TWIN-HMM-BASED NON-INTRUSIVE SPEECH INTELLIGIBILITY PREDICTION

Mahdie Karbasi, Ahmed Hussen AbdelAziz, and Dorothea Kolossa

Institute of Communication Acoustics, Cognitive Signal Processing Group,
Ruhr-Universität Bochum, 44801 Bochum
Email: {Mahdie.Karbasi, Ahmed.HussenAbdelAziz, Dorothea.Kolossa}@rub.de

ABSTRACT

Most of the objective measures employed for speech intelligibility prediction require a clean reference signal, which is not accessible in all realistic scenarios. In this paper, we propose to re-synthesize the relevant features of the clean signal using only the noisy speech signal and utilize them inside an intelligibility prediction framework which requires a reference. A statistical model called twin hidden Markov model (THMM) is used to synthesize the clean speech features. For the intelligibility prediction framework, the short-time objective intelligibility (STOI) measure is used as an accurate and well-known method. The experimental results show a high correlation between the twin-HMM-based STOI (THMMB-STOI) and the human speech recognition results, even slightly outperforming the conventional STOI predictions computed using the actual clean reference signals.

Index Terms— Speech intelligibility prediction, twin HMM, non-intrusive method, objective measures

1. INTRODUCTION AND RELATION TO PRIOR WORK

Speech intelligibility is a measure assessing in how far a speech signal is recognizable. The most reliable way to estimate this quantity is to conduct intelligibility tests with the help of human listeners. However these tests are time consuming and costly. Therefore, there have been many efforts in the last decades to automatically estimate this measure.

Some early introduced and widely-known intelligibility measures include the articulation index (AI) [1], the speech intelligibility index (SII) [2], and the speech transmission index (STI) [3]. These measures have been reported to have an acceptable accuracy but in a limited number of degradation types like linear filtering and additive noise. Later, in order to cope with more complex distortions, new methods like the speech-based envelope power spectrum model (EPSM) [4] were introduced. Also, the short time objective intelligibility (STOI) measure [5], and the mutual information based [6] methods have been proposed more recently.

All of the aforementioned measures and also the majority of the published models so far, fall into the category of intrusive methods, which means that they need a clean reference signal to estimate the intelligibility of the corrupted signal. Since the clean signal is not available in all situations, the requirement to access the reference signal would be a severe limitation for an intelligibility prediction algorithm. Thus, non-intrusive methods have been introduced in order to estimate the intelligibility without a need for the clean signal. For example, speech-to-reverberation modulation energy ratio (SRMR) [7] is a non-intrusive method, which computes the intelligibility of reverberated speech. This measure assumes that lower modulation bands are carrying the speech signal and the higher bands contain the room acoustic information such as reverberation. However, the SRMR method is limited to the assessment of reverberated speech signals and is not suitable for other types of degradations.

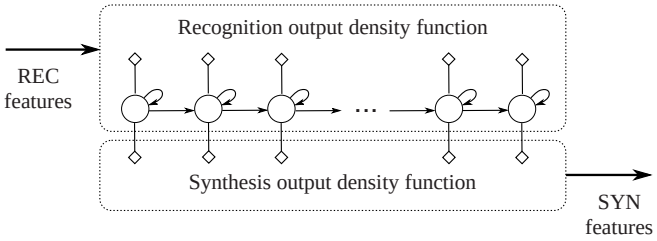

Fig. 1. Concept of twin HMMs.

Low cost intelligibility assessment (LCIA) [8, 9] is another non-intrusive framework, which uses frame-based speech feature extraction and selection methods in combination with Gaussian mixture models (GMMs) for prediction. In [10] a joint acoustic and phonological framework has been suggested to predict the speech intelligibility non-intrusively. The authors use an auditory-model-based feature extraction along with a hybrid speech model to overcome the need for a clean signal.

The above methods use clean generative or discriminative models to estimate the intelligibility from a degraded speech signal. The clean models are trained on clean data during an off-line phase. In this current work, we also propose to use a statistical-model-based approach, but not to extract the intelligibility measure directly. We propose instead to synthesize the relevant features of the clean signal using the statistical model trained on clean data. Then the synthesized features can be used inside any intrusive framework for predicting the intelligibility. This is the main difference between the current work and the previously proposed non-intrusive methods. The proposed method is not limited to the intelligibility estimation procedure that we used and can be integrated in any intrusive method. Therefore the proposed method can take advantage of accurate and reliable intrusive methods and predict intelligibility without requiring the clean signal.

The twin hidden Markov model (THMM) [11, 12] is a statistical model which was previously applied to speech signal enhancement. The main concept of twin HMMs is presented in Figure 1. Here, we propose to use this model for obtaining clean features (instead of clean speech) out of a degraded speech signal. They are needed inside an intrusive framework to predict the intelligibility. THMMs can be optimized for synthesizing specific clean speech features while other simple speech enhancement methods like the Wiener filter or spectral subtraction do not have this capability. The results show a strong correlation between the predicted intelligibility and the human recognition accuracy.

The remainder of this paper is organized as follows: In Section 2,

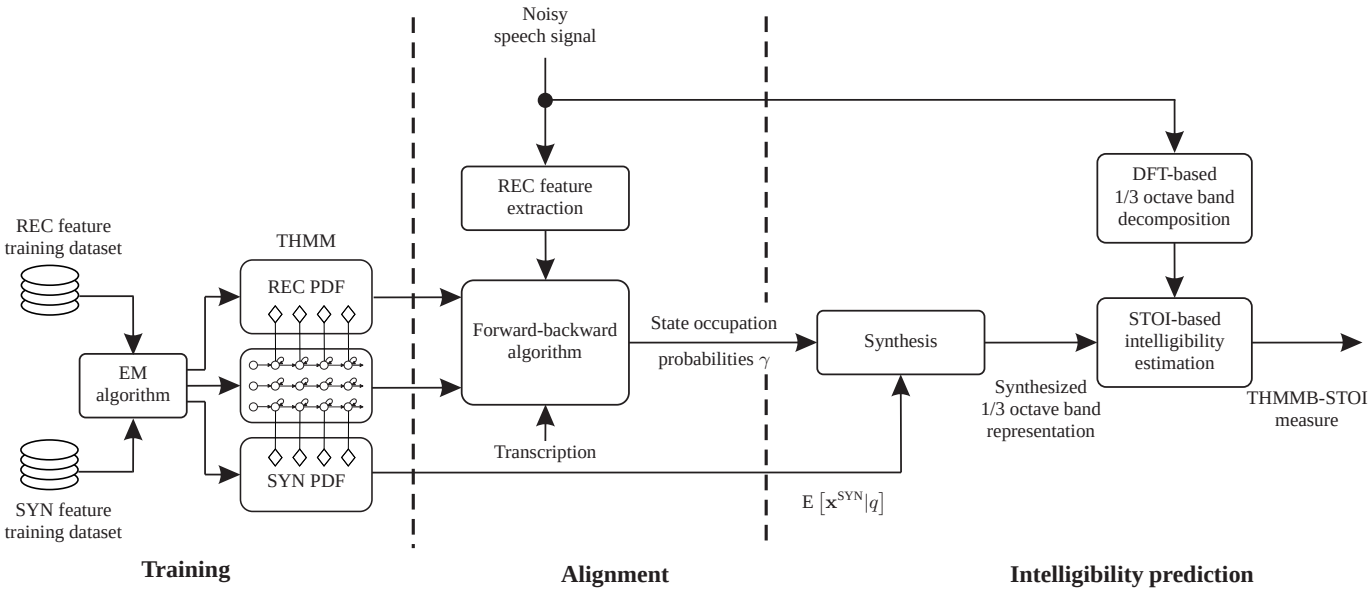

Fig. 2. Framework of speech intelligibility prediction using twin hidden Markov models.

the general concept of THMMs is introduced and the basic structure of our framework is described. Section 3 gives a detailed explanation of all experimental settings and analyzes the results. Finally the conclusions are presented in Section 4.

2. TWIN-HMM-BASED INTELLIGIBILITY PREDICTION

The idea of using twin HMMs (THMMs) in predicting speech intelligibility is a general approach. THMMs can be integrated into the framework of other intrusive intelligibility predictors in order to estimate the relevant clean features from the noisy test signal. This is a notable advantage of the proposed method over other non-intrusive ones. The new approach can work in combination with many other intelligibility prediction methods and compensate their problem of requiring the clean reference signal. To implement our idea, we have chosen the STOI [5] algorithm as the intelligibility prediction measure in this paper. It has been shown in several studies that the STOI is a reasonable predictor of intelligibility and it is currently in wide use [6, 9, 13, 14].

2.1. Twin Hidden Markov Models (THMMs)

As can be seen in the conceptual representation of the twin HMMs in Figure 1, there is only one state sequence in a THMM. However each state is associated with two output density functions (ODF); one for recognition and one for synthesis. The state sequence represents the temporal evolution of speech. The recognition ODFs are trained using recognition features (REC features) and the synthesis ODFs are trained using synthesis features (SYN features). THMMs have been introduced in order to use ASR systems for synthesizing a cleaner version of a speech signal and for speech enhancement. The REC features, which are appropriate for maximizing the recognition accuracy, are used in the THMM framework to decode the best state sequence [11]. Then the decoded state sequence in combination with the synthesis ODFs is used to synthesize a clean reference signal.

2.2. THMM-based STOI

Figure 2 shows the global scheme of the proposed framework. This approach is composed of three main phases; training, alignment, and intelligibility prediction, which are explained in detail in the following paragraphs.

In the first phase, the standard expectation maximization (EM) algorithm is used to train a THMM set. Using the iterative EM algorithm and the REC features, the recognition output density functions of the THMM set are learned. In the last iteration, the occupation probabilities of all states over time γ, are stored for later computations. The training procedure of the synthesis output density functions accumulates the SYN features weighted with the stored γ from the final iteration of the REC distribution training.

In the alignment phase, the REC features are extracted from the noisy speech signal. Then, the features along with the transcription data are fed into a forced alignment system. In this system a forward-backward algorithm is used to estimate the state occupation probabilities γ per time frame. The REC distributions trained during the first phase are used in this estimation process. Using the transcription data and employing a forced alignment algorithm leads to a high accuracy of the estimated γ matrix.

In the third and last phase, at first the relevant clean features are synthesized in order to be used in the intelligibility prediction step. The synthesized features are thus obtained in the SYN feature domain, which is the one-third octave frequency band here. The DFT-based one-third octave band decomposition is also applied to the noisy speech signal. Later, the extracted noisy features together with the synthesized clean features are used inside the STOI-based intelligibility estimation block to estimate the intelligibility measure called THMMB-STOI. This last block implements short-time segmentation, normalization, clipping, and correlation computation between its two inputs, exactly as described in the STOI algorithm [5].

To synthesize the clean signal in this phase, the synthesis output density distribution learned in the training phase, and the state occupation probability obtained in the alignment phase are used. In fact the mean of the SYN distribution in each state $\mathrm{E}[\mathbf{x}_t^{\text{SYN}}|q_t = i]$ is weighted by the occupation probability of the same state $P(q_t = i|\mathbf{x}_t^{\text{REC}})$ and summed over all states to obtain the synthesized ($\mathbf{x}_t^{\text{SYN}}$):

$$\mathbf{x}_t^{\text{SYN}} = \sum_{i=1}^{N} \mathrm{E}(\mathbf{x}_t^{\text{SYN}}|q_t = i)\mathrm{P}(q_t = i|\mathbf{x}_t^{\text{REC}}). \qquad (1)$$

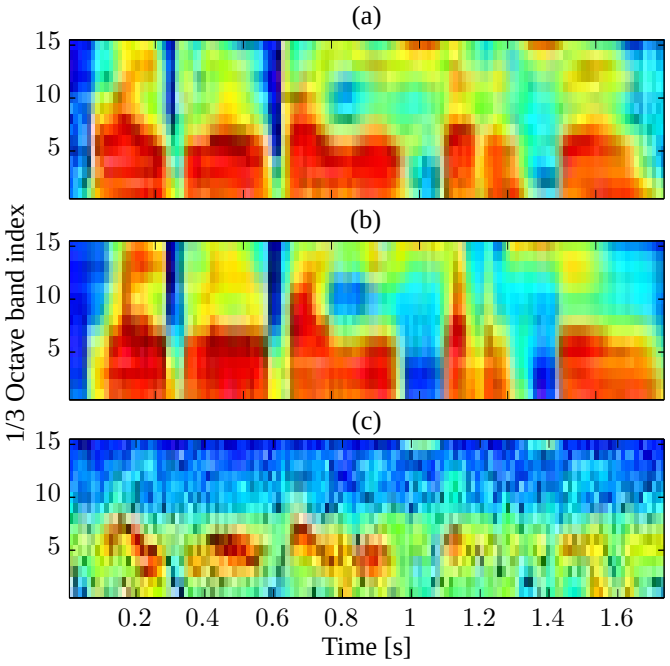

Fig. 3. 1/3 octave band representation of (a) the clean Grid sentence "lay blue by u 7 soon", (b) the same signal synthesized using a THMM, and (c) the corresponding distorted signal at 0 dB SNR.

Here, N is the number of states, t is the frame index and q_t is the state at time index t.

As mentioned earlier, THMMs were primarily introduced for speech enhancement inside an audio-visual speech recognition system. However, there are main structural differences between the current framework and the THMM-based speech enhancement system. The proposed method utilizes the transcription information instead of the automatic speech recognition output to decode the speech content of the signal. Besides, the speech enhancement system receives both audio and video signals as input and uses a combination of the synthesized and the enhanced signals to produce its output signal. It also must be noted that in our proposed method, the relevant clean features are synthesized instead of the clean speech signal.

Figure 3 shows the one-third octave band representation of a noisy signal at 0 dB SNR and its equivalent clean and synthesized versions. One can observe that the clean version of the noisy signal has been retrieved very well using the THMM approach and is in line with its actual clean counterpart. In Figure 4 the estimated THMMB-STOI has been plotted versus the conventional STOI. These results have been computed over all SNRs using the test data set. As expected, a strong and almost linear correlation between the two measures is observable.

3. EXPERIMENTS AND RESULTS

3.1. Data set

In this study, the Grid corpus [15] has been used as the speech database. The original corpus contains 34000 clean speech utterances in total. In addition to the original data, there is also a noisy version of the corpus at 12 different SNRs in the range from 40 down to -14 dB. At each SNR, there are in total 2000 noisy speech signals created by adding speech shaped noise (SSN) to the clean speech

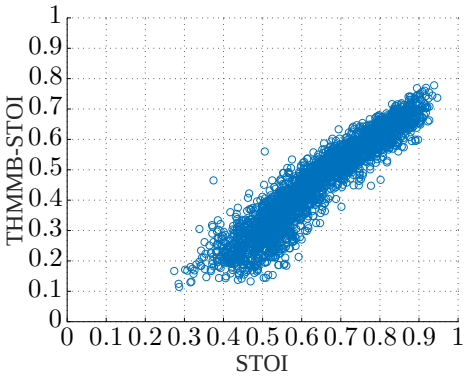

Fig. 4. Scatter plot of the conventionally estimated STOI measure against the THMMB-STOI measure.

utterances from the original Grid corpus. This part also contains results of listening tests conducted by Jon Barker at the University of Sheffield with 20 listeners at each SNR. The sentences in the Grid database comprise six words, following the structure: verb-color-preposition-letter-digit-adverb. However, in the listening tests, the listeners were only asked to recognize the words in the positions of color, letter, and digit. The speech signals at each SNR have been divided randomly into training (80 %), development (10 %) and test sets (10 %) for the following experiments. Also the clean version of the files from each of these sets have been collected to create the equivalent clean training, development and test sets. Since we are proposing to synthesize the clean relevant intelligibility features using the noisy signal and considering the fact that the data at 40 dB SNR can almost be considered clean, this SNR was excluded in the following experiments.

3.2. Experimental setup

In order to obtain a consistent framework with comparable results, all common feature extraction parameters (like frame length, analysis window type, etc.) have been selected as suggested in the STOI framework [5]. This also applies to the frequency sampling of the speech signals, which have been down-sampled from $f_s = 25$ kHz to 10 kHz. The REC features are standard mel frequency cepstral coefficients (MFCCs), which are being used widely in automatic speech recognition tasks. These features are 39-dimensional vectors, composed of the first 13 static MFCCs and their corresponding first (Δ) and second time derivatives ($\Delta\Delta$). According to the selected intelligibility measure, which is the STOI, the SYN features are the one-third octave band representation of the signal in the DFT domain. These features are 15 dimensional vectors, which later are used to compute the STOI measure. Each word is modeled using a linear left-to-right HMM. Therefore we have 51 whole-word HMMs and one silence model in each HMM set. The number of states has been chosen as three times the number of phonemes of the word. A 2- and 1-mixture diagonal covariance GMM is used for modeling the state distribution of recognition and synthesis HMMs respectively. For recognition we trained noise-dependent models using training set data at each SNR separately and evaluated the accuracy of these models with development sets. However, for synthesis, a universal clean model was used for all conditions.

3.3. Results

In the following experiments, the performance of the objective measures is being compared to the human speech recognition accuracy

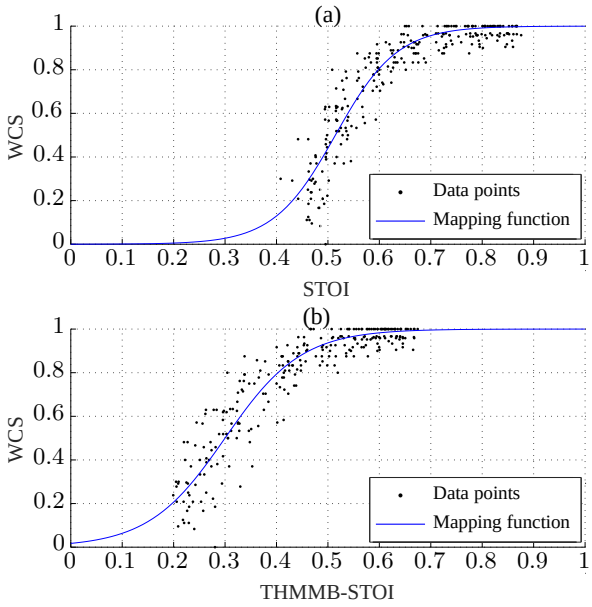

Fig. 5. Scatter plots of (a) the conventionally estimated STOI measure and (b) the THMMB-STOI measure against the word correct scores (WCS), and the corresponding trained logistic functions.

which is stated as the word correct score (WCS). This subjective measure is computed by dividing the number of correctly recognized keywords by the total number of keywords.The reported WCS here was averaged over ten files. Similarly, the results of the objective measures, e.g. STOI, were also computed over the same ten files.

In order to evaluate the performance of the intelligibility measures, three different figures of merit have been used: root mean square error (RMSE), normalized cross-correlation coefficient (NCC), and Kendall's Tau coefficient (τ) [5, 6]. The first two performance measures are valid only when their input variables have a linear relationship. Thus, we used a sigmoid mapping function and the logistic regression method to linearize the relationship between the machine-derived and the human listening results (WCS). The mapping procedure has been performed exactly based on [6].

The mapping functions derived for the STOI and the THMMB-STOI along with the test data are shown in Figure 5. The functions have been derived separately for each intelligibility measure using the development data over all SNRs. As can be seen in Figure 5, the test data appears to fit the mapping function. In Table 1 the accuracy of the proposed measure (THMMB-STOI) as well as the conventional STOI (STOI) are presented. The results have been computed as an overall accuracy using the test data from all SNRs. Larger values of NCC and Kendall's Tau (τ) show higher correlation between the objective measures and the human performance, which consequently means higher accuracy in predicting the intelligibility. In contrast, lower values of RMSE represent a higher accuracy of the intelligibility prediction methods. It can be clearly seen, that the proposed method (THMMB-STOI) has a good performance in terms of all evaluation measures, and is comparable to the conventional STOI. Table 2 shows the prediction errors of the STOI and the THMMB-STOI in each SNR separately. As it was expected, the RMSE of both prediction methods decreases (improves) as the SNR increases. However, there are some inconsistencies between some SNRs. For example, while the SNR improves from -14 to -12 dB, the RMSE value increases for both intelligibility measures.

In total, the proposed THMM-based method was successful in predicting the intelligibility with relatively high accuracy and without having the clean reference signal. However, there is an increased computational complexity in the proposed method. Running on a standard PC (Intel(R) Core(TM) i5-4570 CPU @ 3.20 GHz, 8 GB memory) with Windows 10 and Matlab 2015a, the processing time was 1.92 s on average for a speech signal of 1.76 s average length.

Table 1. Performance of objective measures in terms of NCC (%), RMSE, and Kendall's Tau (τ %) between objective measures and listening test results (WCS)

Measure	NCC (%)	RMSE	τ (%)
STOI	93.55	0.092	74.52
THMMB-STOI	93.57	0.091	74.38

Table 2. RMSE between mapped objective measures and listening test results in different SNRs

SNR (dB)	STOI	THMMB-STOI
-14	0.132	0.108
-12	0.165	0.176
-10	0.109	0.111
-8	0.153	0.149
-6	0.086	0.099
-4	0.052	0.052
-2	0.060	0.064
0	0.046	0.043
2	0.038	0.034
4	0.037	0.032
6	0.039	0.034
All	0.092	0.091

4. CONCLUSIONS

We have proposed a solution to allow the use of intrusive speech intelligibility prediction algorithms in situations when the reference signal is not accessible. The proposed method is based on synthesizing the clean features by means of twin HMMs. This statistical model gives us the freedom to choose the synthesis features according to the intelligibility prediction method that is going to be used. After linearization, the predicted intelligibility using synthesized clean features shows a strong correlation with human data and agrees with the intelligibilities computed with actual reference features. This method can be integrated into the framework of many other intelligibility prediction measures and can provide synthesized clean reference features, which is a requirement in intrusive methods. To synthesize the clean features in this work, transcription data was used. Speech shaped noise was utilized to evaluate on the Grid database, which possesses a specific sentence structure. These facts indicate a direction for future work, namely making THMMs more independent of extra data and more flexible with respect to the task while keeping their accuracy high.This can likely be achieved by training THMMs on large-vocabulary corpora of speech signals, and carrying out speaker adaptation during test time.

5. ACKNOWLEDGMENTS

This research has received funding from the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n°[317521]. The authors would like to thank Jon Barker for providing a noisy version of the Grid database with comprehensive listening test results.

6. REFERENCES

[1] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, p. 90–119, January 1947.

[2] *Methods for the Calculation of the Speech Intelligibility Index*, S3.5-1997, ANSI, New York, NY, USA, 1997.

[3] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[4] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the envelope power signal-to-noise ratio after modulation-frequency selective processing," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2384–2384, 2011.

[5] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[6] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 1, pp. 6–16, Jan 2014.

[7] T. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1766–1774, 2010.

[8] D. Sharma, G. Hilkhuysen, N. Gaubitch, P. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Signal Processing Conference, 2010 18th European*, 2010, pp. 1899–1903.

[9] D. Sharma, P. Naylor, and M. Brookes, "Non-intrusive speech intelligibility assessment," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, 2013, pp. 1–5.

[10] S. Nemala and M. Elhilali, "A joint acoustic and phonological approach to speech intelligibility assessment," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 4742–4745.

[11] A. Abdelaziz, S. Zeiler, and D. Kolossa, "Twin-HMM-based audio-visual speech enhancement," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 3726–3730.

[12] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Using twin-HMM-based audio-visual speech enhancement as a front-end for robust audio-visual speech recognition," in *INTERSPEECH*, 2013, pp. 867–871.

[13] J. Gao and A. Tew, "The segregation of spatialised speech in interference by optimal mapping of diverse cues," in *ICASSP*, 2015, pp. 2095–2099.

[14] C.-C. Hsu, K.-M. Cheong, J.-T. Chien, and T.-S. Chi, "Modulation Wiener filter for improving speech intelligibility," in *ICASSP*, 2015, pp. 370–374.

[15] 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, November 2006.