TWIN-HMM-BASED AUDIO-VISUAL SPEECH ENHANCEMENT

Ahmed Hussen Abdelaziz, Steffen Zeiler, Dorothea Kolossa*

Insititute of Communication Acoustics, Digital Signal Processing Group,
Ruhr-Universität Bochum, 44801 Bochum
{Ahmed.HussenAbdelAziz, Steffen.Zeiler, Dorothea.Kolossa}@rub.de

ABSTRACT

Most approaches for speech signal processing rely solely on acoustic input, which has the consequence that spectrum estimation becomes exceedingly difficult when the signal-to-noise ratio drops to values near 0 dB. However, alternative sources of information are becoming widely available with increasing use of multimedia data in everyday communication. In the following paper, we suggest to use video input as an auxiliary modality for speech processing by applying a new statistical model – the twin hidden Markov model. The resulting enhancement algorithm for audiovisual data greatly outperforms the standard audio-only log-MMSE estimator on all considered instrumental speech quality measures covering spectral and perceptual quality.

Index Terms— Multimodal speech processing, audiovisual speech recognition, state-based speech enhancement

1. INTRODUCTION

Speech enhancement has been using HMM speech models for a long time [1, 2, 3]. The idea of applying speech recognition HMMs for this purpose is appealing due to their rich information sources, which may be tapped to further improve speech enhancement. However, direct implementations of this idea are confronted with two major problems:

Firstly, a good feature space for speech recognition is not typically appropriate for speech processing. While suitable recognition features are characterized by their mutual decorrelation, low dimensionality, and speaker independence, the speech processing features, in contrast, need to have a sufficiently high resolution to yield detailed speech spectrum estimates - a prerequisite that almost directly implies high dimensionality, inter-feature correlations, and thus, a model that describes pitch and other speaker characteristics unrelated to phonetic information.

Secondly, degraded audio data also makes the intermediate step of ASR increasingly difficult, which means that good estimates of the underlying ASR state sequence are extremely hard to obtain by principle.

*This work has been supported by the Ministry of Economic Affairs and Energy of the State of North Rhine- Westphalia, Grant IV.5-43-02/2-005-WFBO-009.

1.1. The Twin HMM

To counteract the first of these difficulties, we suggest the use of a statistical model apt to describe the co-evolution of two streams of data – one of them suitable for recognition and the other for speech processing. We term this model *twin HMM*, to emphasize the reliance on two output distributions for each HMM state.

In this model, the ASR recognition features can be chosen solely for maximum phonetic discriminance. The synthesis features need to have the same temporal evolution as the recognition features, but they do not require the same discriminance properties. Rather, they should contain all information needed to reconstruct (synthesize) speech signals.

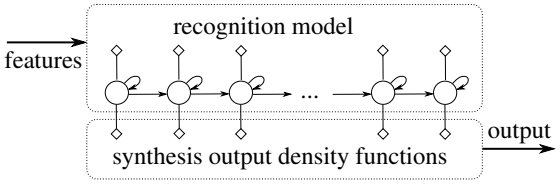

Fig. 1. Concept of twin-HMM for model-based speech enhancement

Figure 1 depicts the core idea of this concept, namely, having *one* underlying state sequence with *two* associated observation models – one for recognition purposes and one dedicated to speech synthesis.

1.2. Audiovisual Speech Processing

The second problem of ASR-based speech processing – degrading performance at decreasing SNRs – can be counteracted by using *audiovisual* data for the recognition phase. Since the video features are independent of the acoustical environment, this helps to obtain maximally reliable state sequences for the final synthesis step.

In the following, we will describe our new framework, comprising both optimizations – twin-HMMs for modeling the co-evolution of recognition and synthesis features, and audiovisual ASR – in Sec. 2. Experiments and results will be shown for a small-vocabulary test setup in Sec. 3 and compared to related state-of-the-art techniques in Sec. 4.

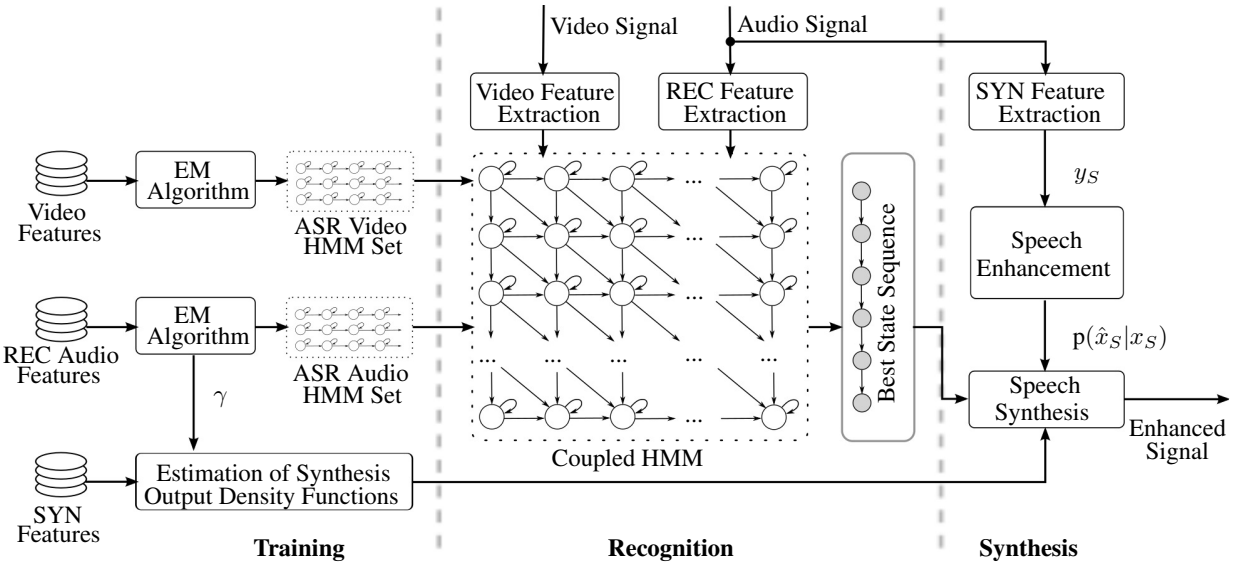

Fig. 2. Framework.

2. FRAMEWORK DESCRIPTION

Speech processing using twin HMMs consists of three main phases, the training, recognition, and synthesis phase, as shown in Figure 2.

2.1. Training

In the training phase, two statistical models are learned, firstly, a video-only HMM set for the visual modality, and secondly a twin HMM set for audio recognition (REC) and synthesis (SYN) features. For video data, the training is performed by the EM algorithm [4]. For audio data, training of the REC features also takes place using the standard EM algorithm, but we additionally store the state occupation probabilities γ from the final expectation step of the EM algorithm.

Estimating the parameters of the SYN distribution is done by applying the reestimation formulas of the maximization step in the EM algorithm to the SYN features, while using the stored state occupation probabilities γ to weight their contributions to all states in the model.

2.2. Recognition

To enhance a noisy audio signal, it is first transcribed. For this purpose, the video features and the twin-HMM REC features are fed to a coupled HMM (CHMM) [5] recognizer. CHMMs are dynamic Bayesian networks that describe the co-evolution of the audio and video features over time, while maintaining their natural correlation. The composition of the CHMMs from the trained ASR audio and video models is performed as described in [6], using only the REC output distributions for the audio part of the CHMM.

Using these CHMMs, it is possible to perform audiovisual speech recognition, for which we use the JASPER system [7]. The hypothesized words are then used to find the best state sequence in the twin HMM, by applying the forward-backward algorithm to the REC features word-by-word and choosing the state with maximal posterior probability at each frame. This two-step procedure of Viterbi recognition followed by forward-backward state probability computation is necessary to avoid excessive computational complexity in the fusion of audio and video information.

2.3. Synthesis

The synthesis phase is composed of three main blocks, namely, SYN feature extraction, speech enhancement and speech synthesis.

After extracting the SYN feature vectors $\mathbf{y}_{S_t}$ from the distorted signal, as shown in Figure 2, a Bayesian speech enhancement algorithm is used to find an unbiased estimate $\hat{\mathbf{x}}_{S_t}$ of the clean SYN feature vector $\mathbf{x}_{S_t}$, with t denoting the time frame index. Since enhancement algorithms cannot perfectly reconstruct the clean feature vector from the noisy data, the enhanced feature vector usually contains residual noise and estimation errors $\mathbf{e}_t$. Thus, we can write

$$\hat{\mathbf{x}}_{S_t} = \mathbf{x}_{S_t} + \mathbf{e}_t. \tag{1}$$

Assuming that the clean feature vector $\mathbf{x}_{S_t}$ in (1) is deterministic and the error vector $\mathbf{e}_t$ is zero-mean Gaussian, i.e. $\mathbf{e}_t \sim \mathcal{N}(\mathbf{e}_t; 0, \mathbf{\Sigma}_{\mathbf{e}_t})$, the estimated feature vector $\hat{\mathbf{x}}_{S_t}$ can also be modeled as Gaussian

$$\mathrm{p}(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t}) = \mathcal{N}(\hat{\mathbf{x}}_{S_t}; \mathbf{x}_{S_t}, \mathbf{\Sigma}_{\mathbf{e}_t}) \tag{2}$$

with mean $\mathbf{x}_t$ and diagonal covariance matrix $\mathbf{\Sigma}_{\mathbf{e}_t}$.

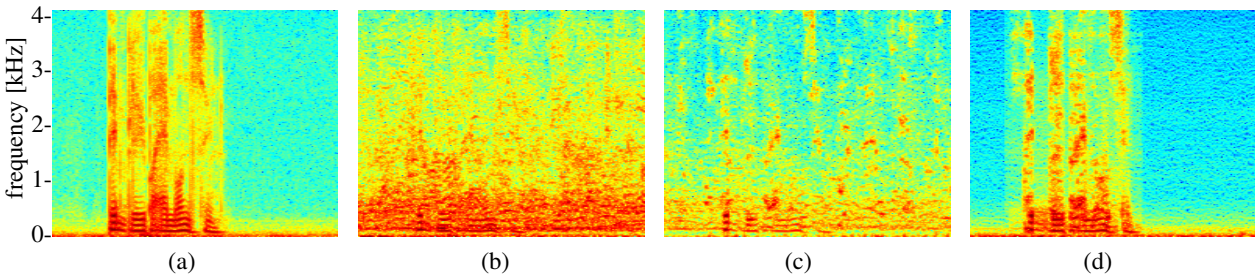

Fig. 3. (a) Spectrum of the three second long GRID sentence "BIN BLUE BY M ONE SOON" uttered in clean conditions. (b) The same sentence with added babble noise at 0 dB SNR. (c) The log-MMSE enhanced spectrum. (d) The filtered noisy spectrum after twin-model-based speech enhancement.

Since this full distribution model is needed in the speech synthesis block, the function of the speech enhancement block is not just to enhance the speech signal but rather to estimate its probability density function (PDF) $\mathrm{p}(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t})$.

In the speech synthesis block, the synthesis output distribution from the training phase, the best state sequence obtained in the recognition phase, and the above PDFs $\mathrm{p}(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t})$ are used to synthesize a new enhanced signal. To do so, we assume that the hidden clean feature vector $\mathbf{x}_{S_t}$ is generated from the state $q(t) = i$ at time t according to the synthesis output distribution $\mathrm{p}\left(\mathbf{x}_{S_t}|q(t) = i\right)$, and that the enhanced feature vector $\hat{\mathbf{x}}_{S_t}$ is generated from the corresponding clean feature vector $\mathbf{x}_{S_t}$ according to the PDF $\mathrm{p}(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t})$.

Using these assumptions, we propose the conditional expectation $E\left[\mathbf{x}_{S_t}|\hat{\mathbf{x}}_{S_t}, q(t) = i\right]$ of the clean feature vector $\mathbf{x}_{S_t}$ given the parameters of state i and the enhanced feature vector $\hat{\mathbf{x}}_{S_t}$ as an estimate for the hidden clean feature vector. We compute this conditional expectation as the mean of the conditional PDF $\mathrm{p}\left(\mathbf{x}_{S_t}|\hat{\mathbf{x}}_{S_t}, q(t) = i\right)$. For the sake of simplicity, we have merged the mixture components of the synthesis output distributions to a single Gaussian with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Since the PDF $\mathrm{p}(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t})$ is Gaussian as well, the needed posterior distribution becomes [8]

$$\mathrm{p}\left(\mathbf{x}_{S_t}|\hat{\mathbf{x}}_{S_t}, q(t) = i\right) = \mathcal{N}\left(\mathbf{x}_{S_t}; \tilde{\boldsymbol{\mu}}_{i,t}, \tilde{\boldsymbol{\Sigma}}_{i,t}\right) \qquad (3)$$

with mean vector

$$\tilde{\boldsymbol{\mu}}_{i,t} = \boldsymbol{\Sigma}_{e_t}\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_{e_t}\right)^{-1}\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_{e_t}\right)^{-1}\hat{\mathbf{x}}_{S_t}. \quad (4)$$

The mean vectors $\tilde{\boldsymbol{\mu}}_{i,t}$ are then used as estimates of the hidden clean SYN features, from which the speech signal are finally synthesized via the overlap-add method.

From (4), it can be noticed that $\tilde{\boldsymbol{\mu}}_{i,t}$ is a weighted sum of the mean vector of the SYN output distribution of the i^{th} state, and the mean vector of the PDF $\mathrm{p}\left(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t}\right)$, which is generated by the speech enhancement block. The weights depend on the covariance matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_{e_t}$, which are the quantitative representation of the uncertainty of each hypothesis.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

Two male and two female speakers are chosen from the Grid audio-visual database [9] to evaluate the presented framework. The speech data of each speaker are divided into a training and a test set. The test sets of all speakers are artificially distorted with three types of noise: white noise, buccaneer jet cockpit noise and speech babble. The noise signals, taken from the NOISEX database [10], are added to the speech data at a signal-to-noise ratio (SNR) in the range from 15 dB down to 0 dB in accordance with ITU-T P.56 [11].

3.2. Experimental setup

All speech signals are downsampled from $f_s = 25$ kHz, the original sampling frequency of the Grid database, to 8 kHz.

The REC features are required to work well in ASR systems, and should be chosen for independence, saliency and acceptable robustness against noise and reverberation. As such properties are available in mel-frequency cepstral coefficients (MFCCs) [12], we are using the first 13 static MFCCs extracted by the ETSI advanced front-end (AFE) [13] and the 26 corresponding Δ and $\Delta\Delta$ coefficients as REC features.

For the SYN features, we have chosen the short-time spectral amplitude with 129 dimensions. As the the twin model's REC and SYN features must use the same framing parameters, the overlap was changed from 120 samples (defined in the ETSI-AFE) to 150 samples, to provide better synthesis quality with a Hanning window of 200 samples.

The video features are 64-dimensional DCT coefficient vectors, encoding the appearance and shape of the speakers mouth. The corresponding mouth region was determined automatically by a Viola-Jones face and mouth detector [14].

All models are trained using the clean signals as described in Section 2.1. Each HMM set consists of 51 whole-word HMMs and one silence HMM. The word HMMs are left-to-right linear models, with the number of states a multiple of the number of phones in the word – a factor of three for the audio and of one for the video models. The states use 4-component diagonal covariance GMMs for the recognition models and 1-component diagonal GMMs for the SYN features.

Table 1. Segmental SNR, PESQ, STOI and Recognition Accuracy

Noise		Segmental SNR			PESQ			STOI			Accuracy	
Type	SNR [dB]	unprocessed	Log-MMSE	Twin HMM	unprocessed	Log-MMSE	Twin HMM	unprocessed	Log-MMSE	Twin HMM	Audio	Audio-Visual
White	15	0.11	1.83	**5.35**	2.37	2.69	**3.02**	0.89	0.87	**0.91**	70.48	93.80
	10	-2.22	-0.39	**2.66**	2.08	2.36	**2.68**	0.82	0.80	**0.85**	49.11	88.92
	5	-4.27	-2.42	**-0.08**	1.82	2.03	**2.31**	0.73	0.71	**0.75**	37.15	79.26
	0	-5.98	-4.27	**-2.49**	1.63	1.72	**1.93**	**0.64**	0.62	0.63	26.95	66.76
Jet	15	0.12	0.79	**4.37**	2.47	2.77	**3.04**	0.90	0.88	**0.92**	85.46	95.04
	10	-2.20	-1.49	**1.60**	2.18	2.44	**2.71**	0.83	0.80	**0.85**	57.09	89.98
	5	-4.26	-3.30	**-1.11**	1.89	2.10	**2.31**	0.72	0.69	**0.73**	38.30	77.40
	0	-5.98	-4.77	**-3.35**	1.67	1.81	**1.90**	**0.61**	0.59	0.60	26.15	64.89
Babble	15	0.35	0.69	**3.99**	2.71	2.88	**3.05**	0.93	0.90	**0.94**	85.63	96.37
	10	-1.99	-1.75	**1.18**	2.41	2.59	**2.74**	0.87	0.83	**0.88**	65.07	91.67
	5	-4.10	-3.52	**-1.39**	2.07	2.26	**2.35**	0.76	0.73	**0.78**	45.21	82.98
	0	-5.88	-4.96	**-3.71**	1.78	1.91	**1.95**	**0.63**	0.60	**0.63**	30.05	69.42
Clean	-	35.00	17.59	18.54	4.50	4.32	4.33	1.00	1.00	1.00	98.76	98.23

In the speech enhancement block, we have used the mean and variance of a Wiener filter as the parameters of the PDF $p(\hat{\mathbf{x}}_{S_t}|\mathbf{x}_{S_t})$ [15]. Improved minima controlled recursive averaging (IMCRA) was used for estimating noise power [16].

3.3. Results

To evaluate the proposed framework, we have used three objective quality measures: segmental SNR [17], PESQ, which correlates well with perceived speech quality in many cases [18], and STOI, which is designed to objectively assess speech intelligibility [19]. Tab. 1 compares the results of the twin HMM framework with those of the conventional log-spectral amplitude MMSE estimator [20] and relates both to the quality of the unprocessed signal.

In almost every test condition the twin-HMM framework clearly outperforms the conventional log-MMSE estimator in terms of all considered quality measures. The de-noised spectra are also much clearer than those obtained with the log-MMSE estimator, as shown in Figure 3.

Tab. 1 also shows the recognition accuracy of the utilized audiovisual ASR system, and compares it to audio-only recognition results that would have been obtained using only the audio REC HMM but no video information. The accuracy of the used audiovisual recognizer can be seen as a main factor influencing performance of the proposed framework, which also highlights the importance of the use of the video information in the more distorted conditions.

4. RELATION TO PRIOR WORK AND CONCLUSIONS

We have presented a new approach for audiovisual speech processing, which uses a fine-grained, precise and synthesis-friendly speech model in the form of a twin-HMM, and applies it in a full audiovisual speech recognition architecture for maximum robustness.

Since audio-visual speech recognition provides a highly robust state estimate in noisy conditions, applicability is less constrained by acoustic conditions than audio-only approaches to state-based signal estimation such as [21, 1, 2]. The use of a full recognizer – rather than of an ergodic, phonetic HMM as in [22, 21] – also distinguishes the approach from other HMM-based speech processing methods and allows for utilizing syntactic and linguistic information.

In addition, the suggested twin-HMMs provide not only a recognition but also a synthesis model. In comparison to prior work, this is a helpful novelty, since we are neither forced to perform recognition in the synthesis domain – as in [1, 2] – nor to find nonlinear transformations between the recognition features and the synthesis domain – as in [3], or in [23], which presents results only for cepstral domain *feature* enhancement. This is important because speech resynthesized from cepstrum features suffers from low quality, due to the loss of information inherent in cepstral feature extraction. Also, the twin-HMM couples the synthesis output distribution directly to the *audiovisual* state. This differs from recent audiovisual speech enhancement methods where state-dependent linear transformations are found between *visual* data and the clean audio spectrum, as in [22, 24].

Combining these optimizations has allowed our approach of twin-HMM-based audiovisual speech processing to notably outperform the standard log-MMSE speech estimator [20] in terms of the segmental SNR, the perceptually motivated PESQ measure, and the intelligibility estimate provided by the STOI measure, making it an interesting alternative framework for speech processing in highly distorted environments.

5. REFERENCES

[1] L. Gagnon, "A state-based noise reduction approach for non-stationary additive interference," *Speech Communication*, vol. 12, pp. 213–219, 1993.

[2] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 40, p. 725–735, 1992.

[3] C. W. Seymour and M. Niranjan, "An HMM based cepstral-domain speech enhancement scheme," in *Proc. Interspeech*, Yokohama, Japan, 1994, p. 1595–1598.

[4] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[5] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1274–1288, 2002.

[6] D. Kolossa, S. Zeiler, A. Vorwerk, and R. Orglmeister, "Audiovisual speech recognition with missing or unreliable data," in *Proc. International Conference on Auditory-Visual Speech Processing*, University of East Anglia, Norwich, UK, 2009, pp. 117–122.

[7] A. Vorwerk, S. Zeiler, D. Kolossa, R. F. Astudillo, and D. Lerch, *Robust Speech Recognition of Uncertain or Missing Data*. Springer, 2011, ch. Use of Missing and Unreliable Data for Audiovisual Speech Recognition, pp. 345–373.

[8] A. H. Abdelaziz and D. Kolossa, "Decoding of uncertain features using the posterior distribution of the clean data for robust speech recognition," in *Proc. Interspeech*, Portland, Oregon, USA, 2012.

[9] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[10] Institute for Perception-TNO and Speech Research Unit-RSRE, retrieved November 2012. [Online]. Available: http://spib.rice.edu/spib/data/signals/noise/

[11] *Objective measurement of active speech level*, International Telecommunication Union Std., 1993.

[12] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, 1980.

[13] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*, ETSI, ES.202.050 Std., 2003.

[14] G. Bradski and A. Kaehler, *Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.

[15] R. F. Astudillo, D. Kolossa, and R. Orglmeister, "Accounting for the uncertainty of speech estimates in the complex domain for minimum mean square error speech enhancement," in *Proc. Interspeech*, Brighton, United Kingdom, 2009.

[16] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio*, vol. 11, no. 5, p. 466–475, 2003.

[17] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of Speech Quality*. Prentice-Hall, 1988.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, Salt Lake City, Utah, USA, 2001.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, Dallas, Texas, USA, 2010.

[20] P. C. Loizou, *Speech enhancement: theory and practice*. CRC Taylor and Francis, 2007.

[21] 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*, 2003, p. 82–86.

[22] I. Almajai and B. Milner, "Visually derived Wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.

[23] G. Potamianos, C. Neti, and S. Deligne, "Joint audio-visual speech processing for recognition and enhancement," in *Proc. AVSP*, 2003.

[24] F. Berthommier, "Characterization and extraction of mouth opening parameters available for audiovisual speech enhancement," in *Proc. ICASSP*, Montreal, Quebec, Canada, May 2004.