Experiments on Far-field Multichannel Speech Processing in Smart Homes

I. Rodomagoulakis[1,3], P. Giannoulis[1,3], Z.–I. Skordilis[1,3], P. Maragos[1,3], and G. Potamianos[2,3]

1. School of ECE, National Technical University of Athens, 15773 Athens, Greece
2. Department of CCE, University of Thessaly, 38221 Volos, Greece
3. Athena Research and Innovation Center, 15125 Maroussi, Greece

irodoma@cs.ntua.gr, maragos@cs.ntua.gr, gpotam@ieee.org

***Abstract*—In this paper, we examine three problems that rise in the modern, challenging area of far-field speech processing. The developed methods for each problem, namely (a) multichannel speech enhancement, (b) voice activity detection, and (c) speech recognition, are potentially applicable to a distant speech recognition system for voice-enabled smart home environments. The obtained results on real and simulated data, regarding the smart home speech applications, are quite promising due to the accomplished improvements made in the employed signal processing methods.**

***Index Terms*—smart homes, microphone arrays, array processing, speech enhancement, voice activity detection, speech recognition**

I. Introduction

The recently emerged intelligent applications for smart domestic environments [1] are designed to offer new opportunities to security, awareness, comfort, and full environmental control in daily indoor life. A major effort [2] in this research field refers to impaired or elderly people with physical disabilities. Among all the employed interaction and sensing technologies, speech processing technology has a great potential to become one of the major interaction modalities, enabling natural and fast human-computer interaction without the necessity of body- or head-mounted microphones. Although voice interfaces enable potentially richer interactions, one of the major issues that prevents the development of speech technology in real home settings is the poor performance of Automatic Speech Recognition (ASR) in noisy environments, as well as the unsolved challenges that emerge in complex acoustic scenes with multiple, possibly overlapping events. The corruption of speech signals is due to interfering sounds and reverberation. These sources of signal degradation can be effectively suppressed by combining multiple microphones for signal processing. The research in the field of microphone array processing deals with problems such as source localization, separation, and enhancement for Distant Speech Recognition (DSR) [3] in acoustic environments with multiple events. Although such array processing techniques have received great attention in the signal processing community over the last years, the research in ASR ignores a great amount of their benefits [4].

This research was supported by the European Union project DIRHA with grant FP7-ICT-2011-7-288121.

A DSR system with a microphone array usually consists of speaker localization, beamforming (BF), post-filtering, and ASR. First, the speaker's position is estimated and then, given the estimation, the beamformer emphasizes the signal coming from a direction of interest. The beamformed signal can be further enhanced by applying post-filtering and, finally, the enhanced signal is fed to the ASR system. A real domestic environment usually involves non-speech acoustic events which must be distinguished from the voice segments by applying Voice Activity Detection (VAD).

The contributions of this paper lie on three problems of the DSR system, namely (a) speech enhancement, (b) voice activity detection, and (c) speech recognition. In Section II, a state-of-the-art multichannel speech enhancement system with beamforming and post-filtering is presented. The system includes a source localization module for data-driven estimation of the source location, which is needed for effective beamforming. The source localization algorithm uses a closed-form source location estimator and is, therefore, fast and introduces small overhead to the enhancement system. Section III presents supervised and unsupervised methods for speech/non-speech classification in multichannel simulations in a realistic home environment. Noisy conditions and multiple acoustic events that may overlap frequently comprise a challenging environment. The proposed classifier performs accurately and close to real time. Finally, Section IV describes the implementation of an ASR system with efficient acoustic and language modelling for a large vocabulary Greek task, targeting spontaneous speech recognition in a reverberant room with noisy conditions.

Overall, the above contributions led to promising results and improvements on a variety of challenging problems in the examined field of DSR for smart home applications. Such applications are explored within the recently launched EU project under the name "Distant-speech Interaction for Robust Home Applications" (DIRHA) [5].

II. Multichannel Speech Enhancement

The use of microphone arrays presents the advantage that spatial information is captured in the recorded signals. Therefore, in addition to spectral characteristics, spatial characteristics of speech and noise signals can also be exploited for

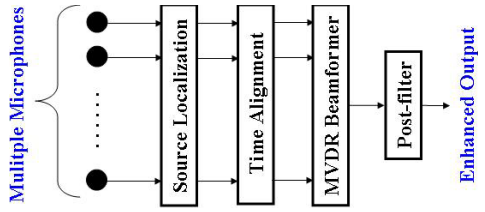

Fig. 1. Multichannel Speech Enhancement System with Post-Filter

speech enhancement. To exploit spatial information, beamforming algorithms have been proposed [6]. In addition to beamforming, post-filtering is often applied to further enhance the desired signal. Commonly used for speech enhancement are the minimum mean-square error (MMSE), the short time spectral amplitude (STSA) [7], and the log-STSA [8] estimators, each of which is equivalent to a Minimum Variance Distortionless Response (MVDR) beamformer followed by a single-channel post-filter [9], [10], [11].

In this paper, a state-of-the-art speech enhancement system is presented, which implements the aforementioned estimators and consists of a source localization and time alignment module, an MVDR beamformer, and a post-filter (Fig. 1).

A. Multichannel Speech Enhancement System

The system inputs are the signals recorded by a set of M microphones in a noisy environment. It is assumed that the signal $s_m(t)$ recorded by microphone m can be modeled as:

$$s_m(t) = s(t) * r_m(t) + v_m(t),\ m = 1, 2, \ldots, M \quad (1)$$

where $s(t)$ is the source signal, $r_m(t)$ is the impulse response of the acoustic path from the source to microphone m, and $v_m(t)$ is an additive noise component. For enhancement purposes, this signal model will be simplified by assuming that reverberations are negligible, namely $r_m(t) = \alpha_m \delta(t - \tau_m)$, where α_m is the attenuation factor and τ_m is the time needed for the source signal to travel to the m-th microphone, so that $s_m(t) = \alpha_m s(t - \tau_m) + v_m(t)$.

The time alignment module temporally aligns the input signals. To do so, the Time Differences of Arrival (TDOAs) of the speech signal to the microphones must be estimated. To compute the TDOAs, speech source localization is first performed. The speech source is localized using a TDOA-based source localization algorithm. First, TDOAs are independently estimated for various microphone pairs of the array, using the Crosspower-Spectrum Phase Coherence Measure (CSP-CM) [12]. For the microphone pair m_1, m_2, the CSP-CM:

$$C_{s_{m_1} s_{m_2}}(\tau, t) = \int_{-\infty}^{\infty} \frac{S_{m_1}(f,t) S^*_{m_2}(f,t)}{|S_{m_1}(f,t)||S_{m_2}(f,t)|} e^{j2\pi f \tau} df, \quad (2)$$

where $S_{m_1}(f,t)$ and $S_{m_2}(f,t)$ are the Short Time Fourier Transforms (STFTs) of $s_{m_1}(t)$ and $s_{m_2}(t)$, respectively, is expected to have a prominent peak at $\tau = \tau_{m_1} - \tau_{m_2}$.

Once the TDOAs have been estimated, the Directions of Arrival (DOAs) of the source signal to each microphone pair are computed. Adopting a far-field propagation model and assuming that the microphones and the source are coplanar, the DOA for each microphone pair m_1, m_2 is a straight line that passes trough the midpoint of the microphone pair baseline with an incident angle of $\theta_{m_1 m_2} = \cos^{-1}(c(\tau_{m_1} - \tau_{m_2})/d_{m_1 m_2})$, where c is the speed of sound and $d_{m_1 m_2}$ is the distance between microphones m_1, m_2 [13]. Ideally, the DOAs intersect at a single point, i.e. the speech source location. However, due to errors in the TDOA estimates, this is not the case in practice and the source location has to be estimated using an error minimization criterion. The approach taken is to find the point on the plane that minimizes the sum of squared distances from the DOA lines. Expressing the DOA line for each microphone pair i in parametric form as $\mathbf{y}_i = \mathbf{x}_i + \lambda \mathbf{r}_i,\ \lambda \in \mathbb{R}, i = 1, 2, \ldots, N$, where N is the number of available microphone pairs, $\mathbf{x}_i$ is the midpoint of the microphone pair baseline, $\mathbf{r}_i$ is the unit vector in the DOA direction, and λ is a parameter that spans $\mathbb{R}$ so that $\mathbf{y}_i$ spans the points on the line, the source location estimator $\mathbf{a}_0$ that satisfies this minimization criterion is:

$$\mathbf{a}_0 = \mathbf{A}^{-1} \sum_{i=1}^{N} (\mathbf{A}_i \mathbf{x}_i),\ \mathbf{A} = \sum_{i=1}^{N} \mathbf{A}_i,\ \mathbf{A}_i = \mathbf{I} - \mathbf{r}_i \mathbf{r}_i^T, \quad (3)$$

where $\mathbf{I}$ denotes the identity matrix. The estimated source location combined with knowledge of the microphone positions enables calculation of the τ_m quantities and consequently alignment of the signals $s_m(t)$.

The MVDR beamformer operates on the aligned signals and produces a single output, which is then processed by the post-filter to obtain the enhanced signal. The MVDR beamformer weights and the post-filter transfer function for each of the MMSE, STSA and log-STSA post-filters are estimated using the estimation procedure proposed in [14].

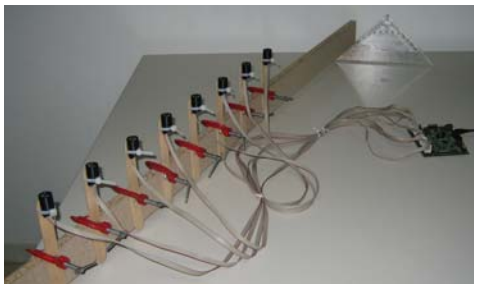

Fig. 2. MEMS Microphone Array

B. Experimental Results

Experiments were conducted on the CMU database and with signals recorded with a MEMS (Micro-Electro-Mechanical System) microphone array. As an objective speech quality measure, the segmental Signal to Noise Ratio (SSNR) was used [15].

TABLE I
SPEECH ENHANCEMENT RESULTS FOR THE CMU DATABASE

Estimator	SSNR Enhancement (dB)
MMSE	14.2320
STSA	13.9078
log-STSA	14.0848

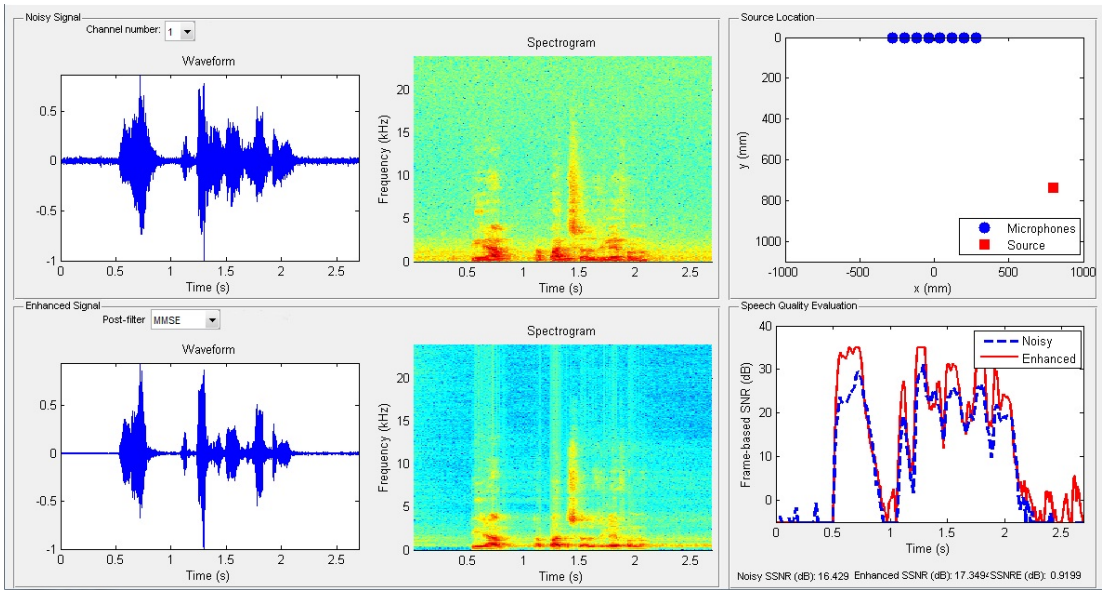

Fig. 3. Speech Enhancement Results using the MEMS Microphone Array: utterance "DIRHA answer the phone" (in Greek)

1) CMU database: The CMU database [16] contains 16kHz recordings of 130 utterances with an 8–element linear microphone array with 7cm microphone spacing. Table I shows the average SSNR enhancement (SSNRE) achieved, which is calculated as the dB difference between the SSNR of the noisy signal at the central microphone of the array and of the enhanced output. A substantial SSNRE of about 14dB was achieved.

2) MEMS Microphone Array: A few preliminary experiments were also conducted with a microphone array consisting of MEMS microphones, a newly developed technology of very compact sensors. Technical details regarding the MEMS sensing elements can be found in [17]. The MEMS array used consists of 8 sensors streaming audio at 48kHz, which can be configured in any desired geometry. For the preliminary experiments, the microphones were configured linearly with 8cm spacing. The configuration is shown in Fig. 2.

A few DIRHA related commands in Greek were uttered by a human talker at various positions relative to the array. Significant enhancement of the speech signal was observed. Indicative results are shown in Fig. 3, which depicts the experiment in which the sentence "DIRHA answer the phone" (in Greek) was uttered by a human speaker standing 1m from the array center at an angle of 45 degrees relative to the array carrier-line. The significant enhancement achieved is evident.

III. Voice Activity Detection

Voice activity detection (VAD) refers to the problem of distinguishing speech from non-speech segments in an audio stream. The non-speech regions could include silence, noise, or a variety of other acoustic signals. Also in case of overlap between speech and other events, there is need of detection and separation. Speech/non-speech segmentation of the acoustic input constitutes a crucial step that provides important information to other system components, such as speaker localization, automatic speech recognition, and speaker recognition. Especially in the case of human-computer interaction, it needs to perform in a highly precise and real-time manner. In our work, we discuss some VAD algorithms, and compare their performance in the DIRHA database environment. Our best effort performs quite fast and accurately.

A. Teager Energy Based Segmentation

This algorithm, reported in [18], was developed in order to achieve accurate speech/non-speech segmentation in highly noisy environments. In contrast to other energy-based algorithms that use traditional energy and zero crossing rate (e.g. [19]), this method employs Teager energy as a feature, combined with an adaptively computed threshold for making a speech/non-speech decision. The Teager energy operator is defined as $\Psi[x(t)] = \dot{x}^2(t) - \ddot{x}(t)x(t)$. The new energy representation is derived through Gabor filtering the signal in various frequency bands, estimating their average Teager energies, and keeping the maximum of them (the most active one). For each frame, the feature computed is $\max_k \{\overline{\Psi(s * h_k)}\}$, where s is the speech signal, h_k is the impulse response of the k^{th} Gabor filter, and $\{\bar{\cdot}\}$ denotes short-time averaging. The algorithm is unsupervised, in the sense that it does not require a training procedure.

B. GMM Classifier Using Mel Band Energy Features

In this approach, a Gaussian mixture model (GMM) based speech/non-speech classifier is trained, and subsequently applied over a short-time sliding window, making a binary decision on whether it corresponds to speech or non-speech. 32 Mel band log-energy (MBLE) features are extracted over short-time windows of 25ms in duration, without proceeding to the DCT based compression/de-correlation stage that yields the traditional Mel-frequency cepstral coefficients (MFCCs) [20]. For each frame, we compute $\text{MBLE}_k = \log \overline{E(s * h_k)}$, where E is the classic energy operator, and h_k denotes here the impulse response of the k^{th} triangular filter of Mel-filterbank.

TABLE II
PERFORMANCE OF THE FOUR SPEECH/NON-SPEECH SEGMENTATION SYSTEMS PRESENTED, DEPICTED IN TERMS OF SPEECH/NON-SPEECH FRAME CLASSIFICATION ERROR.

VAD Approach	test1	test2
Teager Energy	25.25%	27.07%
Teager Energy + GMM/MFBE	19.03%	20.80%
GMM classifier/MFBE features	7.83%	8.80%
GMM/MFBE, 2-mic fusion	6.69%	7.14%

For speech/non-speech modelling, Gaussian mixtures with full covariances are employed, in particular six such mixtures per class. The GMM is trained on a subset of the DIRHA database, using the expectation-maximization algorithm [20]. During testing, the GMM is applied over feature sequences that correspond to short-time sliding windows of 0.5s in duration and a 0.25s overlap. The final speech/non-speech decision is thus obtained every 0.25s, based on the accumulated log-likelihood difference of the two models over the 0.5s window, also biased by an appropriately chosen global threshold that plays the role of a decision confidence. It is worth saying that this VAD implementation permits close to real-time performance.

C. Combined Teager Energy and GMM Based Segmentation

In this approach, the Teager energy based speech/non-speech segmentation system, mentioned earlier, is considered as the first step of a two-stage cascade. The second step of the algorithm applies the GMM based speech/non-speech classifier to provide a final decision as to whether each segment detected as speech by the Teager energy based subsystem should be classified as speech or non-speech. This system thus has the ability to reject erroneously detected speech segments (e.g., segments where other acoustic events are present), however it lacks the ability to further refine such segments into possible speech and/or non-speech subsegments.

D. Multiple Microphone Combination

The above approaches have been considered using data from a single microphone, but can be easily extended to employ data from multiple microphones in the DIRHA scenario. A simple such algorithm has been developed in conjunction with the speech/nonspeech segmentation system "B" above, using data from two microphones under a decision fusion framework. In more detail, under this approach, a particular frame is classified as speech if both microphones classify it as speech with a confidence above a threshold T_a, or if at least one of the microphones does so with a confidence above a rather higher threshold $T_b > T_a$.

In Table II we summarize the results obtained from the different VAD algorithms using two DIRHA testing sets, while in Fig. 4 we show an example of their performance on a single, 1min recording.

IV. LARGE VOCABULARY SPEECH RECOGNITION

This section addresses the Large Vocabulary Continuous Speech Recognition (LVCSR) problem in voice-enabled automated home environments, based on single-channel distant-speech input. Recognition in distant-talking environment is a challenging task due to environmental noise, room acoustics, interfering sources etc.

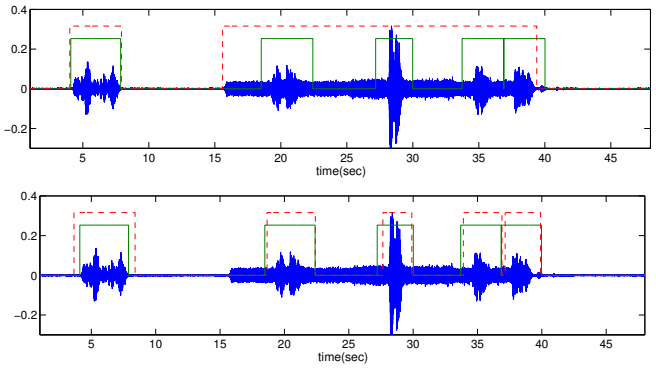

Fig. 4. Examples of application of speech/non-speech segmentation algorithms "A" (upper diagram) and "B" (lower diagram) on a DIRHA recording. Acoustic waveforms (blue), the ground truth (lower, solid green rectangles), and the derived speech segments (higher, dashed red rectangles) are shown. In this recording there is strong overlap between speech, alarm, and water noise.

The implemented Greek LVCSR system aims to recognize speech signals $s(t)$ for multiple combinations of speaker and microphone positions in a room of a smart home. The formed source-microphone channels affect the signal in different ways depending the distance and the room reverberation effects.

The main challenge in building LVCSR systems for such scenarios is to achieve robustness against reverberation and noise effects. One direction to achieve robustness in recognition is to combine multichannel information in signal, feature or decision level. Speech enhancement and voice activity detection, as described before, belong to this class of methods. Another direction is modelling, in which acoustic and language models can be effectively designed to compensate environmental changes and large vocabulary issues. The next paragraphs describe the implementation of language and acoustic modelling for the Greek language. Speaker independent recognition experiments for simulated distant speech are also reported.

A. Language Modelling

Bigram and tri-gram models are build for the Greek journalism domain, in a collection of text mixed up from various sources, containing 12.2 million words, from which the 90% is used for training and the other 10% (`total` set) is used for testing. The vocabulary size of the total text amounts to 242k words, but only a subset of 37k words are considered for modelling, those contained in the transcriptions of the "Logotypographia" [21] corpus in which LVCSR experiments are conducted. A portion of "Logotypographia" transcriptions is included in the language modelling text to decrease the model perplexity when measured in the `logo` set. The `logo` set consists of the corresponding transcriptions of the 1k `eval` set, used for the LVCSR evaluation in the "Logotypographia" corpus.

Language modelling is implemented with the Carnegie Mellon University language modelling toolkit [22] which supports Good-Turing discounting for better modelling of

TABLE III
PERFORMANCE OF THE DEVELOPED GREEK BIGRAM AND TRI-GRAM LANGUAGE MODELS IN TERMS OF PERPLEXITY (PP) MEASURED ON THE TWO EVALUATION SETS `total` AND `logo`.

model type	size	PP-`total`	PP-`logo`
bigrams	335k	297	143
tri-grams	424k	192	30.1

low-frequency word sequences. The performance in terms of perplexity is presented in Table III. Perplexity is defined as $PP = \hat{P}(w_1, w_2, \ldots, w_m)^{-\frac{1}{m}}$, where $\hat{P}(w_1, w_2, \ldots, w_m)$ is the probability estimate assigned to the word sequence $(w_1, w_2, \ldots, w_m)$ by the language model. Out-of-vocabulary (OOV) rates were also measured for both `total` and `logo` sets. The obtained rates were 8% and 0%, respectively. Zero OOV for the `logo` set was due to the closed vocabulary type of the trained language model. Notice that for the LVCSR experiments presented next only tri-grams are incorporated in decoding due to their superior perplexity.

B. Simulated data for distant-speech

Simulations of distant-speech are considered for experimentation due to lack of real recordings in home environments. The simulations have been acquired by applying eq. (1) in a original set of 27 hours close-talk recordings of the multi-speaker Greek journalism database "Logotypographia" [21]. Two simulation sets were produced for experimentation, `reverb1` and `reverbR`. In `reverb1`, conditions are constant, i.e source in position LA (see the map of Fig. 5), microphone LR3 and additive ambient noise with gain 3. In `reverbR`, conditions are randomly changed by applying 10 source-microphone impulse responses (LA-L3R, LA-LC1, LA-L4R, LA-L2R, LC-LC1, LC-L4R, LC-L1R, LD-LC1, LD-L1R, LD-L4R) combined with 3 noise levels (3, 6, 9). With these two sets, we can test the ability of the ASR system to recognize distant speech from multiple speakers in multiple positions inside the simulation room as it is depicted in Fig. 5.

C. Acoustic Modelling

Three sets of acoustic models are developed for distant speech recognition experiments, one for the original `clean` data and the other two for the simulated data `reverb1` and `reverbR`. Using these models, we are able to test the robustness of the LVCSR system in home environments. Also, we are able to test how the system behaves in mismatched training and testing conditions.

The traditional 39-dimensional MFCC and Perceptual Linear Prediction (PLP) front-end is employed [24] for the extraction of 13 MFCCs or PLPs including the coefficient C_0 and augmented by their first and second order time derivatives. Feature extraction is applied on 32ms length windowed segments of the pre-emphasized speech waveform producing features at a 100Hz rate. Utterance-level cepstral mean normalization is also applied to reduce data variability.

The acoustic modelling is based on a set of 28 linguistically approved Greek phonemes. The open source HTK framework

TABLE IV
WER, %, OF THE BASELINE GREEK LVCSR SYSTEM ON THE EVALUATION SET IN BOTH MATCHED AND MISMATCHED TRAINING/TESTING CONDITIONS.

training conditions	testing conditions					
	`clean`		`reverb1`		`reverbR`	
	MFCC	PLP	MFCC	PLP	MFCC	PLP
`clean`	3.34	3.30	83.21	82.86	85.13	84.24
`reverb1`	96.24	93.57	9.53	11.05	12.87	15.13
`reverbR`	94.25	91.24	17.08	16.94	14.56	16.57

[24] is used for the development of 3-state, 16-Gaussian tied-state triphones. The training procedure is applied independently for each set of acoustic models using approximately 20 hours of speech by 58 speakers. First, monophone models are trained by employing "flat-start" initialization due to the absence of time labels in the transcriptions and then, training of triphones, state-tying, and Gaussian mixture splitting are performed resulting to tied-state triphones for the `clean`, `reverb1`, and `reverbR` conditions respectively. The number of models per set ranged from 4000 to 5700. This deviation is due to the decision-tree based state clustering which depends on features. Finally, models for silence and noise have been trained. The noise model aims to capture the transcribed non-speech sounds such as "breath", "clear throat", "paff noise", "side speech", "paper rustle", and "phone ring".

D. Experimental Setup and Results

To evaluate the developed Greek LVCSR system, recognition experiments are conducted in matched and mismatched conditions for the `clean`, `reverb1`, and `reverbR` baseline acoustic models, in a speaker independent framework. In particular, a test (`eval`) set is selected for system evaluation

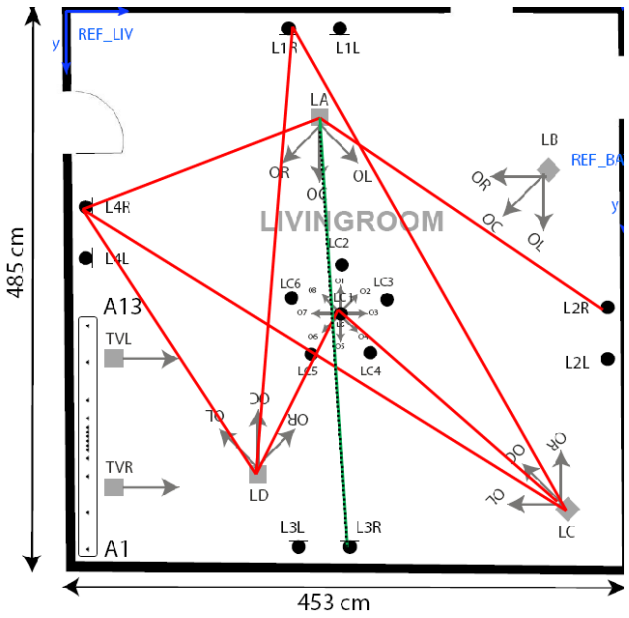

Fig. 5. Simulation map of the DIRHA apartment living room [23]. Green (dashed) and red (solid) lines correspond to the source-microphone location pairs which were simulated in the `reverb1` and `reverbR` sets, respectively.

consisting of 1k utterances which correspond to approximately 2.3 hours of speech by 15 speakers (different from the training set ones). The decoding parameters such as the word insertion penalty, the weights for the acoustic and language models, and the pruning threshold are optimized on a held-out (`dev`) set of 500 utterances. The decoding vocabulary contains 37k words.

Recognition results are reported in terms of WER, %, in Table IV. Overall, the performance under matched conditions is considered quite satisfactory, especially given the large vocabulary size, exhibiting similar WERs for MFCCs and PLPs. The observed low WER of 3.30% for clean speech can be justified by the exhibited low language model perplexity. As expected, there is a performance degradation in the distant speech data that is more pronounced in the more challenging `reverbR` scenario. Moreover, when acoustic models are trained and tested in mismatched conditions, the WER increases significantly compared to the matched condition results. Such degradation is less prominent between the two noisy conditions, compared to the degradation between the noisy and clean conditions. Comparing the two employed feature sets, MFCCs performed slightly better in all noisy matched conditions, although PLPs proved to be more robust in the most mismatched conditions. It is worth noting that if the tri-gram language model is not trained on "Logotypographia" text, its perplexity increases to 598 for the 1k `eval` set and the WER more than doubles by using MFCC features, reaching values of 11.49%, 24.73%, and 32.19% on the `clean`, `reverb1`, and `reverbR` matched training/testing conditions, respectively.

V. Conclusions

This work focused on the problems of multichannel speech enhancement, voice activity detection, and large vocabulary continuous speech recognition (LVCSR). The presented multichannel speech enhancement system achieved a high SSNR enhancement of approximately 14dB for the CMU database. Furthermore, the preliminary experiments with the newly developed technology of MEMS microphones showed very promising results. Regarding the VAD problem, the multichannel approach achieved very satisfying results in a real environment, yielding approximately 7% frame-level speech/non-speech classification error. Finally, the developed LVCSR system performed very satisfactorily for matched training/testing conditions achieving 3.3% WER for clean speech; the performance degraded gracefully in noisy conditions under matched training/testing. In future work, we will explore how the multichannel front-end processing methods, including speech enhancement and voice activity detection, can be used to improve the robustness of the ASR system in adverse conditions.

Acknowledgment

The authors wish to thank M. Omologo, P. Svaizer, and L. Cristoforetti of Fondazione Bruno Kessler for providing the simulated data for distant speech recognition and R. Sannino and L. Spelgatti of STMicroelectronics for kindly providing the MEMS microphone array.

References

[1] M. Chan, E. Campo, D. Estève, and J.-Y. Fourniols, "Smart homes – current features and future perspectives," *Maturitas*, vol. 64, no. 2, pp. 90–97, 2009.

[2] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects," *Personal and Ubiquitous Computing*, vol. 17, pp. 127–144, 2013.

[3] M. Wolfer and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[4] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[5] "The DIRHA (Distance-speech Interaction for Robust Home Applications) EU project." [Online]. Available: http://dirha.fbk.eu/

[6] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, pp. 4–24, 1988.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[8] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[9] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer Verlag, 2001, ch. 3, pp. 39–60.

[10] H. L. Van Trees, *Optimum Array Processing*. Wiley, 2002.

[11] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2002.

[12] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[13] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer Speech and Language*, vol. 9, no. 2, pp. 153–169, 1995.

[14] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Communication*, vol. 49, no. 7–8, pp. 657–666, 2007.

[15] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 1998, pp. 2819–2822.

[16] T. Sullivan, "CMU microphone array database," 1996. [Online]. Available: http://www.speech.cs.cmu.edu/databases/micarray

[17] *MEMS audio sensor omnidirectional digital microphone*, MP34DT01, STMicroelectronics, 2013. [Online]. Available: http://www.st.com/st-web-ui/static/active/en/resource/technical/document/datasheet/DM00039779.pdf

[18] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 14, no. 6, pp. 2024–2038, 2006.

[19] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[20] J. Fiscus, J. Ajot, M. Michel, and J. Garofolo, "The Rich Transcription 2006 Spring meeting recognition evaluation," *Machine Learning for Multimodal Interaction*, pp. 309–322, 2006.

[21] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V. Diakoloukas, "Large vocabulary continuous speech recognition in Greek: Corpus and an automatic dictation system," in *Proc. Eurospeech*, 2003, pp. 1565–1568.

[22] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit," in *Proc. Eurospeech*, 1997.

[23] M. Hagmüller *et al.*, "Experimental task definitions," *Deliverable D2.2, DIRHA Consortium*, Feb. 2013.

[24] S. J. Young *et al.*, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.