INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

HMM-based speech enhancement using sub-word models and noise adaptation

Akihiro Kato and Ben Milner

University of East Anglia
akihiro.kato@uea.ac.uk, b.milner@uea.ac.uk

Abstract

This work proposes a method of speech enhancement that uses a network of HMMs to first decode noisy speech and to then synthesise a set of features that enables a clean speech signal to be reconstructed. Different choices of acoustic model (whole-word, monophone and triphone) and grammars (highly constrained to no constraints) are considered and the effects of introducing or relaxing acoustic and grammar constraints investigated. For robust operation in noisy conditions it is necessary for the HMMs to model noisy speech and consequently noise adaptation is investigated along with its effect on the reconstructed speech. Speech quality and intelligibility analysis find triphone models with no grammar, combined with noise adaptation, gives highest performance that outperforms conventional methods of enhancement at low signal-to-noise ratios.
Index Terms: speech enhancement, HMMs, STRAIGHT, noise adaptation

1. Introduction

The aim of this work is to use hidden Markov models (HMMs) for speech enhancement. HMMs have been very effective in decoding clean and noisy speech into word or phoneme sequences and more recently have been applied successfully to statistical speech synthesis [1]. HMM-based speech enhancement combines these technologies by first decoding noisy speech using a network of HMMs and then, using the same network of HMMs, synthesises clean speech.

Historically, most approaches to speech enhancement use filtering methods that include spectral subtraction, Wiener filtering, statistical and subspace methods [2, 3, 4, 5]. More recently, several approaches have been proposed that instead synthesise or reconstruct a clean speech signal. For example, corpus and inventory methods use noisy speech to identify segments from a database of clean speech which are concatenated to form the enhanced speech signal [6, 7]. Methods that utilise HMMs within the enhancement process fall into both the filtering and reconstruction approaches. Several filtering methods have combined clean speech HMMs and noise HMMs to model noisy speech, with the HMMs subsequently providing speech and noise features that are used to construct filters (e.g. Wiener filters) to enhance the noisy speech [8, 9, 10, 11]. Conversely, reconstruction methods have employed HMM synthesis techniques to synthesise an estimate of the clean speech given a model and state sequence estimated from noisy speech [12, 13].

The work proposed in this paper uses a model of speech production to reconstruct clean speech from a set of parameters obtained from a network of HMMs. Our previous work used whole-word HMMs with a word-level grammar to decode noisy sentences and then reconstructed noise-free sentences, but was constrained to speech conforming to the grammar [13]. The technique is now generalised by using sub-word HMMs and an unconstrained grammar, which removes any dictionary and grammar constraints to enable unconstrained speech input as would be required in a practical deployment. Investigation is made in terms of imposing and relaxing different acoustic model and grammar constraints and examining their effect on speech quality and intelligibility. Furthermore, noise adaptation is also now included and used to adjust the statistics of the clean-trained HMMs to model noisy speech and thereby improve decoding accuracy and the resulting speech signal.

An overview of the proposed HMM-based speech enhancement is given in Section 2. Section 3 describes how a sequence of clean feature vectors is synthesised from a network of HMMs from a noisy input signal and how noise adaptation is applied. Section 4 explains how the parameters needed for speech reconstruction are then obtained from the feature vectors. Experiments are presented in Section 5 that compare the quality and intelligibility of HMM enhancement with conventional methods in white and babble noises.

2. Speech enhancement framework

Noisy speech is first decoded by a network of HMMs that is adapted to the current noise conditions to give a model and state sequence. Using this sequence, the HMMs then output a set of clean observation vectors which are input into a model of speech production to reconstruct enhanced speech. This section describes the speech production model and feature extraction.

2.1. STRAIGHT vocoder

The STRAIGHT vocoder is used for speech reconstruction given its success in HMM-based speech synthesis [1, 14], and requires three input parameters: i) a time-frequency surface, $X(f, i)$, ii) a fundamental frequency contour, f_{0_i}, and iii) a measure of aperiodicity, $A(f, i)$, where f and i represent frequency bin and frame indices, respectively. The challenge for HMM speech enhancement is to estimate these parameters accurately to reconstruct good quality speech.

2.2. Feature extraction

The same HMMs are used for decoding and synthesis, so the speech features must be sufficiently discriminative to provide accurate decoding and also be able to provide the parameters needed for speech reconstruction. To address both criteria the features are based largely around the requirements for STRAIGHT. Frames of speech are extracted at 5ms intervals with a variable duration of $1.2 \times T_0$ for voiced speech, where T_0 is the fundamental period, and a duration of 2.5ms for unvoiced speech. A 1024-point FFT is applied and the resulting power spectrum input into a 23-channel mel filterbank followed by a log and discrete cosine transform (DCT) to produce a 23-D MFCC vector, $\boldsymbol{x}_i$, with no truncation. From the magnitude

http://dx.doi.org/10.21437/Interspeech.2016-928

spectrum the aperiodicity is computed as the ratio between the energy of inharmonic to harmonic components and gives a measure of the relative energy distribution of periodic to aperiodic components. This is input into a 23-channel mel filterbank to give an aperiodicity vector, $\boldsymbol{a}_i$. An estimate of fundamental frequency, f_{0_i}, is computed using PEFAC, which is highly robust at low SNRs [15]. These form static feature vector, $\boldsymbol{c}_i$, with three streams,

$$\boldsymbol{c}_i = [\boldsymbol{x}_i, \boldsymbol{a}_i, \log f_{0_i}] \tag{1}$$

For unvoiced frames, $\log f_{0_i}$ is set to zero.

3. HMM decoding and synthesis

In HMM synthesis a word sequence is used to generate a feature vector stream [16, 17]. Application to speech enhancement is different as no word sequence is available and instead the HMMs must decode the noisy speech into a model and state sequence which is input into the HMMs to generate feature vectors, $\hat{\boldsymbol{c}}_i$. These are transformed into the parameters needed for speech reconstruction.

3.1. HMM training

Static feature vectors, $\boldsymbol{c}_i$, are defined in Eq. 1 and to improve decoding accuracy, and the smoothness of the synthesised feature vectors, a velocity derivative, $\Delta\boldsymbol{c}_i$, is augmented to give the feature vector, $\boldsymbol{o}_i$, used for training

$$\boldsymbol{o}_i = [\boldsymbol{c}_i, \Delta\boldsymbol{c}_i] \tag{2}$$

To incorporate velocity derivatives in the HMM synthesis stage, the set of feature vectors for the entire utterance, $\boldsymbol{O} = [\boldsymbol{o}_0, \boldsymbol{o}_1, \ldots, \boldsymbol{o}_{N-1}]$ is computed from the set of static vectors, $\boldsymbol{C} = [\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N-1}]$

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C} \tag{3}$$

where matrix $\boldsymbol{W}$ contains the regression coefficients to transform the static vectors into the augmented vectors [18].

The acoustic units modelled by the set of HMMs, Λ, can take different forms. In earlier work HMMs were trained on whole words which limited the vocabulary and prevented unconstrained speech input [13]. Shorter duration acoustic units are now considered to allow unconstrained speech input. Monophone HMMs are trained first and then extended to cross-word triphone HMMs. Decision tree clustering was used to restrict the number of triphones to 678 triphones [19]. Whole word models use 16 states, while 5 state HMMs are used for monophone and triphone sub-word models.

3.2. HMM decoding

Estimating a model and state sequence from input noisy speech uses only the MFCC component, $\boldsymbol{x}_i$ of the feature vector, $\boldsymbol{o}_i$, as including aperiodicity and fundamental frequency reduced accuracy. The sequence of noisy MFCC vectors, $\boldsymbol{X} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1}]$, is input into the network of HMMs, Λ, and using Viterbi decoding a state and model sequence, $\hat{\boldsymbol{q}} = [\hat{q}_0, \hat{q}_1, \ldots, \hat{q}_{N-1}]$, is computed

$$\hat{\boldsymbol{q}} = \underset{q}{\operatorname{argmax}}\, p(\boldsymbol{q}|\boldsymbol{X}, \Lambda, G) \tag{4}$$

where G is the grammar. For ease of notation, $\hat{q}_i$, provides the model and state at time i. Depending on the acoustic units being modelled by the HMMs, a grammar, G, can be applied to constrain the decoding and is examined in Section 5.

3.3. Noise adaptation

To improve the robustness of the HMMs when decoding noisy speech, adaptation is applied to adjust the MFCC components of the clean-trained HMMs to model noisy speech. Specifically a mismatch function, $g(.)$, transforms clean speech and noise MFCC vectors, $\boldsymbol{x}$ and $\boldsymbol{d}$, into a noisy MFCC vector, $\boldsymbol{y}$,

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{y}^l = \boldsymbol{C}g(\boldsymbol{C}^{-1}\boldsymbol{x}, \boldsymbol{C}^{-1}\boldsymbol{d}, \boldsymbol{\beta}) \tag{5}$$

where $\boldsymbol{C}$ is a DCT matrix. The mismatch function is defined

$$g(\boldsymbol{x}^l, \boldsymbol{d}^l, \boldsymbol{\beta}) = \boldsymbol{x}^l + \log\left(1 + \exp^{\boldsymbol{d}^l - \boldsymbol{x}^l} + 2\boldsymbol{\beta}\sqrt{\exp^{\boldsymbol{d}^l - \boldsymbol{x}^l}}\right) \tag{6}$$

where the superscript l denotes a log filterbank vector. Vector $\boldsymbol{\beta}$ represents a log filterbank-domain phase component that has been shown to improve adaptation accuracy and is defined in [20]. An explicit value of $\boldsymbol{\beta}$ is not known, however following [20], an estimate is made using a lookup table that is computed offline during a training stage. For a given $\boldsymbol{x}$ and $\boldsymbol{d}$, the lookup table outputs a phase averaged estimate of $\boldsymbol{\beta}$ that is used in Eq. 6. If the phase component is ignored, i.e. $\boldsymbol{\beta} = [\boldsymbol{0}]$, the mismatch function becomes the conventional phase-independent mismatch function. The mismatch function is applied to the means and variances of each state of the clean speech trained HMMs, Λ, to adapt them to model noisy speech. This new set of HMMs, Λ' is then used in the decoding of Equation 4. To obtain the noise statistics needed for adaptation, the method of unbiased MMSE estimation was used [21].

3.4. HMM synthesis

Using techniques from HMM synthesis [1], given the state and model sequence, $\hat{\boldsymbol{q}}$, and clean-trained HMMs, Λ, the most likely sequence of static feature vectors, $\hat{\boldsymbol{C}} = [\hat{\boldsymbol{c}}_0, \hat{\boldsymbol{c}}_1, \ldots, \hat{\boldsymbol{c}}_{N-1}]$, is

$$\hat{\boldsymbol{C}} = \underset{C}{\operatorname{argmax}}\, p(\boldsymbol{W}\boldsymbol{C}|\hat{\boldsymbol{q}}, \Lambda) \tag{7}$$

These feature vectors can now be transformed into the parameters needed by STRAIGHT for speech reconstruction.

4. Extraction of STRAIGHT parameters

From the sequence of synthesised static vectors, $\hat{\boldsymbol{C}}$, the time-frequency surface, $\hat{X}(f,i)$, aperiodicity, $\hat{A}(f,i)$, and fundamental frequency, $\hat{f}_{0_i}$, needed by STRAIGHT can be extracted.

The time-frequency surface, $\hat{X}(f,i)$, is obtained by first equalising each MFCC vector in $\hat{\boldsymbol{C}}$, for spectral tilt introduced in feature extraction, and then applying an inverse DCT and exponential to give filterbank features [22]. Cubic spline interpolation then creates a 513-point spectral representation which forms the time-frequency surface. Similarly, aperiodicity vectors from $\hat{\boldsymbol{C}}$ are inverted to form the aperiodicity, $\hat{A}(f,i)$.

Two methods to obtain fundamental frequency, $\hat{f}_{0_i}$, are considered. The first uses the stream of fundamental frequency estimates in $\hat{\boldsymbol{C}}$ generated by the HMMs. Whilst this is aligned with the time-frequency surface there is no guarantee that it is an accurate representation of the original f_0 values. For example, specific intonation introduced by the speaker is not reproduced in the contour from the HMMs. This leads to the second approach which estimates fundamental frequency from the noisy speech using PEFAC [15]. Preliminary tests compared the two approaches and found the second produced a more accurate f_0 contour and a more representative speech signal. Consequently, PEFAC is used to provide $\hat{f}_{0_i}$ in testing.

5. Experimental results and analysis

Experiments examine speech quality and intelligibility using different acoustic model and grammar configurations, first in clean conditions and then noisy conditions, where noise adaptation is investigated. Experiments use speech from four speakers in the GRID database (two male and two female) [23]. Sentences conform to a structure of *command→colour→preposition→letter→digit→adverb*. For each speaker, 800 sentences are used for training and 200 for testing. Six combinations of acoustic model (whole word, monophone and triphone) and grammar (word grammar and no grammar) are considered as shown in Table 1.

Table 1: Acoustic model and grammar configurations.

Method	Acoustic model	Word grammar
WORD_N	Word	None
WORD_G	Word	GRID
MONO_N	Monophone	None
MONO_G	Monophone	GRID
TRI_N	Triphone	None
TRI_G	Triphone	GRID

Some configurations are inappropriate for practical scenarios due to grammar constraints or using whole-word models (e.g. WORD_G, WORD_N, MONO_G and TRI_G). However, they provide useful analysis while configurations with no grammar and using sub-word models are able to be deployed practically (e.g. MONO_N and TRI_N).

5.1. Tests in clean speech

A first set of tests examined the quality and intelligibility when clean speech is input into HMM-based enhancement to gauge baseline performance and is presented in Table 2. The first column of results (RAW) shows performance when feature vectors, $\boldsymbol{c}_i$, are extracted from clean speech and input directly into STRAIGHT with no HMMs involved. The quality (PESQ) score of 3.52, in comparison to 4.50 for the original speech, shows a reduction in quality arising from the STRAIGHT synthesis which is similar to other studies [24]. Intelligibility, measured by the normalised covariance metric (NCM [25]), reduces slightly to 0.98 from 1.00 with original speech.

The remaining columns in Table 2 show, for the combinations in Table 1, decoding accuracy, speech quality and intelligibility when clean speech is input into HMM enhancement and provide a baseline on performance. Extension _F shows results with forced alignment and consequently decoding accuracy is 100% which gives highest quality and intelligibility. Using a word grammar (_G) gives very high decoding accuracies across all three acoustic models. For word models, decoding accuracy of 99.92% gives quality and intelligibility equal to forced alignment. For triphone and monophone models, although decoding accuracy is still very high (>97%), quality and intelligibility are lower than with forced alignment. This is due to the timings of the decoded phoneme sequence being different to that of the original speech which gives lower objective scores. When no grammars are used (WORD_N, TRI_N and MONO_N) decoding accuracy reduces but this doesn't effect the quality and intelligibility. This is an important result which shows that quality and intelligibility are robust to grammar constraints being removed which is necessary for unconstrained input in practical scenarios. The remaining analysis considers the performance of HMM-based speech enhancement in noisy conditions.

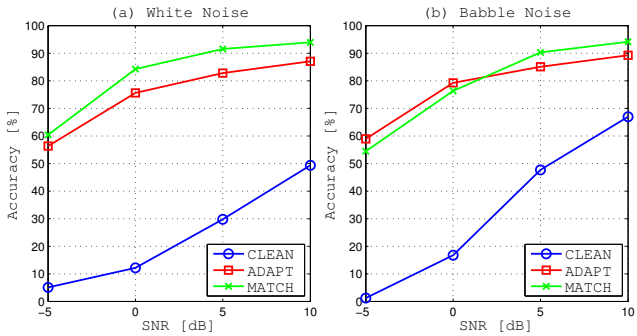

Figure 1: *HMM decoding with clean models, matched models and the noise adaptation in a) white noise and b) babble noise.*

5.2. Decoding accuracy

The first test examines how decoding accuracy in noisy conditions can be maximised using the noise adaptation methods described in Section 3.3. Figures 1a and 1b show decoding accuracy in white noise and babble noise at SNRs from -5dB to +10dB using the original clean trained HMMs (CLEAN), models adapted to noise (ADAPT) and models trained under matched noise conditions (MATCH) which provide an indication of the upper bound on accuracy. For all tests, triphone HMMs are used with no word grammar, i.e. TRI_N. Decoding accuracy using the clean models deteriorates rapidly as SNRs reduce. However, applying noise adaptation improves accuracy substantially with performance close, and in some cases exceeding, matched condition models. In all subsequent tests noise adaptation is applied to the clean models.

A second set of tests is now performed to examine the effect that the model and grammar combinations of Table 1 have on decoding accuracy. Figures 2a and 2b show decoding accuracy in white and babble noise at SNRs from -5dB to 10dB for the different model and grammar configurations (WORD_N had similar performance to WORD_G and is removed for clarity). The WORD_G and TRI_G methods have highest decoding accuracy as they both use a word grammar. Removing the grammar from the triphone system (TRI_N) reduces decoding accuracy but has the important advantage of now being able to decode unconstrained speech input which is necessary in a practical enhancement scenario. The two monophone configurations perform worst.

5.3. Speech quality

Figures 2c and 2d show PESQ scores for HMM enhancement where word-level HMMs (WORD_G) and triphones (TRI_G and TRI_N) attain highest quality. This is due to their high decoding accuracy and the good synthesis quality when using whole-word or triphone models. For the triphone system, removing the grammar constraint (and consequently allowing a practical implementation) has almost no effect on quality even though it had a larger difference in decoding accuracy. This is attributed to the decoding error metric reporting an error irrespective of how acoustically similar the incorrectly chosen model is to the correct model. In many instances the erroneous model is still acoustically similar and so has much less effect on the resulting speech quality than the decoding error rate may suggest. This is particularly true with the large number of triphone models for each phoneme. The two monophone-based systems perform worst and this is expected due to their lower

Table 2: Decoding accuracy, speech quality (PESQ) and intelligibility (NCM) for clean speech input into HMM-based enhancement using different acoustic models and grammar constraints, and for direct synthesis (RAW).

	RAW	WORD_F	WORD_G	WORD_N	TRI_F	TRI_G	TRI_N	MONO_F	MONO_G	MONO_N
Acc. %	-	100	99.92	98.58	100	99.79	98.83	100	97.78	74.90
PESQ	3.52	2.63	2.63	2.63	2.74	2.44	2.44	2.41	2.20	2.22
NCM	0.98	0.77	0.77	0.77	0.80	0.73	0.73	0.70	0.65	0.64

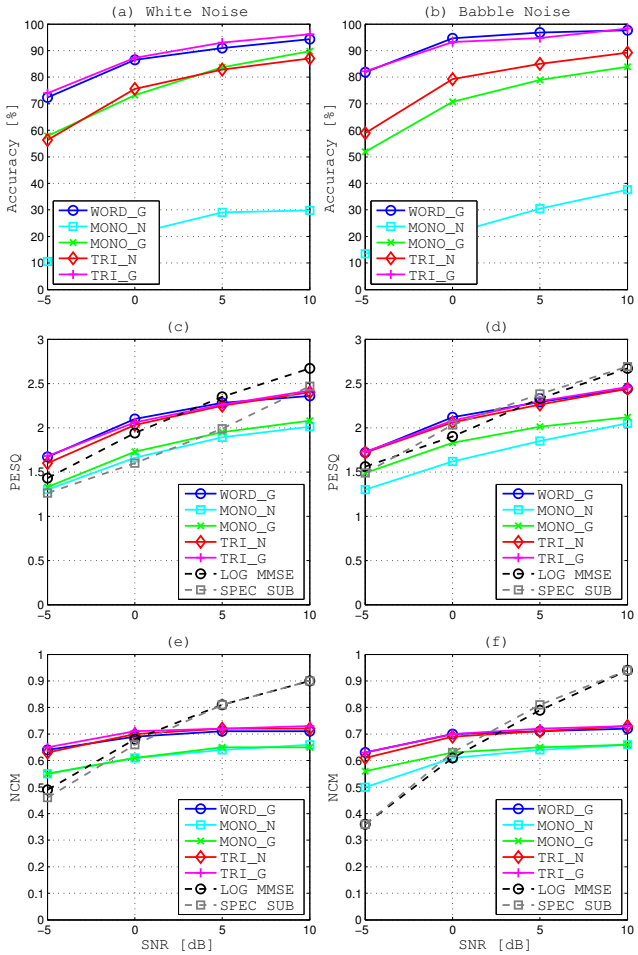

Figure 2: *HMM enhancement in white noise (left column) and babble noise (right column) showing: a) & b) decoding accuracy, c) & d) speech quality (PESQ), e) & f) speech intelligibility (NCM).*

decoding accuracy and the lack of context in synthesis. Also shown for comparison are results for the log MMSE and spectral subtraction methods of enhancement [2, 4], which perform better at higher SNRs but fall below the PESQ scores of the HMM-based enhancement at lower SNRs.

To investigate further the characteristics of the enhanced speech the source-to-interference ratio (SIR) and source-to-distortion ratio (SDR) are shown in Figures 3a and 3b for white and babble noises, and compared with log MMSE [26]. The SIR shows HMM enhancement to be more effective at removing interfering noise than log MMSE and is attributed to the HMMs/STRAIGHT reconstructing noise-free speech. Conversely, the SDR is lower compared to log MMSE and is attributed to the more artificial speech quality produced by

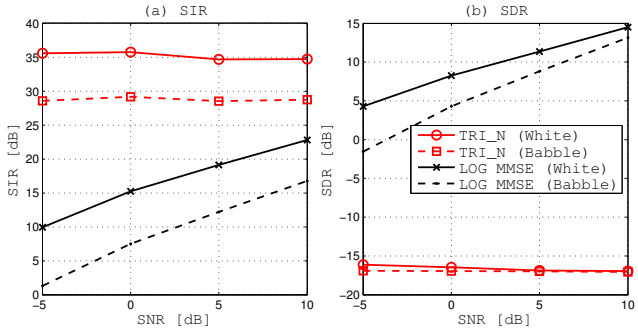

Figure 3: *a) SIR and b) SDR of HMM enhanced speech and log MMSE in white and babble noises at SNRs from -5dB to 10dB.*

STRAIGHT. Listening to the speech confirms these results.

5.4. Speech intelligibility

Figures 2e and 2f show speech intelligibility (NCM). Triphone models (TRI_N and TRI_G) and word models (WORD_G) attain highest intelligibility which remains very stable even at low SNRs. The slight reduction at -5dB is attributed to the reduction in decoding accuracy that is observed in these high levels of noise. Intelligibility of the conventional methods falls more rapidly and at SNRs of 0dB and below the HMM enhancement achieves higher intelligibility.

6. Conclusions

This work has presented a method of speech enhancement that uses HMMs to first decode input noisy speech and then synthesise parameters to reconstruct clean speech. Analysis has shown that triphone-based systems maintain performance without the need for a word grammar which enables a practical system for enhancing unconstrained speech. Furthermore, quality and intelligibility are found to not be too sensitive to decoding errors as often a similarly sounding acoustic model is selected. To bring decoding accuracy in noise to a level capable of providing good synthesis of parameters, noise adaptation was found to be as effective as matched training which again enables a practical deployment. In comparison to log MMSE, the HMM enhanced speech was in general found to be free from background noise but more distorted. HMM enhancement was found to have higher quality and intelligibility at low SNRs, and remains more stable as SNRs reduce. However, tests at higher SNRs and in clean conditions show that quality and intelligibility are restricted, compared to the original speech, which puts an upper limit on performance. Further work is concentrated on improving synthesis quality in clean conditions which should improve quality in noise.

7. References

[1] 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.

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

[3] N. Hadir, F. Faubel, and D. Klakow, "A model-based spectral envelope Wiener filter for perceptually motivated speech enhancement," in *Interspeech*, 2011.

[4] 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.

[5] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," vol. 11, no. 4, pp. 334–341, July 2003.

[6] X. Xiao and R.M. Nickel, "Speech enhancement with inventory style speech resynthesis," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1243–1257, Aug. 2010.

[7] J. Ming and D. Crookes, "Speech enhancement from additive noise and channel distortion - a corpus-based approach," in *Interspeech*, Sept. 2014, pp. 2710–2714.

[8] M. Nilsson, M. Dahl, and I. Claesson, "HMM-based speech enhancement applied in non-stationary noise using cepstral features and log-normal approximation," *in Proc. DSPCS*, pp. 82 – 86, 2003.

[9] D.Y. Zhao and W.B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 882 – 892, 2007.

[10] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, no. 12, pp. 1846–1856, Dec. 1989.

[11] H. Veisi and H. Sameti, "Cepstral-domain HMM-based speech enhancement using vector Taylor series and parallel model combination," in *Proc. ISSPA*, 2012, pp. 298–303.

[12] J.L. Carmona, J. Barker, A.M. Gomez, and N. Ma, "Speech spectral envelope enhancement by hmm-based analysis/resynthesis," *IEEE Signal Processing Letters*, vol. 20, no. 6, pp. 563–566, 2013.

[13] A. Kato and B.P. Milner, "Using hidden Markov models for speech enhancement," in *Interspeech*, 2014.

[14] 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, pp. 187–207, Apr. 1999.

[15] S. Gonzalez and M. Brookes, "PEFAC - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

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

[17] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, Y. Guan 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, July 2010.

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

[19] K. Shinoda and T. Watanabe, "MDL-based context-dependent sub word modelling for speech recognition," *J. Acoust. Soc. Jpn.*, vol. 21, no. 2, pp. 79–86, 2000.

[20] F. Faubel, J. McDonough, and D. Klakow, "A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-mel domain," in *Interspeech*, 2008, pp. 553–556.

[21] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[22] X. Shao and B. Milner, "Predicting fundamental frequency from mel-frequency cepstral coefficients to enable speech reconstruction," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1134–1143, Aug. 2005, DOI: 10.1121/1.1953269.

[23] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 150, no. 5, pp. 2421–2424, Nov. 2006.

[24] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," *in 8th ISCA Speech Synthesis Workshop*, pp. 135 – 140, 2013.

[25] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[26] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462 – 1469, 2006.