A fragment-decoding plus missing-data imputation ASR system evaluated on the 2nd CHiME Challenge

Ning Ma

MRC Institute of Hearing Research,
Nottingham, NG7 2RD, UK
n.ma@ihr.mrc.ac.uk

Jon Barker

Department of Computer Science,
University of Sheffield,
Sheffield, S1 4DP, UK
j.barker@dcs.shef.ac.uk

Abstract

This paper reports on our entry to the small-vocabulary, moving-talker track of the 2nd CHiME challenge. The system we employ is based on the one that we developed for the 1st CHiME challenge, the latest results of which are reported in (Ma and Barker, 2012). Our motivation is to benchmark the system on the new CHiME challenge and to measure the extent to which it is robust against speaker motion, a feature of the second challenge that was absent in the first. The paper presents a brief overview of our fragment-decoding plus missing-data imputation system and then makes a component-by-component analysis of the system performance on both the 1st and 2nd CHiME challenge datasets. We conclude that due to its reliance on pitch and spectral cues the system is robust against the introduction of small speaker motions. We achieve an average keyword recognition score of **85.9%** compared to 86.3% for the stationary speaker condition.

Index Terms: Missing feature imputation, noise-robust speech recognition, mask estimation.

1. Introduction

For automatic speech recognition (ASR) to work reliably it is typically necessary for the speech signal to be free from interference from competing noise sources and, ideally, free from the distorting effects of reverberation. These conditions are usually ensured by employing a microphone that is close to the mouth of the speaker. For example, ASR systems work well with head-mounted microphones, mobile device held up to the face, and to a less extent with lapel microphones. However, for a wide range of applications these 'close-talking microphone' configurations are artificial and inhibit natural communication. There has therefore been growing interest in the more challenging 'distant microphone' scenario [1, 2].

In 2011 the 1st CHiME challenge was organised to promote research into robust automatic speech recognition in distant microphone settings [3]. The challenge employed command sentences from a small vocabulary corpus reverberantly mixed into 'multisource' background noise recordings collected using a binaural manikin situated in a domestic living room. The challenge attracted entries from 13 teams, a representative sample of which are reported in a recent Special Issue of Speech Communication [3]. However, a limitation of this original challenge was that the target talker was mixed into the backgrounds using a constant binaural room impulse response (measured 2 m in front of the manikin) and hence the task failed to model the variability in the receiver-source geometry that would be observed in a real application scenario (e.g. variabilty due to speaker motion). The new, 2nd CHiME challenge, that is being considered in this paper relaxes this assumption. The talker is still assumed to be standing in a 'sweet spot' at a position 2 m in front of the manikin but the talker now has the freedom to make small head movements within a region of 20 cm by 20 cm around this location. This has been modelled by selecting random start and end locations for each utterance and interpolating between impulse responses measured on a find grid in the room. Full details of the challenge construction are provided in [4].

For the original CHiME challenge we developed a system based on a combination of spectro-temporal fragment decoding plus missing data imputation. Results of this system are published in [5] and [6]. The purpose of this current paper is to re-evaluate this system on the new challenge in order to assess how well it copes with the increased difficulty of the more realistic mixing conditions. In particular, we break the system down into a number of components and directly compare the gain bought by each component on the stationary-speaker 1st CHiME challenge (CHiME-1) and the moving-speaker 2nd CHiME challenge (CHiME-2). The paper is not intended to introduce original techniques but rather to serve as a benchmark for our existing system on a new dataset and to provide some insight into the robustness of our previously reported results.

Section 2 will provide an overview of the fragment-

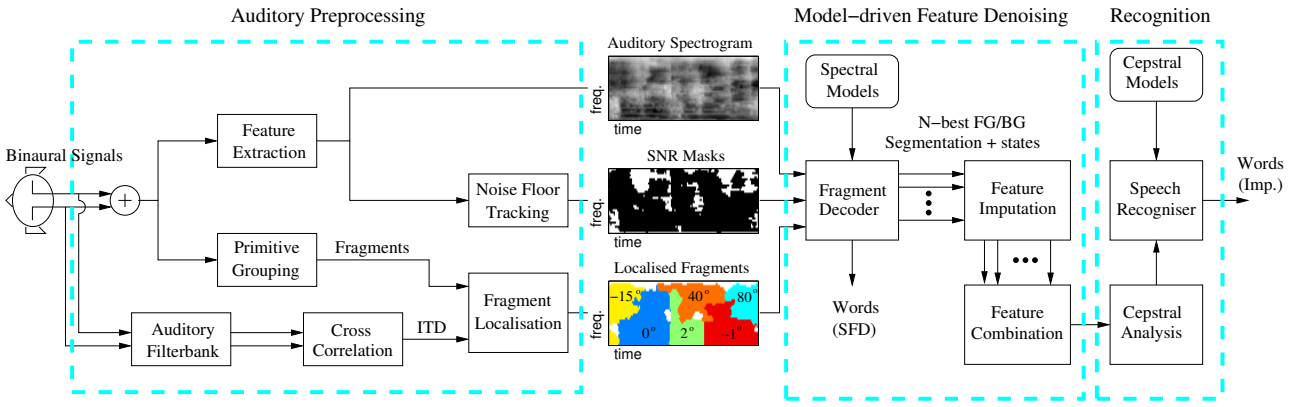

Figure 1: Overview of the fragment-decoding plus missing-data imputation ASR system.

decoding and imputation system. This has been kept deliberately brief and non-technical because a detailed presentation can be found in [5]. Section 3 describes the experiments that have been run on the new challenge. Changes that have been necessary to tune the system to the new data set are discussed. Comparative results are discussed in Section 4 and an attempt is made to provide explanation for differences in system behaviour on the two tasks. Finally, we put the performance of the system into the context of previously reported CHiME systems and discuss the potential impact of the work.

2. System Overview

For the 2nd CHiME challenge we have re-employed the system developed for the 1st CHiME challenge described in detail in [5] and used an improved imputation algorithm described in [6]. The system is illustrated in Figure 1 and described in overview here. For fuller technical details the reader is referred to the earlier papers [5, 6].

The system can be described as having three stages: an 'auditory' pre-processing stage that operates on the binaural acoustic signals and generates a set of spectro-temporal representations. This is followed by a model-based spectral-temporal feature denoising stage driven by a process called 'fragment decoding'. The heart of this stage is a speech recognition pass working in the spectral domain and recognition output can be evaluated directly at this point. De-noised spectral features are then transformed into the cepstral domain and processed using a conventional speech recogniser (the 2nd recognition pass). These three stages are described in the sections that follow.

2.1. Auditory pre-processing

The front-end processing computes three representations of the signal that are required for the denoising stage,

i) **The auditory spectrogram** This is the basic representation used to train models for the 1st recognition pass. The left and right channels are summed and passed through a Gammatone filterbank. The log magnitude of the filterbank outputs are smoothed and sampled at a 100 Hz frame rate to form a spectro-temporal representation (an 'auditory spectrogram'). Note that by summing the left and right channels we are taking advantage of the fact that the target source is known to come from a direction roughly directly in front on the manikin, i.e. a simple beam-forming.

ii) **The noise floor SNR Mask** This is a binary mask which estimates the spectral temporal regions where 'signal' dominates a quasi-stationary noise floor. Similar masks have formed the basis of many previous missing data ASR systems (e.g. [7,8]). The noise floor is estimated using a technique based on the minimum tracking-based methods popularly used in speech enhancement [9, 10]. Our implementation fits a slowly time-varying GMM-based noise floor model to the energy minima observed in the noisy auditory spectrogram. The mask is then computed by comparing the noise floor estimate and the noisy auditory spectrogram: regions that lie above the noise floor estimate (i.e. that are above 0 dB SNR) are labelled as dominated by 'signal'. Note, these regions are not necessarily dominated by the target speech signal, they are just not masked by the noise floor.

iii) **The localised fragments** The 'fragments' are spectro-temporal regions that are believed to be dominated by a single environmental sound source. They are generated by a 'primitive grouping' module that first uses multi-pitch analysis to track the pitch of multiple harmonic sound sources through time. The pitch estimates at each time frame are then used to bind Gammatone filterbank channels across frequency. Finally a simple image-segmentation algorithm operates on remaining regions in order to

isolate any energy peaks in the auditory spectrogram that have not yet been accounted for, i.e. the regions dominated by non-periodic energy (e.g. fricative speech regions). A 'fragment localisation' module then uses both the left and right binaural signals to estimate an azimuthal direction for each fragment. The directions are obtained by averaging interaural time difference (ITD) estimates for each time-frequency element within the fragment.

2.2. Model-based spectro-temporal feature denoising

The core of the feature denoising block is a 'fragment decoder'. This decoder is an extension of the missing-data approach to ASR [11]. Missing data ASR systems take noisy spectro-temporal representations and a 'mask' indicating which spectro-temporal elements are reliable. In contrast, the fragment decoder takes a set of fragments and then considers all masks that can be generated by the various foreground (i.e. reliable) versus background (i.e. masked) labellings of the fragments. The decoder simultaneously searches for the fragment labeling and speech state sequence that best matches the noisy data to a set of clean speech models.

We employ a couple of extensions to the basic fragment decoding approach. First, regions that are dominated by noise according to the noise floor SNR mask are constrained to be labeled as background. This means that the fragment decoder is only making foreground/background decisions about regions that stand clear of the noise floor. Second, the fragment location estimates are used to bias the fragment decoder against labelling fragments as foreground if they appear to come from a direction that is too far from 0 degrees (i.e. because in the CHiME scenario the talker is known to be standing approximately directly in front of the binaural manikin). If the location estimates were reliable then this bias could be very strong, e.g. any fragment originating from outside a narrow beam around 0 degrees could be reliably labelled as part of the background. However, room reverberation makes the location estimates very unreliable – even allowing for the fact that ITD estimates are integrated over complete fragments – so a small empirically-derived bias is used that allows the foreground/background decision to be dominated by the goodness of the fragment's match to the clean speech models.

The decoder outputs a speech model state sequence and a foreground/background segmentation that are employed in the denoising stage. The spectro-temporal features in the foreground region are those that are dominated by the target speech source and are at a favourable local SNR. These features remain unchanged. The features in the masked regions are dominated by noise. These are denoised by replacement with MMSE estimates of the noise-free speech derived from the clean speech models [12], and specifically from the model state that the decoding process has aligned to the frame being denoised. Of course, if the decoder has estimated the model sequence incorrectly the imputed estimates will be incorrect. In [6] we found that this problem could be reduced by using the N best decodings to form multiple estimates and then averaging the estimates weighted by decoding confidence measures.

2.3. Speech recognition

In the final stage a DCT-transform is employed to convert the reconstructed spectral features into a set of 13 features in the cepstral domain, i.e. Gammatone filterbank cepstral coefficients (GFCCs). Delta and delta-delta features are added to form a 39-dimensional feature vector. Word-level HMMs with the same structure as those of CHiME challenge baseline system are trained using the training data sets specified by the challenge: a reverberated but noise-free set and a noise-added set.

3. Experiments and Results

3.1. Experimental setup and system tuning

The CHiME-1 and CHiME-2 challenges use identical training and test sets and differ only in the manner that the speech and background are mixed (stationary speaker vs. moving speaker). The similarity of the two challenges allows results to be directly compared. Further, for both CHiME-1 and CHiME-2, we employ the same HMM set up and training regime as employed in the CHiME baseline system: in particular we use word-based HMMs and train 34 speaker dependent models matched to the talkers in the CHiME test set.

The configuration of the recognition system was almost exactly the same as for the CHiME-1 evaluation as described in [5] (and the N-best extension in [6]). The remainder of this section details notable differences.

Sampling rate In the previous CHiME challenge the data was distributed at 48 kHz sampling rate and our filterbank was designed with 32 filter channels with filter centre frequencies evenly spaced between 50 Hz and 8 kHz on an equivalent rectangular bandwidth scale [13]. The current challenge data was distributed at 16 kHz. We wished to use an identical representation, so in order to avoid aliasing in the highest frequencies band, the signals were first upsampled to match the 48 kHz rate of the earlier challenge.

Fragment localisation As discussed earlier the fragment decoding system is able to use a fragment location estimate to bias the labelling of the fragment towards either being foreground or background. In the previous system this was achieved by tuning three parameters: an azimuth threshold, T; a foreground/background bias for 'lateral' fragments, (i.e. those with absolute azimuth estimates *greater* than the threshold) expressed as proba-

bility P_l; and a foreground/background bias for 'central' fragments (i.e. those with absolute azimuth estimates *less* than the threshold), expressed as probability P_c. The azimuth threshold was tuned by first using knowledge of the premixed speech and background signals to correctly assign the foreground/background label to a set of candidate fragments. Then the histograms of the estimated azimuths for each class were examined. It was seen that fragments labeled with absolute azimuths greater than 18 degrees were mostly coming from competing sources. Once this 18 degree threshold was selected, P_l and P_c were tuned empirically by running experiments on the development test set. For the new data the same analysis was performed and it was seen that the distributions of estimated fragment azimuths for the foreground and background classes were less divergent. Many foreground fragments had very large azimuth estimates. This is possibly due to the fact that, i) the speech target motion makes the fragments harder to localise reliably, and ii) the lower sampling rate reduces the accuracy of the ITD estimates from which the localisation estimates are derived. When using the previous parameters the localisation information failed to improve recognition performance. Widening the threshold to 30 degrees and retuning P_l and P_c allowed localisation cues to once again confer a modest benefit (see next section).

N-best decoding In [6] we reported improved results using smoothed imputations constructed by taking a weighted average of individual imputations obtained from the N-best speech fragment decodings. Our earlier experiments on the CHiME-1 development test set showed the optimal value of N to be 5. Using the same value on the current task proved to provide no benefit to the development test set performance. A value of 3 provided a better result and was therefore used for the final system evaluation on the test set.

3.2. Results and analysis

Table 1 presents results for variations of the system evaluated on the CHiME-2 *development* test set, and tables 2 and 3 compare *final* test set performance for CHiME-2 and the earlier CHiME-1 respectively. All figures represent keyword accuracies as required by the challenge protocol (see [4]). The result labeled **MFCC** is the recognition performance obtained using the baseline non-robust recognition system that is distributed with the challenge data. (Although the baseline system is not designed to be robust it operates on features extracted from the sum of left and right channels and hence noise from lateral directions is somewhat suppressed, i.e. by beam-forming). Note that when averaged across conditions the baseline performance is nearly 2% greater for CHiME-2. In CHiME-1 the target talker is stationary and a constant impulse is used for mixing the data sets but *different* recordings of the 2 m 0 degree impulse response were used to prepare the training set data and test set data. Differences between the two impulse responses introduce a small amount of model mismatch. In contrast, in CHiME-2 the talker makes small movements that are simulated by interpolating between pairs of impulse responses measured at positions chosen within a small area directly ahead of the recording manikin. Although the movement makes the task more challenging (i.e. it is harder to use spatial filtering to separate the target and masker), the impulse response statistics are matched across the training and test sets and the variability in impulse response observed in the training data introduces some robustness against small changes in the impulse responses observed in the test data.

Table 1: Keyword accuracies (%) using different system configurations for the CHiME-2 development test set.

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	ave.
SFD	72.58	74.75	79.83	85.42	87.83	88.58	81.50
+NF	73.75	75.67	81.33	86.08	87.58	88.92	82.22
+Loc	75.08	75.67	82.08	86.00	88.00	89.08	82.65
+Loc+NF	75.42	77.58	83.58	86.92	89.08	90.17	83.79
Imp.1	69.83	74.75	81.00	84.42	89.17	89.83	81.50
Imp.1 MC	76.58	78.33	84.00	87.83	90.50	90.83	84.68
Imp.3 MC	**77.00**	**78.25**	**84.33**	**87.67**	**89.75**	**91.17**	**84.70**

Table 2: Keyword accuracies (%) using different system configurations for the CHiME-2 evaluation test set.

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	ave.
MFCC	32.17	38.33	52.08	62.67	76.08	83.83	57.52
SFD	72.25	75.33	80.92	84.58	87.17	89.00	81.54
+NF	74.58	77.75	83.17	85.33	88.67	90.17	83.28
+loc	74.00	76.83	81.50	85.00	88.17	89.83	82.56
+NF+loc	76.50	78.25	84.33	86.25	89.00	89.75	84.01
Imp.1	72.50	74.67	83.08	85.83	88.92	91.00	82.67
Imp.1 MC	77.25	79.92	85.50	88.92	90.42	92.17	85.70
Imp.3 MC	**77.75**	**80.08**	**85.25**	**88.83**	**91.08**	**92.50**	**85.92**

Table 3: Keyword accuracies (%) using different system configurations for the CHiME-1 evaluation test set.

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	ave.
MFCC	30.33	35.42	49.50	62.92	75.00	82.42	55.93
SFD	71.75	72.75	78.75	82.83	85.08	87.25	79.74
+NF	74.17	76.33	81.25	85.00	86.92	87.00	81.78
+loc	74.00	75.83	81.33	85.33	87.50	88.67	82.11
+NF+loc	76.00	78.00	83.17	86.08	88.33	88.42	83.33
Imp.1 MC	78.08	80.58	85.75	88.08	91.00	91.50	85.83
Imp.3 MC	**78.50**	**81.25**	**85.58**	**88.25**	**91.08**	**93.33**	**86.33**

The result labeled **SFD** is the output of the baseline speech fragment decoding system, i.e. without use of adaptive noise flooring, fragment localisation, or spectral imputation. Using SFD, the average CHiME-2 final test set performance is increased from 57.5% to 81.5%. Note that the better performance of the CHiME-2 baseline with respect to CHiME-1 is also reflected in the CHiME-1 and CHiME-2 SFD results. Introducing the adaptive noise floor component (**+NF**) improves performance for CHiME-2 by 1.7%. This is comparable to the 2.0% improvement that the adaptive noise floor brought to the

CHiME-1 evaluation. In contrast, using the fragment localisation component (**+loc**) which previously produced a 2.4% improvement is now only earning an additional 1.0%. This is not surprising given the decreased discriminability between the azimuth estimates of foreground and background fragments and the widening of the lateral fragment rejection threshold discussed earlier. As found in our previous work, the adaptive noise floor and localisation systems can be combined (**+NF+loc**), and doing so provided a total performance increase of 2.5% over the SFD baseline for CHiME-2 compared to 3.6% for CHiME-1.

The SFD system **+NF+loc** was used to provide state sequence and mask estimates to drive the imputation system. Decoding cepstral transforms of the imputed masks using models training on the reverberated noise-free data, **Imp.1**, led to a *drop* in performance of 1.5%. This can be explained by a failure of the SFD denoising to remove all fragments of noise. Ideally this mismatch should be avoided in the final pass by employing models trained on a denoised version of the noise-added training data. Here though we followed the approach we used previously, i.e. we increased robustness by retraining the models using a multiconidition training set. The multicondition training set was made by combining the supplied noise-free training set and the noise-added training set (**Imp.1 MC**). Using the new models produced an increase in performance of 3.0% relative to using the noise-free models and an improvement of 1.7% over the **+NF+loc** system. The same step in the previous evaluation led to an improvement of 2.5% over **+NF+loc**. The difference can perhaps be attributed to differences in the multicondition training data set. For CHiME-1, training data was mixed at the target SNR levels employed in the test sets, i.e. -6 dB to +9 dB. For CHiME-2, the supplied noise-added training set consists of utterances that are mixed at random locations in the CHiME noise background with no regard to the SNRs produced.

Finally, combining N-best imputations provided a disappointing 0.2% improvement compared to 0.5% improvement for the CHiME-1 task. Note, tuning of N on the development test set led to an optimum of 5 previously and 3 for the new data. The 0.2% does not represent a statistically meaningful improvement. It is unclear why the N-best decoding technique has failed to convey an advantage on the new task, perhaps the greater variability in the models caused by the variable target position reduces the impact of mismatch due to poor imputation and hence lessens the advantage of averaging over N-best imputations.

The overall result for the final system is 85.92% for the moving talker CHiME task (CHiME-2) compared to 86.33% for the stationary task (CHiME-1).

4. Discussion and Conclusions

Despite the fact that the CHiME-2 challenges introduces speaker motion as an additional source of target speech variability our overall performance is only reduced by 0.4%. The fact that the performance drop is small is largely due to the fact that the system is making small advantage of spatial filtering in the first place. Comparing the +NF and +NF+loc the additional benefit of fragment localisation was 1.5% and reduced to 0.7% in the current challenge.

It may be noted that our overall system performance is somewhat below that of the very best performances previously reported for competing systems on the 1st CHiME challenge – some of which approached human speech recognition performance (e.g. [14, 15]). However, in contrast to these highly optimised systems, our system is using a comparatively simple 'back-end'. In fact, the final recognition pass uses nothing more than the somewhat naive training and recognition set-up of the baseline CHiME recogniser. Once the features have been denoised by the fragment-decoding and imputation stage there are a multitude of conventional techniques that can be applied to further increase performance, e.g. model optimisations such as state-clustering, discriminative model training, more sophisticated speaker adaptation, supervised and unsupervised noise adaptation, and robust decoding strategies such as dynamic variance adaptation or uncertainty decoding (back-end techniques that have been applied with success by other CHiME challenge systems [3]). In particular, the 2nd pass recognition models should be retrained on data that has been processed by the denoising stage to reduce potential mismatch between denoised and noise-free speech.

Finally, it may be argued that because our system employs an unconventional denoising strategy that relies heavily on multi-pitch tracking, auditory representations and fragment decoding, the system may have quite different strengths and vulnerabilities in comparison with more established techniques. In this respect the denoised features that the system generates may be suitable candidates for including in multistream systems that take advantage of feature complementarity (e.g. [15]). This possibility will be open for investigation because following the CHiME challenge rules the system's recognition outputs have been submitted alongside the correctness results. In the same spirit we also plan to share the features themselves and to make our CHiME system code available on request.

5. References

[1] M. Wöelfel and J. McDonough, *Distant speech recognition*. Wiley, 2009.

[2] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research

developments and directions in speech recognition and understanding, Part 1," *IEEE Signal Processing Magazine*, vol. 26, pp. 75–80, 2009.

[3] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[4] 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. IEEE ICASSP*, Vancouver, Canada, 2012.

[5] N. Ma, J. Barker, H. Christensen, and P. Green, "A hearing-inspired approach for distant-microphone speech recognition in the presence of multiple sources," *Computer Speech and Language*, in press.

[6] N. Ma and Barker, "Coupling identification and reconstruction of missing features for noise-robust automatic speech recognition," in *Proc. Interspeech*, Portland, Oregon, 2012.

[7] P. Renevey and A. Drygajlo, "Detection of reliable features for speech recognition in noisy conditions using a statistical criterion," in *Proc. CRAC*, Aalborg, Denmark, 2001.

[8] C. Cerisara, S. Demange, and J. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Computer Speech and Language*, vol. 21, pp. 443–457, 2007.

[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 504–512, 2001.

[10] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 466–475, 2003.

[11] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and uncertain acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[12] B. Raj, M. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.

[13] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[14] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech and Language*, vol. 27, no. 3, pp. 851–873, 2013.

[15] M. Wöllmer, J. Geiger, B. Schuller, and G. Rigoll, "Noise robust ASR in reverberated multisource environments applying convolutive NMF and long short-term memory," *Computer Speech and Language*, vol. 27, no. 3, pp. 780–797, 2013.