Using Hidden Markov Models for Speech Enhancement

Akihiro Kato and Ben Milner

University of East Anglia

akihiro.kato@uea.ac.uk, b.milner@uea.ac.uk

Abstract

This work presents an approach to speech enhancement that operates using a speech production model to reconstruct a clean speech signal from a set of speech parameters that are estimated from the noisy speech. The motivation is to remove the distortion and residual and musical noises that are associated with conventional filtering-based methods of speech enhancement. The STRAIGHT vocoder forms the model for speech reconstruction and requires a time-frequency surface and fundamental frequency information. Hidden Markov model synthesis is used to create an estimate of the time-frequency surface and this is combined with the noisy surface using a perceptually motivated signal-to-noise ratio weighting. Experimental results compare the proposed reconstruction-based method to conventional filtering-based approaches of speech enhancement.

Index Terms: speech enhancement, hidden Markov models, STRAIGHT

10.21437/Interspeech.2014-576

1. Introduction

The aim of this work is to enhance a speech signal by reconstructing it from a speech model and a set of parameters extracted from the noisy speech. This moves away from conventional methods of speech enhancement that apply filtering to remove noise from the noisy speech. Such methods can introduce distortion and musical noise and leave residual noise in the enhanced speech. The motivation for the reconstruction-based approach is that if the speech model and parameters are sufficiently accurate then the resulting speech should be free from noise and distortion.

Conventional filtering approaches to speech enhancement are normally a two stage process of first estimating the contaminating noise, or signal-to-noise ratio (SNR), and then removing the noise from the noisy speech. Many methods have been proposed for speech enhancement and can be broadly categorised into spectral subtraction, Wiener filtering, statistical methods and subspace methods [1]. Spectral subtraction methods are the most simple and in principle require just a noise spectrum estimate which is subtracted from the noisy signal to give a clean speech spectrum [2]. Inaccuracies in noise estimation lead to over or under subtraction which can introduce musical noise and speech distortion or leave residual noise, although numerous methods have attempted to reduce these artefacts [3, 4]. Wiener filtering gives generally higher speech quality than spectral subtraction although its implementation is more complex as an estimate of the SNR is required rather than just a noise estimate. Iterative approaches and decision directed methods to obtain the *a priori* SNR have been shown effective [5, 6]. Statistical methods, and in particular log MMSE, estimate directly log spectral magnitudes from the noisy speech [7]. These give further improvements in speech quality and are considered as being some of the most effective filtering-based methods of speech enhancement. Subspace methods operate differently and transform the noisy speech into a new space that comprises speech and noise subspaces [8, 9]. Truncation aims to retain speech components and remove noise components. Retaining too few speech components oversmooths the speech, while retaining too many components leaves residual noise. As the speech and noise spaces are not entirely separable it is necessary to apply further filtering to the speech components to remove residual noise. Whilst these methods are effective, they do not perform as well as the statistical methods [10]. Evaluation of all of these different methods of speech enhancement shows them to be effective in improving speech quality but reveals them to be dependent on the accuracy of noise estimation. The effect of such errors is to introduce unwanted artefacts into the enhanced speech such as musical noise, residual noise and distortion.

This work aims to address these issues by using a model of speech production to reconstruct clean speech from a set of parameters extracted from the noisy speech. No filtering of the speech takes place and so no artefacts from filtering should be introduced. This introduces to two main challenges – i) to find a sufficiently good model for speech reconstruction and ii) to develop robust methods to accurately estimate noise-free acoustic parameters needed by the model. The STRAIGHT vocoder is used as the speech model given its success in speech synthesis and this requires a time-frequency surface and fundamental frequency contour [11]. In this work it is proposed to estimate the time-frequency surface by combining the original noisy time-frequency surface with a noise-free time-frequency surface produced from the output of a network of HMMs using HMM synthesis techniques [12]. The two time-frequency surfaces are combined according to a localised non-linear function of the signal-to-noise ratio. This gives more weight to time-frequency regions from the original speech signal in high SNR regions, while at low SNRs the HMM-synthesised surface is given more weight.

The remainder of the paper is organised as follows. Section 2 gives an overview of the proposed reconstruction-based speech enhancement system and also the STRAIGHT speech model. Sections 3 and 4 describe how the parameters needed by STRAIGHT, namely the time-frequency surface and fundamental frequency/aperiodicity, are estimated from the noisy speech. Experimental results comparing the proposed method with conventional filtering methods using PESQ are given in Section 5.

2. Speech enhancement

Speech enhancement is achieved by reconstructing a clean speech signal using a set of speech parameters estimated from the noisy speech. Such an approach has two main challenges. First, a suitable speech production model is needed that can synthesise high quality, distortion-free speech from a set of speech parameters. Second, robust algorithms are required to extract

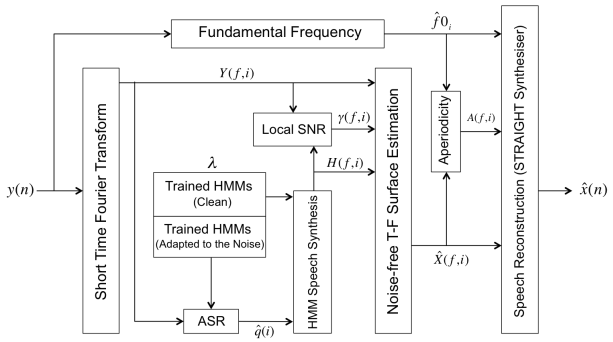

Figure 1: *Framework for reconstruction-based speech enhancement.*

the speech parameters from the noisy speech signals.

A wide variety of engineering models of speech production have been proposed across the areas of speech coding, speech recognition and speech synthesis and many could be considered for use in this work. These include vocoder/LPC-based models such as CELP, ACELP and RELP [13] and sinusoidal and harmonic plus noise models [14, 15]. A variant of the vocoder is STRAIGHT which has found widespread use in HMM-based speech synthesis [11, 16] and as such is chosen as the speech model in this work. The next two subsections give a brief overview of STRAIGHT and then describe the proposed reconstruction-based method of speech enhancement.

2.1. STRAIGHT vocoder

STRAIGHT is a vocoder designed to eliminate periodicity and phase effects and is able to synthesise good quality speech. A variable filter models vocal tract and a series of impulses represent the excitation signal [11]. Quality is improved by all-pass filtering the excitation signal to modify its phase which reduces the typical buzzy sound associated with vocoders. The vocal tract filter is implemented as a minimum phase impulse response filter. To reconstruct a speech signal, STRAIGHT requires three input parameters – the fundamental frequency, $f0_i$, a time-frequency surface, $X(f,i)$ and a measure of aperiodicity, $A(f,i)$, where f and i represent the frequency and frames indices. A fine temporal resolution for these parameters is necessary with typically 1000 frames extracted per second.

2.2. Speech enhancement framework

Extracting the parameters needed by STRAIGHT from a clean speech signal is relatively straightforward and results in good quality speech being synthesised. However, when the speech is noisy, the parameters extracted become distorted and when input into STRAIGHT the resulting speech quality deteriorates. The challenge for enhancement is therefore to provide noise-free estimates of the parameters from noisy speech signals. Figure 1 shows the proposed framework to estimate a set of clean speech parameters. From a noisy speech signal, $y(n)$, processing stages to obtain noise-free estimates of the time-frequency surface, $\hat{X}(f,i)$, the fundamental frequency, $f0_i$ and aperiodicity, $A(f,i)$ are shown. These signals are input into STRAIGHT to reconstruct the enhanced speech signal. Section 3 explains the procedure for estimating the clean time-frequency surface while Section 4 describes fundamental frequency and aperiodicity estimation.

3. Clean time-frequency surface estimation

The estimate of the noise-free time-frequency surface, $\hat{X}(f,i)$ is obtained from an SNR-dependent weighting of the time-frequency surface extracted from the noisy signal, $Y(f,i)$, and a clean time-frequency surface, $H(f,i)$, that is synthesised by an HMM

$$\hat{X}(f,i) = \gamma(f,i)Y(f,i) + (1-\gamma(f,i))\,H(f,i) \quad (1)$$

$\gamma(f,i)$ is a time and frequency dependent non-linear weighting function of the local SNR. In regions with high local SNRs $\gamma(f,i)$ should be close to one so that relatively noise-free regions contribute strongly to the clean estimate of the time-frequency surface. At low SNRs, $\gamma(f,i)$ should be close to zero to attenuate noisy regions of $Y(f,i)$ and replace them with noise-free estimates synthesised from the HMMs. Implementation therefore requires the HMM synthesised time-frequency surface, $H(f,i)$, and estimation of the SNR function $\gamma(f,i)$.

3.1. HMM estimate of time-frequency surface

HMM estimation of the time-frequency surface, $H(f,i)$, is based on techniques used in HMM-based speech synthesis [12, 17]. For synthesis a network of HMMs is trained on features containing spectral envelope and excitation information and duration information is also modelled. To synthesise a speech signal the HMMs output the maximum likelihood set of observation vectors given the word sequence to be synthesised. For a system trained on only static features the output vectors are a piecewise set of state mean vectors which result in poor quality speech. Instead, temporal derivatives are included in the feature vectors with the result that the HMMs produce a smoother set of features for synthesis.

The application to speech enhancement has several differences. First, no word sequence is available as is the case with speech synthesis. Second, the HMMs only need to model spectral envelope as excitation information can be provided by fundamental frequency estimates from the speech signal itself. This is advantageous as more simple HMMs can be used that do not need to model the discontinuities of voiced and unvoiced parts of speech and need no explicit durational modelling.

3.1.1. HMM training

A set of HMMs is trained on feature vectors, $\boldsymbol{c}_i$, that are extracted from a clean speech training database. Essentially, each feature vector can be viewed as comprising a row vector taken from the time-frequency surface, $X(f,i)$, extracted using a short-time Fourier transform with a log operation applied

$$\boldsymbol{x}_i = [\log(X(0,i), \log(X(1,i), \ldots, \log(X(F-1,i)] \quad (2)$$

To improve the smoothness of the synthesised feature vectors a velocity derivative, $\Delta\boldsymbol{x}_i$, is augmented to the static feature to give the feature vector used for training, $\boldsymbol{c}_i$

$$\boldsymbol{c_i} = [\boldsymbol{x}_i, \Delta\boldsymbol{x}_i] \quad (3)$$

Currently, the HMMs are whole-word models taken from the GRID database which is described in more detail in Section 5 [18]. In practice to allow unconstrained input the HMMs could be phoneme-based which would allow any speech to be decoded.

3.1.2. HMM generation

Generation of the HMM-based time-frequency surface, $H(f,i)$, from an input time-frequency surface extracted from noisy speech, $Y(f,i)$, is a two-stage process. First the model and state sequence of the HMMs is determined and second this is used to output the most likely set of observation vectors.

The noisy time-frequency surface, $Y(f,i)$, is first transformed into a sequence of feature vectors, $\boldsymbol{C} = [\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_{I-1}]$, using equations 2 and 3. Using Viterbi decoding the most likely state sequence, $\hat{\boldsymbol{q}} = [\hat{q}_0, \hat{q}_1, \ldots, \hat{q}_{I-1}]$, given the set of noisy observation vectors, $\boldsymbol{C}$, is computed

$$\hat{\boldsymbol{q}} = \underset{q}{\operatorname{argmax}}\, P(\boldsymbol{q}|\boldsymbol{C}, \lambda) \tag{4}$$

For ease of notation, q_i implicitly provides the state and model at time frame i. From the state sequence and set of HMMs, the most likely sequence of static feature vectors is determined

$$\boldsymbol{X} = \underset{X}{\operatorname{argmax}}\, p(\boldsymbol{W}\boldsymbol{x}|\hat{\boldsymbol{q}}, \lambda) \tag{5}$$

where $\boldsymbol{X} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{I-1}]$ and $\boldsymbol{W}$ is a matrix that transforms the stream of static vectors into the augmented vectors [19]. The HMM synthesised time-frequency surface, $H(f,i)$, is created from each vector in $\boldsymbol{X}$

$$H(f,i) = \exp\left(x(f)_i\right) \quad \forall\ f \text{ and } i \tag{6}$$

$x(f)_i$ is the fth element of the ith log spectral vector, $\boldsymbol{x}_i$.

3.2. Local SNR estimation

An estimate of the localised time-frequency SNR, $SNR(f,i)$, is computed in decibels as

$$SNR(f,i) = 20\log_{10}\frac{GH(f,i)}{Y(f,i) - GH(f,i)} dB \tag{7}$$

A scaled HMM synthesised time-frequency surface, $GH(f,i)$, provides the clean speech part of the SNR calculation, while the noise component is computed by subtracting this from the noisy time-frequency surface, $Y(f.i)$. The scaling term, G, is found by minimising the difference between the noisy and HMM synthesised time-frequency surfaces over all time and frequency points in an utterance [20]

$$G = \underset{G}{\operatorname{argmin}} \sum_{i=0}^{I-1}\sum_{f=0}^{F-1} \left(Y(f,i) - GH(f,i)\right) \tag{8}$$

Whilst the SNR gives a measure of noise contamination in each time-frequency region it must be transformed before being used in equation 1. A perceptually motivated transformation of the SNR into a weighting function $\gamma(f,i)$, was found to be

$$\gamma(f,i) = \frac{1}{1 + e^{-\alpha(SNR(f,i)-\beta)}} \tag{9}$$

Parameters α and β have been estimated empirically and are currently set to $\alpha = 0.1$ and $\beta = 3.0$ to give $\gamma(f,i)$ close to one for high SNRs and close to zero for low SNRs.

4. Fundamental frequency and aperiodicity

Estimation of fundamental frequency has been the subject of much research with many successful algorithms reported [21, 22, 23, 24, 25]. Rather than developing a new method of fundamental frequency estimation, this work uses YIN to provide an fundamental frequency estimate, $\hat{f0}_i$, at each time instant [24]. STRAIGHT uses a measure of aperiodicity, $A(f,i)$, to determine the relative energy distribution of periodic to aperiodic components in the speech and this is calculated as the ratio between the energy of inharmonic to harmonic components in the speech spectrum [26].

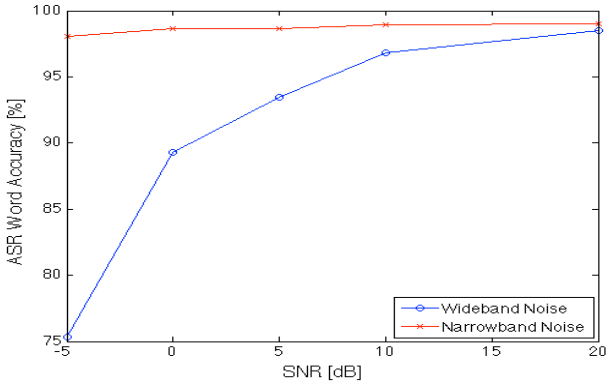

Figure 2: *Word accuracy in narrowband and wideband noise at SNRs from -5dB to +20dB.*

5. Experimental results and analysis

This section examines the effectiveness of the proposed reconstruction-based speech enhancement. Two noises have been selected for evaluation – narrowband noise and wideband noise. The narrowband noise is created by filtering white noise using four bandpass filters, centered at 500Hz, 1kHz, 2kHz and 3kHz, with bandwidths of 200Hz to give localised regions of low SNR. White noise forms the wideband noise and gives broad areas of contamination.

Tests first examine the accuracy of HMMs at decoding the speech which is subsequently used to provide the $H(f,i)$ contribution to the the time-frequency surface in equation 1. Secondly, the quality of the enhanced speech is compared with conventional filtering approaches using PESQ analysis across a range of SNRs. The speech used for these experiments is taken from four speakers in the GRID database, two male and two female, and is downsampled to 8kHz [18]. This has a rigid grammar where each utterance contains six words of the following structure *command→colour→preposition→letter→digit→adverb*. From the 1000 utterances from each speaker, 800 were used for training and the remainder for recognition and enhancement.

5.1. Narrowband noise

Figure 2 shows word accuracy in narrowband noise at SNRs from -5dB to +20dB for HMMs adapted to the noise conditions. The narrowband noise affects only a few spectral regions which allows recognition accuracy to remain very high across all SNRs. Even at -5dB, comparing the model and state sequence with that obtained using forced recognition on clean speech revealed little difference. This allows an accurate time-frequency surface, $H(f,i)$, to be synthesised. PESQ scores for the reconstructed speech are shown in figure 3 at SNRs from -5dB to +20dB. For comparison PESQ scores for no noise compensation, spectral subtraction, Wiener filtering and log MMSE

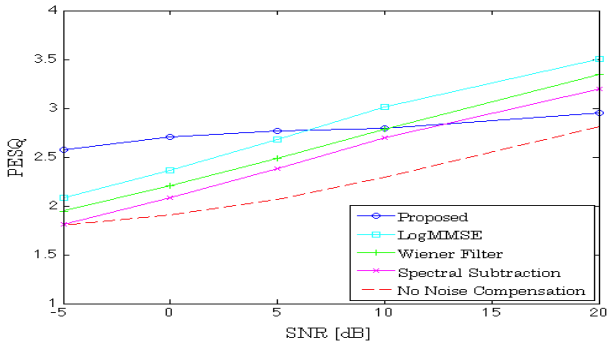

Figure 3: *Speech enhancement PESQ scores in narrowband noise at SNRs from -5dB to +20dB.*

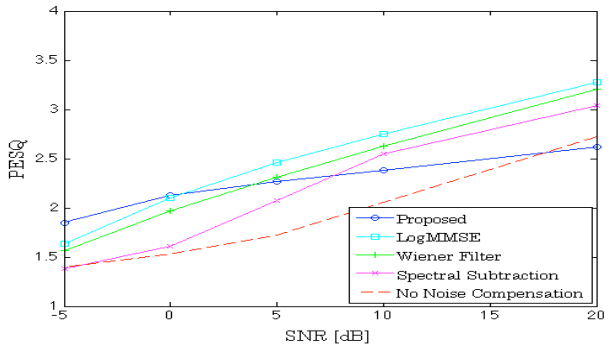

Figure 4: *Speech enhancement PESQ scores in wideband noise at SNRs from -5dB to +20dB.*

are also shown [1]. The results show that in clean conditions the reconstruction-based enhancement is less effective. However, as SNRs decrease reconstruction-based enhancement outperforms the conventional enhancement methods by utilising noise-free information from the HMMs.

5.2. Wideband noise

Figure 2 shows that word accuracy in wideband noise is significantly lower than in narrowband noise due to the widespread contamination introduced. Comparing the model and state sequences at lower SNRs to those obtained with forced recognition on clean speech showed that a significant number of decoding errors were introduced. Figure 4 shows PESQ analysis for reconstruction-based enhancement and conventional methods. This shows that the performance of all the enhancement methods has reduced in comparison to their performance in narrowband noise as the wideband noise has contaminated large regions of the time-frequency surface. At higher SNRs the quality of the reconstruction-based enhancement has deteriorated in comparison to the conventional methods, although at SNRs of 0dB and below it produces highest quality speech.

To give further insight into the enhancement process figure 5 shows time-frequency surfaces for the utterance *"place red at C 3 please"* spoken by a male speaker in white noise at an SNR of 0dB. The figure shows the original noisy speech, $Y(f,i)$, the HMM synthesised speech, $H(f,i)$, the weighted SNR, $\gamma(f,i)$, (light regions indicate high SNR and dark regions low SNR) and the enhanced speech, $\hat{X}(f,i)$.

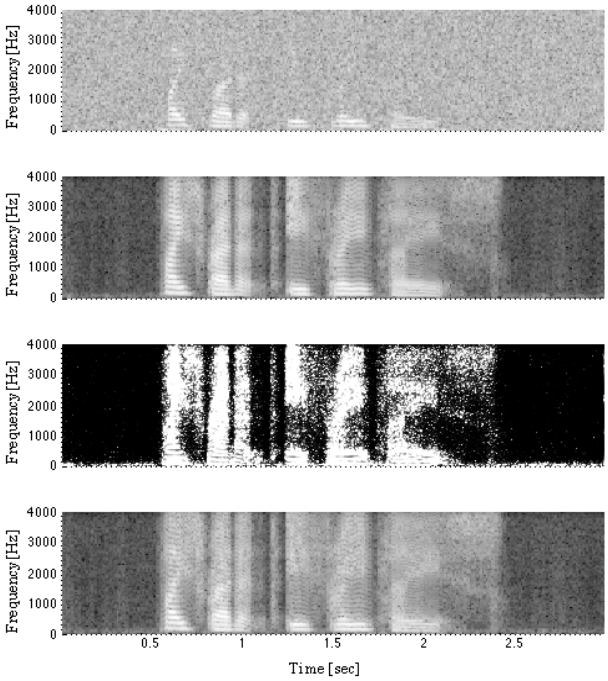

Figure 5: *Spectrograms of noisy speech, HMM synthesised speech, localised SNR and enhanced speech at an SNR of 0dB in white noise for the utterance "place red at C 3 please".*

5.3. Discussion

Analysis of the reconstruction-based enhanced speech shows it to be higher quality than conventional enhancement methods at low SNRs but worse at higher SNRs. The analysis also shows that the reconstruction-based enhancement is more effective in localised noise rather than in wideband noise where large areas of the time-frequency surface are contaminated. When the noise is localised, as with narrowband noise, large areas of the input noisy speech have high SNRs so only a small part of the time-frequency surface needs to be combined with the HMM synthesised signal. In wideband noise, the corruption is much more widespread meaning that large parts of the time-frequency surface are at low SNRs and require combination with the HMM synthesised signal.

6. Conclusions

This work has presented a first step in using HMMs to enhance a noisy speech signal. A framework has been developed that combines an HMM synthesised clean speech signal with the original noisy signal according to a perceptual weighting of the local SNR. This provides a time-frequency surface that when combined with fundamental frequency allows an enhanced speech signal to be reconstructed. Tests shows that the reconstruction-based enhancement improves the quality of speech and at lower SNRs outperforms conventional filtering methods.

The methods proposed and the analysis carried out have revealed many areas for improvement of this initial framework. In particular mapping of SNR to the perceptual weighting function needs refinement to optimise the information available in the noisy speech and HMM synthesised speech. This may also extend to exploiting the HMMs to provide fundamental frequency information in situations where estimates are erroneous.

7. References

[1] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, Inc., 2007.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] H.-T. Hu, F.-J. Kuo, and H.-J. Wang, "Supplementary schemes to spectral subtraction for speech enhancement," *Speech Communication*, vol. 36, no. 3, pp. 205–218, Mar. 2002.

[4] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *ICASSP*, vol. 4, Orlando, FL, May 2002.

[5] J. Lim and A. Oppenheim, "All pole modelling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, 1978.

[6] P. Scalart and J. Vieira-Filho, "Speech enhancement based on a priori signal to noise estimation," in *ICASSP*, vol. 2, Atlanta, GA, USA, May 1996, pp. 629–632.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[8] Y. Ephraim and H. Van-Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[9] M. Dendrinos, S. Bakamides, and G. Carayannis, "Speech enhancement from noise; a regenerative approach," *Speech Communication*, vol. 10, pp. 45–57, 1991.

[10] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *ICASSP*, France, 2006, pp. 153–156.

[11] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "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, pp. 187–207, Apr. 1999.

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

[13] 3G TS 26.071, "3rd generation partnership project (3GPP) TSG-SA codec working group mandatory speech codec speech processing functions amr speech codec; general description," 1999.

[14] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[15] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Eurospeech*, 1995, pp. 451–454.

[16] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the Blizzard Challenge 2007," in *Proc. Blizzard Challenge 2007*, Aug. 2007.

[17] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, Y. G. J. Tian, R. Hu, K. Oura, Y.-J. Wu, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for hmm-based speech synthesisanalysis and application of tts systems built on various asr corpora," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 984–1004, 2010.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *JASA*, vol. 150, no. 5, pp. 2421–2424, nov 2006.

[19] K. Tokuda, T. Yoshimura, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *ICASSP*, 2000, pp. 1315–1318.

[20] J. Ming, T. Hazen, and J. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modelling for speech separation," *Computer Speech and Language*, vol. 24, no. 1, pp. 67–76, Jan. 2010.

[21] M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio and Electroacoustics*, vol. 16, no. 2, pp. 262–266, 1968.

[22] L. Rabiner, M. Cheng, A. Rosenburg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 399–418, Oct. 1976.

[23] R. McAulay and T. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal representation," in *ICASSP*, vol. 1, Albuquerque, NM, USA, Apr. 1990, pp. 249–252, DOI: 10.1109/ICASSP.1990.115585.

[24] A. d. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[25] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm," ETSI STQ-Aurora DSR Working Group, ES 202 212 version 1.1.1, Nov. 2003.

[26] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.