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

Blind Non-Intrusive Speech Intelligibility Prediction using Twin-HMMs

Mahdie Karbasi[1], *Ahmed Hussen Abdelaziz*[2], *Hendrik Meutzner*[1], *Dorothea Kolossa*[1]

[1]Cognitive Signal Processing Group, Ruhr-Universität Bochum, 44801 Bochum, Germany
{mahdie.karbasi, hendrik.meutzner, dorothea.kolossa}@rub.de
[2]International Computer Science Institute, Berkeley, USA
ahmedha@icsi.berkeley.edu

Abstract

Automatic prediction of speech intelligibility is highly desirable in the speech research community, since listening tests are time-consuming and can not be used online. Most of the available objective speech intelligibility measures are intrusive methods, as they require a clean reference signal in addition to the corresponding noisy/processed signal at hand. In order to overcome the problem of predicting the speech intelligibility in the absence of the clean reference signal, we have proposed in [1] to employ a recognition/synthesis framework called twin hidden Markov model (THMM) for synthesizing the clean features, required inside an intrusive intelligibility prediction method. The new framework can predict the speech intelligibility equally well as well-known intrusive methods like the short-time objective intelligibility (STOI). The original THMM, however, requires the correct transcription for synthesizing the clean reference features, which is not always available. In this paper, we go one step further and investigate the use of the recognized transcription instead of the oracle transcription for obtaining a more widely applicable speech intelligibility prediction. We show that the output of the newly-proposed blind approach is highly correlated with the human speech recognition results, collected via crowdsourcing in different noise conditions.

Index Terms: Speech intelligibility prediction, twin-HMM, speech recognition, speech synthesis, non-intrusive methods, objective measures.

1. Introduction

The need to automatically assess the speech intelligibility is growing in the field of speech signal processing. During the last decades, many efforts have been made to present an accurate machine-computable metric of speech intelligibility. Most of the published measures are referred to as intrusive methods, as they require a clean reference speech signal—in addition to the degraded/processed speech signal—for predicting the speech intelligibility. Measures like the articulation index (AI) [2], the speech transcription index (STI) [3], the speech intelligibility index (SII) [4], the short-time objective intelligibility (STOI) [5], and mutual-information-based techniques such as [6], all belong to the category of intrusive metrics. These measures can not be used in situations where the clean reference signal is inaccessible. This property poses a disadvantage for such algorithms and reduces their applications. Another general disadvantage of such methods is that they are not applicable to scenarios like voice conversion or even bandwidth extension, because the comparison to the clean reference signal will indicate unreasonable low intelligibilities in such scenarios [7]. In contrast to these metrics, there are non-intrusive methods, which do not require the clean reference speech signal. For instance, in [8], a hybrid discriminative-generative statistical model is used in combination with auditory models in order to predict the intelligibility without using the clean reference signal. In another work [9], a large set of speech specific features and feature dimensionality reduction techniques are used to train a Gaussian mixture model (GMM), from which the intelligibility of unseen data is predicted. Employing a combination of long- and short-term features in a binary tree regression model to predict the intelligibility is another approach suggested in [10]. Falk et al. [11] use an auditory model to compute the speech to reverberation modulation energy ratios (SRMR) for predicting the intelligibility. The performance of the SRMR measure is investigated in [12], for predicting the intelligibility results of hearing-impaired listeners. The SRMR was later improved by taking into account the sources of speech variation including pitch and speech content [13]. It was shown that these variations can degrade the performance of the measure.

In a completely new and different framework, we have recently proposed the so-called THMM-based non-intrusive speech intelligibility prediction algorithm [1]. This algorithm re-synthesizes the clean features that are required in an intrusive speech intelligibility prediction framework, instead of directly estimating the intelligibility. It has been shown that this method can successfully estimate the clean features that are required by the STOI algorithm. As the THMM-based method requires the transcription of the noisy speech signal for synthesizing the clean features, we used the true transcriptions in [1] for synthesis. However, in many scenarios the oracle transcription would not be available. In order to overcome this limitation, we propose to use the THMM approach for first recognizing the transcription of the noisy speech signal, and then for synthesizing the clean features.

Our results show that the proposed blind THMM-based intelligibility prediction approach is highly correlated with the human intelligibility scores for various noise conditions.

2. THMM-based Speech Intelligibility Prediction

The twin HMM is a statistical model introduced by Abdelaziz et al. [14] for audio-visual speech enhancement. Like conventional HMMs, the THMM is composed of a sequence of states modeling a time series of observations, e.g., the feature observation sequence of speech signals. However, in the THMM, each state is associated with two output density functions (ODFs): Recognition (REC) ODFs are used for modeling the distribution of features suitable for recognition (REC features), and synthesis (SYN) ODFs are employed for modeling the distribution of

http://dx.doi.org/10.21437/Interspeech.2016-155

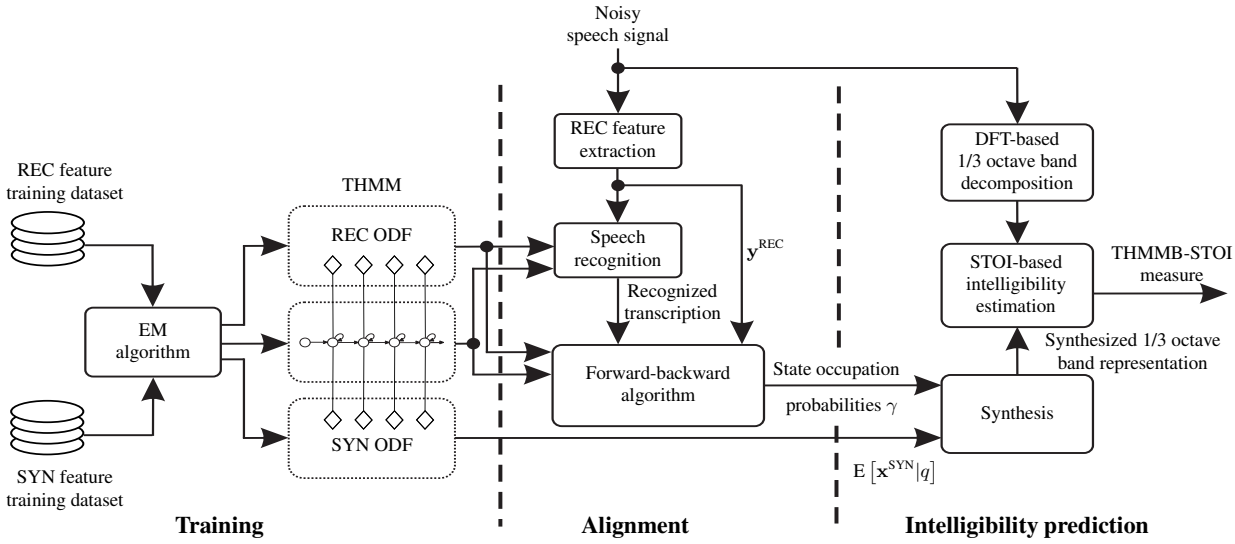

Figure 1: Framework of blind speech intelligibility prediction using twin hidden Markov models.

features appropriate for synthesis (SYN features). This characteristic of the THMM approach makes it a convenient tool for synthesizing clean speech from a given noisy speech signal. The REC features that are used for speech recognition are chosen to have a high accuracy in decoding the best state sequence in HMMs and hence in automatically recognizing speech signals. In contrast, suitable features for synthesis do not have the same strength in speech recognition tasks. The ability of THMMs to share the same state sequence with two different distributions makes it possible to use the REC and the SYN features simultaneously for synthesizing a clean signal, leading to a model that can optimize both recognition and synthesis performance through an appropriate choice of features.

The main idea behind using THMMs in predicting the intelligibility is to synthesize the clean reference features for an intrusive intelligibility prediction method such as STOI [1]. In current framework, THMMs are used to recognize the unknown transcription of a given noisy signal, prior to synthesizing the clean reference features. Applying this idea to the THMM-based method makes it fully blind and independent of any additional data during the intelligibility prediction phase. In order to implement this idea, we have chosen the STOI method [5] for the intelligibility prediction part.

The framework of the proposed blind method is shown in Figure 1. It is explained in detail in the following sections.

2.1. Training

During the off-line training phase shown on the left-hand-side, two sets of output density functions are trained. At first, the recognition distributions (REC ODFs) are trained using the REC features and the iterative expectation maximization (EM) algorithm. In the last iteration of this algorithm, the state occupation probabilities γ are stored.

Then, for training the synthesis distributions (SYN ODFs), the SYN features are weighted with the state posterior probabilities γ and finally accumulated to form the SYN ODFs, as in [14].

2.2. Alignment

In this phase, shown in the center of Figure 1, the REC features are first extracted from the noisy or degraded speech signal, the intelligibility of which we need to assess. Then, the REC features together with the REC ODFs, trained in the training phase, are used in a Viterbi algorithm to recognize the speech content of the signal. The transcription information, resulting from the speech recognition step, is then used in the forward-backward algorithm, which uses the REC features in order to compute the state posterior probabilities, the γ-matrix, defined by $\gamma(j,t) = \mathrm{p}\left(q_t = j|\mathbf{y}_t^{\text{REC}}\right)$. Here, p(.) represents the probability distribution function. q_t and $\mathbf{y}_t^{\text{REC}}$ are the THMM state and the noisy REC feature vector at time frame t, respectively.

The main difference between this new framework and the previous one in [1] occurs in this phase. In [1] the oracle transcriptions were used in order to compute the state occupation probabilities in the forward-backward algorithm. Requiring the transcription data can still be restrictive for some applications. Hence, we propose to use the automatic speech recognition model, embedded in the THMMs, to create this information. This modification makes the THMM-based method effectively self-reliant.

2.3. Intelligibility Prediction

The actual intelligibility prediction is shown in the right part of Figure 1. It uses the γ-matrix from the alignment phase together with the SYN ODFs of the THMM speech model to estimate the reference features for an intrusive intelligibility prediction. As mentioned above, the STOI algorithm has been chosen for this purpose. Therefore, the STOI-relevant features, the DFT-based one-third octave band decomposition, are at first extracted from the noisy speech signal in this phase. Then, the clean version of the same type of features is synthesized from the THMM. Lastly, the noisy and clean features are compared inside the STOI-based intelligibility prediction block and the THMM-based-STOI (THMMB-STOI) measure is computed. The intelligibility prediction block is exactly the same as in the original STOI method after time-frequency analysis. It consists of short-time segmentation, clipping and normalization and finally correlation coefficient computation.

To derive the equations for estimating the clean SYN feature vectors, a minimum mean square error (MMSE) estimator has been used. The cost function for the MMSE criterion, C_{MMSE}, is defined as follows:

$$C_{\text{MMSE}}\left(\hat{\mathbf{x}}_t^{\text{SYN}}\right) = \mathbb{E}\left[\left(\mathbf{x}_t^{\text{SYN}} - \hat{\mathbf{x}}_t^{\text{SYN}}\right)^2\right] \quad (1)$$

where $\mathbb{E}[.]$ stands for the expectation value, $\hat{\mathbf{x}}_t^{\text{SYN}}$ represents the estimate of the clean SYN feature vector at time frame t and $\mathbf{x}_t^{\text{SYN}}$ is the corresponding true feature vector. Minimizing C_{MMSE} given only the observed features $\mathbf{y}_t^{\text{REC}}$, which are the noisy REC feature vector, results in the following optimum estimate:

$$\hat{\mathbf{x}}_t^{\text{SYN}} = \mathbb{E}\left[\mathbf{x}_t^{\text{SYN}} | \mathbf{y}_t^{\text{REC}}\right]. \quad (2)$$

Equation (2) can be marginalized over all states of the THMM:

$$\hat{\mathbf{x}}_t^{\text{SYN}} = \sum_{j=1}^{N} \text{p}\left(q_t = j | \mathbf{y}_t^{\text{REC}}\right) \mathbb{E}\left[\mathbf{x}_t^{\text{SYN}} | q_t = j, \mathbf{y}_t^{\text{REC}}\right]. \quad (3)$$

Here, N is the number of states and q_t is the state at time frame t. Since the THMM assumes that $\mathbf{x}_t^{\text{SYN}}$ and $\mathbf{y}_t^{\text{REC}}$ are conditionally independent when q_t is known, Equation (3) can be reformulated as follows to obtain the final equation utilized for synthesizing the clean feature vectors:

$$\hat{\mathbf{x}}_t^{\text{SYN}} = \sum_{j=1}^{N} \text{p}\left(q_t = j | \mathbf{y}_t^{\text{REC}}\right) \mathbb{E}\left[\mathbf{x}_t^{\text{SYN}} | q_t = j\right]. \quad (4)$$

Hence, to estimate the clean SYN features, the mean of the SYN ODFs in each state, $\mathbb{E}\left[\mathbf{x}_t^{\text{SYN}} | q_t = i\right]$, is weighted with the posterior probability of occupying this state $\text{p}\left(q_t = j | \mathbf{y}_t^{\text{REC}}\right) = \gamma(j,t)$, and summed over all states.

Figure 2 shows the one-third octave band representation of a distorted signal with white noise 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 quite well using the THMM-based approach and is mostly similar to the actual clean counterpart, despite the unavailability of the reference transcription.

3. Experiments and Results

3.1. Dataset

The speech database used in this work was the Grid corpus [15]. In this corpus, there are 34 speakers, each of whom has uttered 1000 clean speech signals with a simple grammar, 6 words per sentence, each of the form verb-color-preposition-letter-digit-adverb. We also used a noisy version of this database, created using speech-shaped noise (SSN) at 9 different SNRs in the range from 6 dB down to -10 dB with steps of 2 dB. At each SNR, there are in total 2000 noisy speech signals. The intelligibility listening test results have been collected by Cooke et al. [16] for this noisy version of the corpus. In addition, two other noisy versions of Grid were created specifically for this study using white and babble noise at the same SNRs of the SSN data. Before starting the experiments, the data were randomly divided into training (80%), development (10%), and test (10%) sets at each SNR and each noise type separately. The training sets were used to train the THMM, and the development sets were used during the training phase to verify the accuracy of the THMM

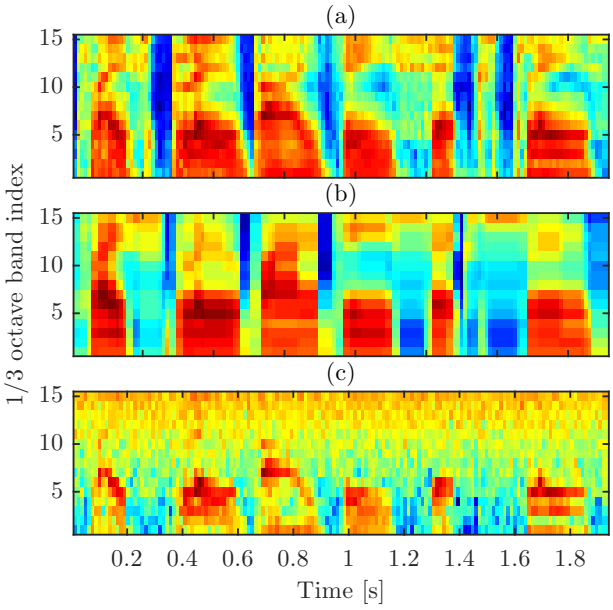

Figure 2: 1/3 octave band representation of (a) the clean Grid sentence "bin blue by v 6 please", (b) the same signal synthesized using a THMM, and (c) the corresponding distorted signal with white noise at 0 dB SNR.

distributions. Finally, the performance of the proposed and the baseline methods in predicting the speech intelligibility were evaluated using the test set data. To obtain the human speech recognition results, three separate listening tests were carried out over the test and development set signals of each noise type.

3.2. Listening Tests

We measure the speech intelligibility of the above-mentioned three noisy datasets by means of a large-scale listening experiment, using crowdsourcing tests at CrowdFlower [17]. Each test participant was asked to transcribe a set of 22 audio signals, containing different SNR conditions, ranging from -10 dB to 6 dB in steps of 2 dB, where the noise type was fixed for a given test set. In order to prevent memorization, we ensured that the same utterance text was only utilized once within a given test set.

Each test set also contained 4 clean utterances that were used for quality control[1]. Those clean utterances were randomly located between the actual test signals. Based on the transcription results of the control utterances, only those participants have been considered for the experiment that correctly transcribed more than 50 % of the clean digits.

The transcriptions were recorded using a multiple-choice approach by providing selections forms, i.e., radio buttons and drop down menus. Each contributor was allowed to participate multiple times but restricted to solve at most 6 test sets. The payment for transcribing a single utterance was $0.01.

We have collected the responses from 849 individual participants, considering only those participants who have passed the quality control requirements during the test. This corresponds to an overall number of 36018 transcribed utterances.

[1] Quality control can be helpful to identify contributors that are not working fairly and to exclude those that are not sufficiently qualified for the task (e.g., due to missing language skills in English).

3.3. Experimental setup

The first 13 Mel frequency cepstral coefficients (MFCCs) plus their first (Δ) and second order derivatives ($\Delta\Delta$) were used as REC features. For synthesis, the DFT-based one-third octave band decomposition of the signal was used, as it is the type of feature needed in the STOI algorithm. All other feature extraction parameters were also set as suggested in [5] for the STOI algorithm.

The REC distributions were trained noise-dependently using the training datasets. In contrast, the SYN distributions were trained using only the clean data. Using these settings, we could get the highest possible accuracy in decoding the state posterior probabilities and in synthesizing the clean features. It must be noted that the requirement of clean data is limited to the off-line training phase and no clean data is needed during the intelligibility prediction phase.

3.4. Results and Discussion

To evaluate the performance of the objective speech intelligibility measures, their predictions were compared to the human speech recognition accuracy, which is computed in terms of the word correct score (WCS). This is computed by dividing the number of correctly recognized keywords by the total number of keywords. Here, the WCS was averaged over ten randomly selected files. Similarly, the results of the objective measures, e.g. STOI, were also computed over the same ten files. In total, we obtained 20 groups of ten files at each SNR. The comparisons were performed between the WCS and the objective intelligibility measure computed for each group of ten files. Finally, these comparison values were averaged over all SNRs and were reported in Tables 1 and 2.

To perform the comparisons, it was first necessary to map the output of the instrumental measure to the domain of the listening test results. To achieve this, a logistic regression function was employed, as described in [6]. Three figures of merit, as suggested in [6, 5], were utilized for evaluation: the normalized cross correlation coefficient (NCC), the root mean square error (RMSE), and Kendall's Tau (τ).

We have analyzed the performance of three speech intelligibility prediction measures, namely the conventional STOI, the THMMB-STOI using oracle transcriptions, and the blind THMMB-STOI using the recognized transcription, in different noise conditions. For speech-shaped noise (SSN), the performance of the above metrics has additionally been evaluated based on the listening test results collected by Cooke et al. [16]. The results, in Tables 1 and 2, demonstrate that the THMMB-STOI with oracle transcriptions has a strong correlation with both listening test results. For the blind THMMB-STOI, slight decreases in accuracy were found relative to the STOI and the original THMMB-STOI.

Table 1: Comparison of objective speech intelligibility measures with listening test results (WCS), collected by Cooke et al. [16], for SSN in terms of NCC (%), RMSE, and τ (%).

Measure	NCC (%)	RMSE	τ (%)
STOI	93.20	0.095	73.94
THMMB-STOI (Oracle)	93.17	0.095	73.94
THMMB-STOI (Blind)	92.74	0.098	73.76

Table 2: Comparison of objective speech intelligibility measures with listening test results (WCS), collected by crowdsourcing, for SSN, Babble, and White noise in terms of NCC (%), RMSE, and Kendall's Tau (τ).

Noise	Measure	NCC (%)	RMSE	τ (%)
SSN	STOI	92.46	0.072	73.24
	THMMB-STOI (Oracle)	92.10	0.073	72.69
	THMMB-STOI (Blind)	92.43	0.071	72.67
Babble	STOI	94.29	0.073	77.05
	THMMB-STOI (Oracle)	93.71	0.076	74.87
	THMMB-STOI (Blind)	92.95	0.080	73.82
White	STOI	85.69	0.065	66.11
	THMMB-STOI (Oracle)	84.45	0.067	64.54
	THMMB-STOI (Blind)	83.71	0.069	63.73

However, it is evident that the blind THMMB-STOI is still in good agreement with human data, even though it did not use any extra information like the clean signal or the reference transcriptions. It can also be seen that both THMMB-STOI methods are performing well in comparison to the standard STOI measure in all noise types.

Overall, the THMM-based methods have successfully estimated the clean features for the STOI method, even in the blind version without a need for transcriptions. Furthermore, it was shown that both blind and oracle-based THMMB-STOI measures have a strong correlation with human intelligibility data in three different noise types.

4. Conclusions

In this paper, we have presented a new blind and non-intrusive approach for predicting the speech intelligibility, which is based on using THMMs and an automatically recognized transcription. Whereas the original THMM-based method needs the speech transcription for clean speech synthesis, here, it was proposed to use the THMMs for both automatically recognizing the transcriptions and synthesizing the clean features. The experimental results revealed that the proposed method has a good accuracy, leading to high correlations to human speech recognition results in different noise conditions. In comparison to the oracle-based THMMB-STOI and the STOI measures, the newly-proposed method had a slightly lower accuracy in predicting the speech intelligibility, but with added benefit of not requiring any extra information such as a clean reference signal or an oracle transcription. This is a significant advantage in comparison to the original THMM-based approach. It also has the potential to be integrated into the framework of other intrusive intelligibility prediction methods and hence provide an estimate of the clean reference features for them.

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] M. Karbasi, A. H. AbdelAziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," in *Proc. ICASSP*, Mar. 2016, pp. 624–628.

[2] N. French and J. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[3] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

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

[5] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 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," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, Jan. 2014.

[7] T. Fingscheidt and P. Bauer, "A phonetic reference paradigm for instrumental speech quality assessment of artificial speech bandwidth extension," in *Proc. 4th International Workshop on Perceptual Quality of Systems*, 2013.

[8] S. Nemala and M. Elhilali, "A joint acoustic and phonological approach to speech intelligibility assessment," in *Proc. ICASSP*, Mar. 2010, pp. 4742–4745.

[9] D. Sharma, G. Hilkhuysen, N. Gaubitch, P. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *Proc. EUSIPCO*, Aug. 2010, pp. 1899–1903.

[10] D. Sharma, P. Naylor, and M. Brookes, "Non-intrusive speech intelligibility assessment," in *Proc. EUSIPCO*, Sept. 2013, pp. 1–5.

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

[12] T. Falk, S. Cosentino, J. Santos, D. Suelzle, and V. Parsa, "Non-intrusive objective speech quality and intelligibility prediction for hearing instruments in complex listening environments," in *Proc. ICASSP*, May 2013, pp. 7820–7824.

[13] J. Santos, M. Senoussaoui, and T. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. IWAENC*, Sept. 2014, pp. 55–59.

[14] A. Abdelaziz, S. Zeiler, and D. Kolossa, "Twin-HMM-based audio-visual speech enhancement," in *Proc. ICASSP*, May 2013, pp. 3726–3730.

[15] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.

[16] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, vol. 49, no. 5, pp. 402–417, 2007.

[17] CrowdFlower, Inc, "CrowdFlower," as of February/March 2016, http://www.crowdflower.com.