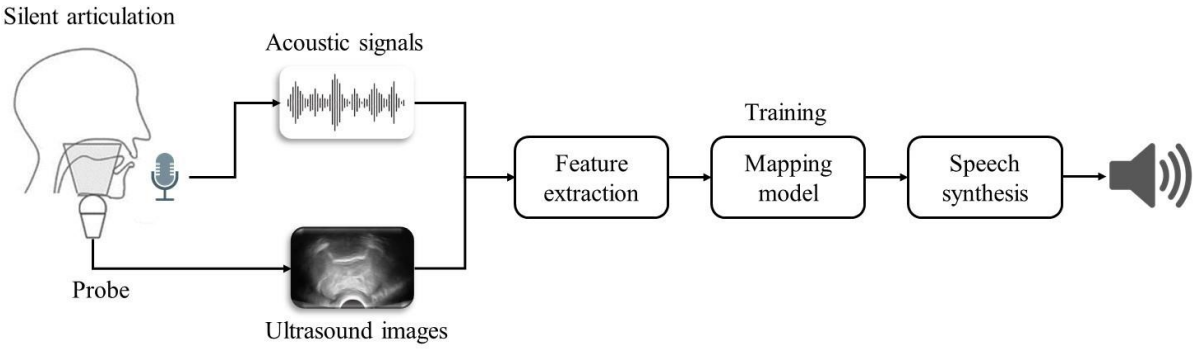

Fig. 1. Workflow of an ultrasound-based silent speech.

unvoiced. We attained a correlation rate of 0.74 between the original and the predicted F0 curve in the experiments. And in subjective listening tests, our subjects could not distinguish between the sentences synthesized using the DNN-estimated or the original F0 curve and ranked them as having the same quality. However, only a single F0 estimation algorithm based on Idiap [19] was implemented [17].

Here, we extended our study by investigating different robust F0 estimation techniques: Yaapt [20], Rapt [21], DIO [22] and Yin [23]. In contrast with our recent work where Idiap worked as a continuous pitch algorithm implemented with a continuous vocoder, the new four algorithms are discontinuous and implemented with a discontinuous vocoder. We discovered in our experiments that all discontinuous algorithms got better values than Idiap (being the baseline of the current paper) in objective and subjective measurements.

II. Methods

A. Data Acquisition Protocol

Two Hungarian male and two female subjects with normal speaking abilities were recorded while reading sentences aloud (altogether 209 sentences each), and the data of a female speaker was used in our current experiments. The sentences are divided into two distinct sets, 200 were selected for training and validation sets, 9 for the test set. The tongue movement was recorded in midsagittal orientation using the "Micro" ultrasound system of Articulate Instruments Ltd. at 82 fps. The speech signal was recorded with a Beyerdynamic TG H56c tan omnidirectional condenser microphone.

Moreover, a single-speaker dataset with data of one professional voice talent, a male native speaker of English is also tested in this work. The speaker was fitted with the UltraFit stabilising helmet, which held the video camera and the ultrasound probe. Data was recorded using the Articulate Assistant Advanced (AAA) software. Ultrasound was recorded using Articulate Instruments' Micro system at ~80fps with a 92o field of view. A single B-Mode ultrasound frame has 842 echo returns for each of 64 scan lines, giving a 64 × 842 "raw" ultrasound frame that captures a midsagittal view of the tongue. The speaker was seated in a hemi-anechoic chamber and audio was captured with a Sennheiser HKH 800 p48 microphone with a 48KHz sampling frequency at 16 bit [62]. In the experiment, the recorded audios were resampled to 22KHz and the ultrasound images were resized to 64 × 128.

The ultrasound data and the audio signals were synchronized using the tools provided by Articulate Instruments Ltd. In the experiments below, the raw scanline data of the ultrasound was used as input data for the DNNs. The images were reduced to 64128 pixels (for details, see [6]).

B. Feature Extraction and Speech Synthesis

The general workflow of ultrasound-based silent speech interface is shown in Fig. 1. We applied the SPTK vocoder for the analysis and synthesis of speech (http://sptk.sourceforge.net). The speech signal was lowpass filtered and resampled to 22 050 Hz. The F0 curve was extracted by Idiap, Yaapt, Rapt, Dio and Yin, respectively. We extracted 12 Mel-Generalized Cepstrum-based Line Spectral Pair (MGC-LSP) features along with the gain, which resulted in a 13-dimensional feature vector. This vector served as the training target during DNN training. In the synthesis phase, we replaced all parameters required by the synthesizer with the estimates produced by the DNN. The vocoder generated an impulse-noise excitation according to the F0 parameter and applied spectral filtering using the MGC-LSP coefficients and a Mel-Generalized Log Spectral Approximation (MGLSA) filter [24] to reconstruct the speech signal.

C. DNN-based Fundamental Frequency Estimation

DNNs were used in two major machine learning components, one dedicated to making the voiced/unvoiced decision, while the second was to estimate the actual F0 value for voiced frames.

Since the V/U decision for each frame has a binary output, we treated it as a classification task. While working on the same input images, the second DNN seeks to learn the F0 curve. This second task was viewed as a regression problem, and it was trained with the voiced segments from the training data. The outputs of the two DNNs were merged during the evaluation (synthesis) step. For Idiap, this is achieved by taking the output value of the F0 predictor network where the voicing network decided in favor of voicing and returning a constant value for frames judged to be unvoiced. For Yaapt and another three algorithms, only those predicted F0 values from voiced frames are used.

We trained DNNs with five hidden layers of 1000 ReLU neurons. The F0 parameter was predicted together with the gain and the 12 LSP parameters. This DNN contained 14 linear neurons in its output layer. The network trained for the binary U/V decision task had the same structure but with a binary classification output layer.

To evaluate the best F0 predicting algorithm via subjective listening test, we synthesized 2 reference sentences. To have an upper glass ceiling, we synthesize

TABLE I. AVERAGE OBJECTIVE SCORES BASED ON HUNGARIAN SYNTHESIZED SPEECH SIGNALS. THE BOLD VALUE DENOTES THE BEST RESULTS

Method	Evaluation Metric				
	IS	***LLR***	***CEP***	***fwSNRseg***	***ESTOI***
Idiap (baseline)	4.4821	0.6078	4.5801	5.7718	0.3645
Rapt	1.1673	0.5014	3.9928	6.9196	0.3897
Yaapt	**0.5664**	**0.4772**	**3.8166**	**7.1242**	**0.4134**
DIO	1.4039	0.5103	3.9604	7.0647	0.3881
Yin	3.0025	0.5397	4.0710	6.8494	0.3754

TABLE II. AVERAGE OBJECTIVE SCORES BASED ON ENGLISH SYNTHESIZED SPEECH SIGNALS. THE BOLD VALUE DENOTES THE BEST RESULTS

Method	Evaluation Metric				
	IS	***LLR***	***CEP***	***fwSNRseg***	***ESTOI***
Idiap (baseline)	6.7277	0.6727	4.4453	5.3645	0.2711
Rapt	**4.6177**	**0.6314**	**4.2897**	**5.4879**	0.2783
Yaapt	9.1106	0.6869	4.5444	5.2307	**0.2853**
DIO	18.4918	0.9091	5.4055	4.3135	0.2809
Yin	4.6803	0.6369	4.3115	5.4340	0.2852

sentences using the original F0 curve. To have a benchmark/lower anchor version, we synthesized sentences using a constant F0, where the V/U network predicted the voicing of the actual ultrasound images.

III. RESULTS AND DISCUSSION

A. Objective Evaluation

The performance of F0 detection algorithms is evaluated by comparing their synthesized speech and original speech. 5 metrics are used: (1) IS (Itakura–Saito) [25] as

$$d_{IS}(\vec{a}_p, \vec{a}_c) = \frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\vec{a}_p R_c \vec{a}_p^T}{\vec{a}_c R_c \vec{a}_c^T}\right) + \log\left(\frac{\sigma_c^2}{\sigma_p^2}\right) - 1 \qquad (1)$$

where σ_p^2 and σ_c^2 are the LPC gains of the clean and processed signals, respectively; (2) LLR (log-likelihood ratio) [25] as

$$d_{LLR}(\vec{a}_p, \vec{a}_c) = \log\left(\frac{\vec{a}_p R_c \vec{a}_p^T}{\vec{a}_c R_c \vec{a}_c^T}\right) \qquad (2)$$

where $\vec{a}_c$ is the LPC vector of the clean speech signal, $\vec{a}_p$ is the LPC vector of the processed enhanced speech signal, and R_c is the autocorrelation matrix of the noise-free speech signal; (3) CEP (cepstrum distance measures) [26] as

$$d_{CEP}(\vec{c}_c, \vec{c}_p) = \frac{10}{\log 10}\sqrt{2\sum_{k=1}^{p}[c_c(k) - c_p(k)]^2} \qquad (3)$$

where $\vec{c}_c$ and $\vec{c}_p$ is are the CEP coefficient vectors of the noise-free and processed signals, respectively; (4) fwSNRseg (frequency-weighted segmental SNR) [27] as

$$fwSNRseg = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)\log_{10}\frac{X(j,m)^2}{(X(j,m)-\hat{X}(j,m))^2}}{\sum_{j=1}^{K} W(j,m)} \qquad (4)$$

where W(*j*, *m*) is the weight placed on the *j*th frequency band, K is the number of bands, M is the total number of frames in the signal, X(*j*, *m*) is the critical-band magnitude (excitation spectrum) of the clean signal in the *j*th frequency band at the *m*th frame, and $\hat{X}(j, m)$ is the corresponding spectral magnitude of the enhanced signal in the same band; and (5) ESTOI (Extended ShortTime Objective Intelligibility) [28]. IS and LLR directly calculate the distance between two sets of linear prediction coefficients (LPC) on the original and the predicted speech. In contrast, CEP distance provides an estimate of the log spectral distance between two speeches. fwSNRseg was adopted in the time domain for the error criterion. ESTOI calculates the correlation between the temporal envelopes of original and predicted speech. For all measures, a calculation is done frame-by-frame and a smaller value indicates better performance except for the fwSNRseg measure (higher value is better). This objective evaluation was done on test data (9 sentences).

TABLE I. and II list the results of various measurement methods (note that our goal is to minimize IS, LLR and CEP, while maximizing fwSNRseg and ESTOI). It can be seen that the Yaapt performs extremely well in each metrics in the Hungarian corpus, whereas Rapt could be seen as the best one in the English corpus.

Comparing the baseline with others, we can observe that discontinuous algorithms get a better score than the baseline in every metrics. It shows that speech signal synthesized by the predicted discontinuous F0 curve are much closer to the original speech signal. F0 predicted by discontinuous algorithms with discontinuous vocoder have better performance than the baseline.

B. Subjective Evaluation

To find out which investigated model is closer to natural speech, we conducted an online MUSHRA-like (Multi-Stimulus test with Hidden Reference and Anchor) listening test [29]. The advantage of MUSHRA is that it allows the evaluation of multiple samples in a single trial without breaking the task into many pairwise comparisons. We aimed to compare natural and synthesized baseline sentences with the synthesized sentences using four discontinuous F0 extraction algorithms. We used a benchmark/ lower anchor

sentence with constant F0 and a distorted version of the original MGC features. Five sentences were selected for the test, which is not included in the training database. All sentences appeared in randomized order (different for each listener). In the MUSHRA test, the listeners had to rate the naturalness of each stimulus in a randomized order relative to the reference (which was the natural sentence), from 0 (highly unnatural) to 100 (high natural).

Altogether 16 listeners participated in the main test (6 females, 10 males). None of them indicated any hearing loss. The subjects were between 21-47 years (mean 24 years). On average, the whole test took 12 minutes to complete. Fig. 2 shows the average naturalness score for these experimented algorithms. The benchmark version (const F0) achieved the lowest score, while the natural sentences (natural) were rated the highest, as expected. Comparing with other discontinuous algorithms, the baseline Idiap get the lowest score, which means all discontinuous algorithms based predicted sentences sound more natural than baseline. We also noticed that the score of the four discontinuous algorithms is very similar. The reason might be their synthesized sentences are relatively close, and it is hard for a human being to distinguish their subtle differences. To check the statistical significance of the differences, we conducted Mann-Whitney-Wilcoxon rank-sum tests with a 95% confidence level, showing that the result of the Yin algorithm was significantly different from the baseline. In contrast, the other differences are not significant.

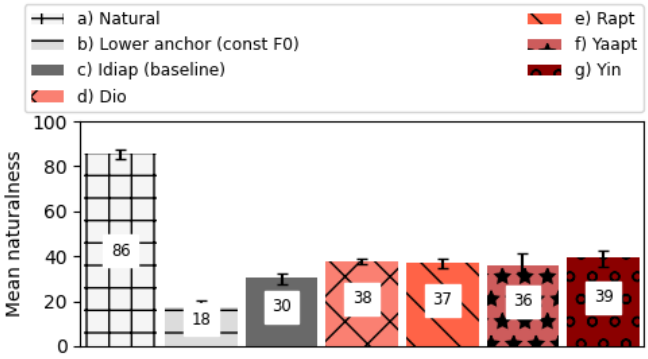

Fig. 2. Results of the subjective listening test. The error bars show the 95% confidence intervals.

IV. Conclusions

In this work we described our experiments comparing several discontinuous F0 estimation algorithms with a continuous baseline one in ultrasound-based articulatory-to-acoustic mapping. We used four accurate discontinuous F0 estimation algorithms to predict the F0 value of voiced frames. The objective and subjective evaluation results demonstrated that F0 predicted by discontinuous algorithms and the synthesized sentences outperform the one based on continuous F0 (baseline). The experiments were run on the voice of only one Hungarian female speaker. We plan to repeat our experiments with more speakers (both male and female) and with English data. Besides, it will be worth applying recurrent neural networks to consider the sequential nature of articulatory and speech data. For a practical Silent Speech Interface, it will be necessary to use speaker adaptation techniques, i.e. in the future, we plan to test how the UTI-to-F0 algorithms trained on one speaker work with other speakers or with real silent articulation.

Acknowledgment

The research was partly supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825619 (AI4EU). The authors were partially funded by the National Research, Development and Innovation Office of Hungary (FK 124584, PD 127915 grants). The Titan X GPU for the deep learning experiments was donated by the NVIDIA Corporation. We would like to thank Gábor Gosztolya for his comments on this manuscript. We thank the listeners for participating in the subjective test.

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg, "Silent speech interfaces," Speech Communication, vol. 52, no. 4, pp. 270–287, 2010.

[2] B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," in Proc. ICASSP, Montreal, Quebec, Canada, pp. 685–688, 2004.

[3] T. Hueber, E. Benaroya, B. Denby, and G. Chollet, "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface," in Proc. Interspeech, Florence, Italy, pp. 593–596, 2011.

[4] T. Hueber, G. Bailly, and B., "Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface," in Proc. Interspeech, Portland, OR, USA, pp. 723–726, 2012.

[5] A. Jaumard-Hakoun, K. Xu, C. Leboullenger, P.R. Ragot, and B. Denby, "An Articulato-ry-Based Singing Voice Synthesis Using Tongue and Lips Imaging," in Proc. Interspeech, pp. 1467–1471, 2016.

[6] T.G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "DNN-Based Ultra-sound-to-Speech Conversion for a Silent Speech Interface," in Proc. Interspeech, Stockholm, Sweden, pp. 3672–3676, 2017.

[7] J. Wang, A. Samal, and J. Green, "Preliminary Test of a Real-Time, Interactive Silent Speech Interface Based on Electromagnetic Articulograph," in Proc. SLPAT, pp. 38–45, 2014.

[8] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces," PLOS Computational Biology, vol. 12, no. 11, pp. e1005119, nov 2016.

[9] K. Nakamura, M. Janke, M. Wand, and T. Schultz, "Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0," in Proc. ICASSP, Prague, Czech Republic, pp. 573–576, 2011.

[10] J. Freitas, A. Ferreira, M. A. Figueiredo, A. Teixeira and M. Dias, "Enhancing multimodal silent speech interfaces with feature selection,"in Proc. Interspeech, Singapore, Singapore, pp. 1169–1173, 2014.

[11] J.A. Gonzalez, L.A. Cheah, P.D. Green, J.M. Gilbert, S.R. Ell, R.K. Moore, and E. Holdsworth, "Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary," In Proc. ISCA, Interspeech, Stockholm, Sweden, pp. 3986–3990, 2017.

[12] J. Jiang, A. Alwan, L. E. Bernstein, P. Keating, and E. Auer, "On the correlation between facial movements, tongue movements, and speech acoustics," EURASIP Journal on Applied Signal Processing, vol. 2002, no. 11, pp. 174–1188, 2002.

[13] C.A. Bickley and K.N. Stevens, "Effects of a vocal tract constriction on the glottal source: experimental and modeling studies," Journal of Phonetics, vol. 14, pp. 373–382, 1986.

[14] J.R. Westbury and P.A. Keating, "On the naturalness of stop consonant voicing," Journal of Linguistics, vol. 22, pp. 145–166, 1986.

[15] Z.C. Liu, Z.H. Ling, and L.R. Dai, "Ar-ticulatory-to-acoustic conversion with cascaded prediction of spectral and excitation features using neural networks," in Proc. Interspeech, San Francisco, CA, USA, pp. 1502–1506, 2016.

[16] C. Zhao, L. Wang, J. Dang, and R. Yu, "Prediction of F0 based on articulatory features using DNN," in Proc. ISSP, Tienjin, China, 2017.

[17] T.G. Csapó, M.S. Al-Radhi, G. Németh, G. Gosztolya, T.G., L. Tóth, A. Markó, "Ultrasound-based Silent Speech Interface Built on a Continuous Vocoder", in Interspeech, pp. 894-898, 2019.

[18] T. Grósz, G. Gosztolya, L. Tóth, T.G. Csapó, and A. Markó, "F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces," in Proc. ICASSP, pp. 291-295, 2018.

[19] P.N. Garner, M. Cernak, P. Motlicek, "A simple continuous pitch estimation algorithm," IEEE Signal Processing Letters, vol. 20, no. 1, pp. 102–105, 2013.

[20] S.A. Zahorian, and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking." The Journal of the Acoustical Society of America, vol. 123, no .6, pp. 4559-4571, 2008.

[21] T. David, and W.B. Kleijn. "A robust algorithm for pitch tracking (RAPT)." Speech coding and synthesis, pp. 495-518, 1995.

[22] M. Morise, H. Kawahara, and T. Nishiura, "Rapid f0 estimation for highsnr speech based on fundamental component extraction," IEICE Transactions on Information and Systems, (Japanese Edition), vol. J93-D, no.2, pp.109-117, 2010.

[23] D.C. Alain and H. Kawahara. "YIN, a fundamental frequency estimator for speech and music." The Journal of the Acoustical Society of America vol. 111, no. 4, pp. 1917-1930, 2002.

[24] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," Electronics and Communications in Japan (Part I: Communications), vol. 66, no. 2, pp. 10–18, 1983.

[25] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements, Objective Measures of Speech Quality, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[26] K. Nobuhiko, N. Hiromi, and I. Kenzo, "Objective quality evaluation for low bitrate speech coding systems", IEEE Journal on Selected Areas in Communications, vol. 6, pp. 262–273, 1988.

[27] J. M. Tribolet, P. Noll, B. J. McDermott and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," in Proc. ICASSP, Oklahoma, USA, pp.586-590, 1978.

[28] J. Jensen and C.H. Taal, "An Algorithm for Pre-dicting the Intelligibility of Speech Masked by Modulated Noise Maskers," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 11, pp. 2009–2022, 2016.

[29] "ITU-R Recommendation BS.1534: Method for the subjec-tive assessment of intermediate audio quality," 2001.