NOISE-ROBUST AUTOMATIC SPEECH RECOGNITION WITH EXEMPLAR-BASED SPARSE REPRESENTATIONS USING MULTIPLE LENGTH ADAPTIVE DICTIONARIES

Emre Yılmaz, Jort F. Gemmeke, and Hugo Van hamme

Dept. ESAT, KU Leuven, Leuven, Belgium

ABSTRACT

In this work, we apply our recently proposed sparse representations based speech recognition system on the small vocabulary track of the 2nd 'CHiME' Speech Separation and Recognition Challenge. This system uses exemplars of different length to approximate noisy speech segments as a linear combination of the speech and noise exemplars with sparse weights. The exemplars are labeled speech segments extracted from the training data, each representing half words and they are organized in multiple dictionaries based on their class and length. A reconstruction error-based decoding is adopted to find the best matching class sequence. After the initial experiments on AURORA-2, we further apply our system on the CHIME data which is a more challenging task addressing not only non-stationary noise but also reverberation. Moreover, the structure of the CHIME data allows speaker-dependent acoustic modeling and sampling noise segments from the immediate acoustic context of the target utterances. Using speaker-dependent dictionary sets, several recognition experiments are conducted on the development and test sets to evaluate the system performance with different kinds of noise dictionaries. These experiments show that combined noise dictionaries containing noise exemplars extracted from both the immediate acoustic context of the test utterances and noise-only segments in the training data provide better recognition accuracies compared to fixed and adaptive dictionaries.

Index Terms— Exemplar-based recognition, sparse representations, non-negative sparse coding, multiple dictionaries

1. INTRODUCTION

The performance of automatic speech recognizers is reduced by non-stationary noise and reverberation in everyday applications. The degeneration of the spectro-temporal structure of speech signals due to reverberation highly depends on the characteristics of the enclosed acoustic space and the location of the speaker and recording device. Several front end approaches, e.g. linear filtering, feature and spectrum enhancement, and back end approaches, e.g. hidden Markov models (HMM) adaptation and acoustic context-dependent likelihood evaluation, have been proposed to mitigate the adverse effect of reverberation on the speech recognizers [1].

Considerable amount of research is devoted to tackle the reduced recognition accuracies due to non-stationary noise resulting in a number of approaches which can mainly be classified under robust feature extraction [2], signal and feature enhancement [3], model compensation [4] and missing data techniques [5, 6, 7]. These techniques are used together with HMM-based speech recognizers which are known to perform poorly in case of noise due to mismatches between the training and testing conditions.

As an alternative to HMM-based systems, there is an emerging interest in exemplar-based approaches, and several exemplar-based sparse representations (SR) techniques have been proposed in the last years for feature extraction [8], speech enhancement [9] and noise-robust speech recognition [10, 11, 12] tasks. These approaches model the acoustics using fixed length exemplars which are labeled speech and noise segments from the training corpus and stored in a single overcomplete dictionary. The noisy speech segments are jointly approximated as a sparse linear combination of these speech and noise exemplars with exemplar weights obtained by solving a regularized convex optimization problem. Enforcing sparsity results in only a very few exemplars with non-zero weights. Consequently, a realistic approximation of noisy speech segments are obtained without overfitting. The obtained weights mapped to HMM state likelihoods [10] and the noisy speech is decoded by applying the Viterbi algorithm.

We have recently proposed an alternative SR-based recognition system which uses different length exemplars organized in separate dictionaries based on their length and class (the associated speech unit) [13]. The input speech segments are approximated as a linear combination of the exemplars in each dictionary. Compared to a system using fixed-length exemplars stored in a single dictionary, using separate dictionaries for each class provides better classification as input speech segments are approximated as a combination of exemplars belonging to the same class only. Moreover, each exemplar is associated with a single speech unit and the natural duration distribution of each speech unit in the training data is preserved yielding exemplars of different lengths. A reconstruction error-based decoding is adopted, hence the system is capable of decoding unseen speech. Considering the sparse representation model using different length exemplars associated with a single class and the reconstruction error-based back end, this approach combines two exemplar-based speech recognition frameworks, i.e. exemplar matching-based recognition techniques [14, 15] and the aforementioned SR-based systems using a fixed length dictionary.

As an extension to the system described in [13], we performed noisy digit recognition using dictionaries which contain both speech and noise exemplars [16]. This noise-robust recognition system provided promising results on the AURORA-2 database [17] which contains clean and noisy digit utterances.

In this paper, we investigate the performance of this novel SR-based technique on the small vocabulary track of the 2nd 'CHiME' Speech Separation and Recognition Challenge. The provided data contains utterances recorded in a noisy living room and corrupted by both noise and reverberation. This recognition task is more challenging than the previous one performed on AURORA-2, due to the highly non-stationary room noise and the impact of reverberation on the spectro-temporal structure of speech. Clean, reverberated and noisy recordings from several speakers are provided in the training data which allow speaker-dependent acoustic modeling. Furthermore, the target utterances are provided both in isolated and embedded forms, the latter revealing useful information about the

immediate acoustic context of each utterance. This information allows 'noise sniffing' during the recognition, i.e. on-the-fly extraction of noise exemplars that lie temporally close to the target utterances [18, 19]. We have compared the performance of our system using fixed and adaptive noise dictionaries which use noise exemplars extracted from the embedded training data and from the neighborhood of every target utterance respectively. A third kind of noise dictionary which combines exemplars from fixed and adaptive dictionaries is also evaluated.

The rest of the paper is organized as follows. The exemplar-based sparse representations system using multiple dictionaries is explained in Section 2. The evaluation setup and implementation details are discussed in Section 3. Section 4 presents the recognition results and comments on the system performance. In Section 5, the conclusions and thoughts for future work are discussed.

2. SPARSE REPRESENTATIONS WITH MULTIPLE LENGTH EXEMPLARS

2.1. Noisy speech model

The noise-robust recognizer described in [16] models the noisy speech segments as a sparse linear combination of speech and noise exemplars that are stored in multiple dictionaries. Speech exemplars spanning multiple frames are reshaped into a single vector and stored in the columns of a speech dictionary $\mathbf{S}_{c,l}$: one for each class c and each length l. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where D is the number of Mel frequency bands in a frame and $N_{c,l}$ is the number of available speech exemplars of length l and class c. Similarly, a single noise dictionary $\mathbf{N}_l$ for each length l is formed by reshaping the noise exemplars. Each speech dictionary is concatenated with the noise dictionary of the same length to form a single dictionary $\mathbf{A}_{c,l} = [\mathbf{S}_{c,l}\, \mathbf{N}_l]$ of dimensionality $Dl \times M_{c,l}$ where $M_{c,l}$ is the total number of available speech and noise exemplars. For any class c, a reshaped noisy speech vector $\mathbf{y}_l$ of length Dl is expressed as a linear combination of the exemplars with non-negative weights:

$$\mathbf{y}_l \approx \sum_{m=1}^{M_{c,l}} x_{c,l}^m \mathbf{a}_{c,l}^m = \mathbf{A}_{c,l}\mathbf{x}_{c,l} \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \qquad (1)$$

where $\mathbf{x}_{c,l}$ is an $M_{c,l}$-dimensional sparse weight vector. A sparse weight vector implies that the noisy speech is approximated by a few exemplars from the speech and/or noise dictionaries.

2.2. Obtaining the exemplar weights

The exemplar weights are obtained by minimizing the cost function,

$$d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l}) + \sum_{m=1}^{M_{c,l}} x_{c,l}^m \Lambda_m \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \qquad (2)$$

where $\mathbf{\Lambda}$ is an $M_{c,l}$-dimensional vector which contains non-negative values and controls how sparse the resulting vector $\mathbf{x}$ is. Defining $\mathbf{\Lambda}$ as a vector, the amount of sparsity enforced on different types of exemplars can be adjusted. The first term is the divergence between the noisy speech vector and its approximation. The second term is a regularization term which penalizes the l_1-norm of the weight vector to produce a sparse solution. The generalized Kullback-Leibler divergence (KLD) is used for d:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k \qquad (3)$$

The regularized convex optimization problem can be solved by applying non-negative sparse coding (NSC). For NSC, the multiplicative update rule to minimize the cost function (2) is derived in [10] and is given by

$$\mathbf{x}_{c,l} \leftarrow \mathbf{x}_{c,l} \odot (\mathbf{A}_{c,l}^T(\mathbf{y}_l \oslash (\mathbf{A}_{c,l}\mathbf{x}_{c,l}))) \oslash (\mathbf{A}_{c,l}^T\mathbf{1} + \mathbf{\Lambda}) \qquad (4)$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}$ is a Dl-dimensional vector with all elements equal to unity. Applying this update rule iteratively, the weight vector becomes sparser and the reconstruction error between the noisy speech vector and its approximation decreases monotonically.

2.3. Decoding the noisy speech

The first term of Equation (2) expresses the reconstruction error between a noisy speech segment of length l and a class c. Every noisy speech segment of each available exemplar length is approximated as a linear combination of exemplars. This is achieved by applying the sliding window approach [10] to the noisy utterance for each available exemplar length and iteratively applying Equation (4) using the dictionaries containing exemplars of the corresponding length. After a fixed number of iterations, the reconstruction error is calculated. As the label of each dictionary is known, decoding is performed by applying dynamic programming (taking the grammar into account) to find the class sequence that minimizes the reconstruction error.

3. EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

3.1. Database

The small vocabulary track of the 2nd 'CHiME' Speech Separation and Recognition Challenge [20] addresses the problem of recognizing commands in a noisy living room. The clean utterances are taken from the GRID corpus [21] which contains utterances from 34 speakers reading 6-word sequences of the form *command-color-preposition-letter-digit-adverb*. There are 25 different letters, 10 different digits and 4 different alternatives for each of the other classes. The recognition accuracy of a system is calculated based on the correctly recognized letter and digit keywords.

The clean utterances are artificially reverberated using binaural room impulse responses which include the speaker head movement effects. Then, they are mixed with binaural recordings of genuine room noise at SNR levels of 9, 6, 3, 0 ,-3 and -6 dB. The training set contains 500 utterances per speaker (17,000 utterances in total) with clean, reverberated and noisy versions. Noisy utterances are provided both in isolated or embedded form. Embedded recordings contain 5 seconds of background noise before and after the target utterance. The development and test sets contain 600 utterances from all speakers at each SNR level (3600 utterances in total for each set) both in isolated and embedded from. The immediate noise context of the target utterances are available in the embedded recordings. The development set also contains 600 noise-free reverberated utterances. All data has a sampling frequency of 16 kHz.

3.2. Exemplar extraction and dictionary creation

The exemplars and noisy speech segments are represented as Mel-scaled magnitude spectral features extracted with a 26 channel Mel-scaled filter bank ($D = 26$). The frame length is 25 ms and the frame shift is 10 ms. Binaural data is averaged in the spectral domain to obtain 26-dimensional feature vectors.

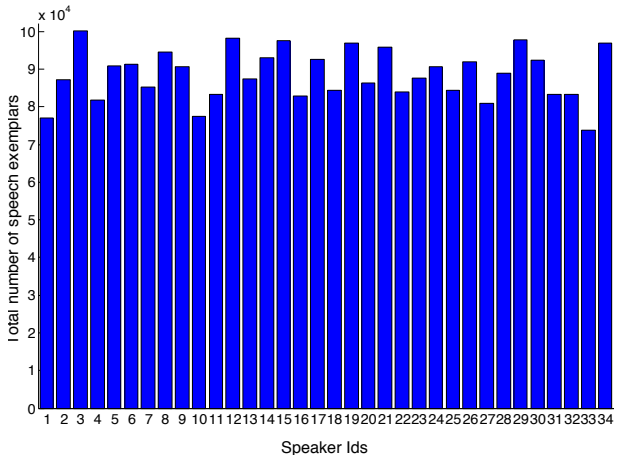

Fig. 1: Total number of speech exemplars for each speaker

The exemplars are extracted from the reverberated utterances in the training set according to the state-based segmentations obtained using the clean acoustic models provided in the toolkit. Exemplars belonging to each speaker are organized in separate dictionary sets for speaker-dependent modeling yielding 34 different dictionary sets. The minimum and maximum exemplar lengths are 2 and 40 frames respectively. Exemplars longer than 40 frames are omitted to limit the number of dictionaries. The usage of very short exemplars is viable due to the existence of a strict grammar and they are indeed observed to be useful during the recognition. Exemplars representing full words turned out to provide poor acoustic modeling in terms of generalizability resulting in a high error rate. Half word exemplars seemed to generalize sufficiently to unseen data under the condition of applying so-called *prewarping*, i.e. removing a small number of frames, except the very first and last frame, from an exemplar of length l to obtain shorter exemplars of length $l_{\text{new}} < l$. Due to the high number of alternatives and hence the small number of exemplars per word, speaker-dependent modeling of letters results in low recognition accuracies even after applying prewarping. Hence, letter exemplars from all speakers are used in all 34 dictionary sets. Dictionary sizes vary with class, but are limited to 200. The total number of speech exemplars for each speaker that are used during the recognition is given in Figure 1.

The silences between the words are assumed to be negligible, hence, dictionaries representing a silence class are not used. This comes with several advantages as the reconstruction error scores obtained using silence dictionaries have to be compensated [16]. However, the isolated utterances in the training, development and test sets contain a different number of silence frames in the beginning of the utterances. To overcome the problems that may occur during the decoding, the number of silence frames in the beginning of the reverberated training data is limited to 10 frames while extracting the exemplars. Furthermore, during the recognition, the decoding is repeated 5 times each time omitting 5 frames from the beginning. The class sequence yielding the minimum reconstruction error per frame is then chosen to be the recognition output.

3.3. Noise dictionaries

We used three different noise dictionaries during the experiments, i.e. fixed, adaptive and finally a combination of these two noise dictionaries. Fixed noise dictionaries contain noise exemplars which are extracted from the embedded recordings in the training set. Since there are large amounts of noise segments available in the embedded training data, only the segments with high energy are selected according to an l_1-norm based criterion in order to eliminate silences. After a collection of noise exemplars for each available speech exemplar length is obtained, 200 noise exemplars are randomly selected and concatenated to the speech dictionaries of the corresponding length.

Adaptive dictionaries contain 200 noise exemplars that are extracted from the neighborhood of each target utterance in both directions until the frames belonging to other target utterances. In this way, the mismatch between the actual noise segments corrupting the target utterance and the noise exemplars in the dictionary is reduced [19]. On the other hand, only sampling noise segments from the neighborhood of the target utterance may limit the diversity of spectrographic content of the noise exemplars. In case of highly non-stationary noise, having a more spectrally diverse noise dictionary might give better separation. Therefore, we use a third kind of dictionary which contains noise exemplars randomly taken from both fixed and adaptive dictionaries. The combined noise dictionaries contain 50 noise exemplars from the fixed dictionaries and 150 noise exemplars from the adaptive dictionaries.

Table 1: Recognition accuracies obtained on the development and test sets. The baseline results are obtained using the acoustic models trained on noisy data.

(a) Development Set

SNR(dB)	-6	-3	0	3	6	9
Baseline	49.67	57.92	67.83	73.67	80.75	82.67
Fixed	50.42	55.42	66.33	77.58	82.42	88.25
Adaptive	54.58	61.33	70.75	**80.83**	**86.00**	**90.00**
Combined	**57.17**	**63.08**	**71.42**	80.08	85.17	**90.00**

(b) Test Set

SNR(dB)	-6	-3	0	3	6	9
Baseline	49.33	58.67	67.50	75.08	78.83	82.92
Fixed	49.25	54.58	65.67	75.42	82.67	87.58
Adaptive	52.08	61.75	71.83	**79.75**	**85.08**	89.42
Combined	**55.92**	**61.92**	**72.83**	**79.75**	84.33	**89.75**

3.4. Implementation details

The whole system is implemented in MATLAB and we used GPUs to accelerate the evaluation of Equation (4). The multiplicative update rule is iterated 50 times to find the exemplar weights. Elements of $\mathbf{\Lambda}$ corresponding to speech exemplars are set to 0.45, and the ones corresponding to noise exemplars are set to 0.3. The l_2-norm of dictionary columns and reshaped noisy speech vectors are normalized to unity.

4. RESULTS

We conducted recognition experiments on the development and test data to evaluate the recognition accuracies using the different noise dictionaries discussed in Section 3.3. The baseline recognition accu-

racies provided at the CHIME website[1] are obtained using a GMM-based speaker-dependent speech recognizer trained on noisy data. The recognition accuracies obtained using fixed, adaptive and combined noise dictionaries are given in Table 1. The best results at each SNR level are given in bold.

The best recognition accuracies, especially at lower SNR levels, are obtained using the combined noise dictionaries which contains noise exemplars from both fixed and adaptive dictionaries. At SNR level of -6 dB, the recognition accuracy obtained on the test data using the combined noise dictionaries is 55.92% compared to 49.25% and 52.08% of the fixed and adaptive noise dictionaries respectively. This significant improvement is mainly due to the limited number of noise exemplars available in the dictionaries. Using only 200 noise exemplars, a noise dictionary containing exemplars with both more diverse spectrographic representations and noise samples from the preceding and following segments of the target utterances provide more effective noise modeling. At higher SNR levels, the adaptive and combined dictionaries provide comparable results.

At all SNR levels, the system using fixed noise dictionaries perform worse than the others which is consistent with the previously reported results. The main reason of these inferior results is the mismatch between the actual noise frames and noise exemplars extracted from the training data due to the highly non-stationary noise. Using adaptive noise dictionaries, which contain noise exemplars that are extracted from the noise segments next to the target utterance, results in a more accurate recognition compared to the fixed noise dictionaries.

The overall performance of the proposed system, which is slightly better than the challenge baseline, mainly suffers from the limited noise modeling. The computational restrictions on using more noise exemplars can be removed by designing an efficient implementation of the system with reduced computational complexity.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have evaluated the performance of a sparse representations-based speech recognition system using multiple length exemplars on the small vocabulary track of the 2nd 'CHiME' Speech Separation and Recognition Challenge. The recognition is performed using half word exemplars organized in separate dictionaries on the basis of their length and class. Moreover, for each speaker, a unique dictionary set containing exemplars belonging to that speaker is used. The impact of using fixed, adaptive and combined noise dictionaries on the recognition accuracy is investigated during the experiments. The combined noise dictionaries which contains noise exemplars from both the neighborhood of the target utterance and a global noise exemplar collection have been shown to provide the best recognition accuracies at lower SNR levels. Considering the limited noise modeling due to computational restrictions, the combined noise dictionaries are more effective than adaptive dictionaries due to increased spectral diversity of noise exemplars and providing better noise modeling than fixed dictionaries since they have noise samples from the immediate acoustic context of the target utterances. The recognition accuracies obtained using the combined noise dictionaries have shown the feasibility of this novel noise-robust speech recognition technique.

Using a few hundreds of noise exemplars compared to the thousands of the techniques using fixed length exemplars organized in a single dictionary yields a lower recognition accuracy especially at lower SNR levels. There are several ways to reduce the computational limitations on the noise modeling. Firstly, multiplicative update iterations, which are the computational bottleneck of the whole system, can be efficiently evaluated by a designated implementation of speech and noise dictionaries. Another approach is to use a shared noise dictionary with the speech dictionaries of the same length. Finally, choosing the most informative frequency bands in an exemplar results in a reduced dictionary size. Experimenting with such efficient implementations remains as a future work.

[1] http://spandh.dcs.shef.ac.uk/chime_challenge/chime2_task1.html

6. ACKNOWLEDGEMENTS

This work has been supported by the KU Leuven research grant OT/09/028 (VASI) and IWT-SBO Project 100049 (ALADIN).

7. REFERENCES

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114 –126, Nov. 2012.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[3] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. ICASSP*, May 1996, vol. 2, pp. 733 –736 vol. 2.

[4] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352 –359, Sept. 1996.

[5] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267 – 285, 2001.

[6] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101 – 116, Sept. 2005.

[7] M. Van Segbroeck and H. Van hamme, "Advances in missing feature techniques for robust large-vocabulary continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 123 –137, Jan. 2011.

[8] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representations features for speech recognition," in *Proc. INTERSPEECH*, Sept. 2010, pp. 2254–2257.

[9] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based speech enhancement and its application to noise-robust automatic speech recognition," in *International Workshop on Machine Listening in Multisource Environments*, Sept. 2011, pp. 53–75.

[10] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067 –2080, Sept. 2011.

[11] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *Proc. ICASSP*, May 2011, pp. 4588 –4591.

[12] Q. F. Tan and S. S. Narayanan, "Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1337 –1346, May 2012.

[13] E. Yılmaz, D. Van Compernolle, and H. Van hamme, "Combining exemplar-based matching and exemplar-based sparse representations of speech," in *Symposium on Machine Learning in Speech and Language Processing (MLSLP)*, Portland, OR, USA, Sept. 2012.

[14] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1377 –1390, May 2007.

[15] L. Golipour and D. O'Shaughnessy, "Context-independent phoneme recognition using a k-nearest neighbour classification approach," in *Proc. ICASSP*, Apr. 2009, pp. 1341 –1344.

[16] E. Yılmaz, J. F. Gemmeke, D. Van Compernolle, and H. Van hamme, "Noise-robust digit recognition with exemplar-based sparse representations of variable length," in *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sept. 2012.

[17] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Tutorial and Research Workshop ASR2000*, Sept. 2000, pp. 181–188.

[18] J. F. Gemmeke and H. Van hamme, "Advances in noise robust digit recognition using hybrid exemplar-based techniques," in *Proc. INTERSPEECH*, 2012, pp. 1–4.

[19] A. Hurmalainen, J. F. Gemmeke, and Virtanen T., "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech & Language*, vol. 27, no. 3, pp. 763–779, 2012.

[20] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[21] 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. 120, no. 5, pp. 2421–2424, 2006.