ROBUST FAR-FIELD SPOKEN COMMAND RECOGNITION FOR HOME AUTOMATION COMBINING ADAPTATION AND MULTICHANNEL PROCESSING

A. Katsamanis[1,3], *I. Rodomagoulakis*[1,3], *G. Potamianos*[2,3], *P. Maragos*[1,3], *A. Tsiami*[1,3]

[1]School of ECE, National Technical University of Athens, 15773 Athens, Greece
[2]ECE Dept., University of Thessaly, 38221 Volos, Greece
[3]Athena Research and Innovation Center, 15125 Maroussi, Greece
{nkatsam, irodoma, maragos}@cs.ntua.gr, gpotam@ieee.org, atsiami@central.ntua.gr

ABSTRACT

The paper presents our approach to speech-controlled home automation. We are focusing on the detection and recognition of spoken commands preceded by a key-phrase as recorded in a voice-enabled apartment by a set of multiple microphones installed in the rooms. For both problems we investigate robust modeling, environmental adaptation and multichannel processing to cope with a) insufficient training data and b) the far-field effects and noise in the apartment. The proposed integrated scheme is evaluated in a challenging and highly realistic corpus of simulated audio recordings and achieves F-measure close to 0.70 for key-phrase spotting and word accuracy close to 98% for the command recognition task.

Index Terms— distant speech recognition, multichannel processing, keyword spotting, adaptation

1. INTRODUCTION

The recently emerged intelligent applications for smart domestic environments [1] are designed to offer new opportunities for security, awareness, comfort, and full environmental control in daily indoor life. Although voice interfaces enable potentially richer interactions, one of the major issues that prevents the development of speech technologies 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. There has been increasing attention to signal processing and ASR methods for more robust recognition [2] in everyday listening conditions. In this direction, the community of microphone array processing [3] has reached significant milestones in several problems of speech processing such as source localization, source separation and speech enhancement but end applications still have to earn from these benefits [4].

The current paper presents our research on multichannel command recognition and keyword spotting from far-field speech [5] targeting their integration for smart home applications. Such applications are designed in the framework of an ongoing EU project under the name "Distant-speech Interaction for Robust Home Applications" (DIRHA) [6]. As its name suggests, DIRHA ambitiously aims at introducing a multichannel, distant-speech-controlled system that would allow human-home interaction in real, everyday conditions. The current research prototype is developed for the apartment shown in Fig. 1. In this context, the success of the integrated system critically depends on the effective solution to the following problems, namely robust speech modeling, channel selection and channel combination. For robust speech modeling, in the absence of speech data for home environments, the mismatch between the complex acoustic environment and generic speech models can be reduced by artificially distorting the training data as in [7, 8], by dereverberation [9] and/or adaptation to a development set [10, 7]. For channel selection, the signal-to-noise ratio (SNR) of the microphones can be used as in [11], other signal-based metrics [12], or decoder-based techniques as the one presented in [13] that is based on the confidence of the speech recognition results for each microphone. Channel combination, similarly can happen either at the signal-level, e.g., by beamforming as in [11, 10], or at the decision-level by techniques such as ROVER [14], or a similar SNR-weighted confusion-network based fusion [11] or by the driven decoding algorithm, presented in [10].

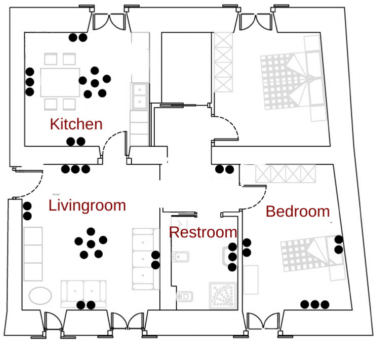

Fig. 1. The floorplan of the apartment for which the DIRHA prototype is being developed. The black dots represent the 40 microphones installed, namely 28 microphones on the walls and two pentagon-shaped arrays on the kitchen and livingroom ceilings.

In this work, we present an integrated approach for keyword spotting and speech recognition by investigating alternative schemes for the solution of the problems mentioned above. More specifically, a) we optionally apply beamforming with or without postfiltering, b) we use properly distorted training data for robust speech modeling, c) adapt the resulting models on a separate development set, d) select the most reliable channels based on signal-to-noise ratio (SNR), and e) then combine these channels via N-best list rescoring. The proposed system has been developed for Greek and is evaluated on a corpus collected for the purposes of the DIRHA project and comprises simulated recordings in the apartment. The recordings are

This research was supported by the EU project DIRHA with grant FP7-ICT-2011-7-288121. We wish to thank the members of the SHINE group at FBK for providing the simulated data and professor V. Digalakis at Technical University of Crete for kindly providing the LVCSR "Logotypografia" corpus. We are grateful to our research collaborators S. Lefkimmiatis and Z. Skordilis for the initial software for multichannel speech enhancement.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

highly realistic: they include speech of various types, background noise, various localized acoustic events, and suffer from significant reverberation. Overall, we achieve a 97.92% word accuracy in a 99-commands recognition task and reach an F-measure close to 0.70 for key-phrase spotting.

2. ROBUST AND MULTICHANNEL PROCESSING

2.1. Robust modeling

Robust modeling refers to the class of approaches that aim to reduce the mismatch between the training and testing conditions either by simulating the testing conditions for the generation of artificial training data or by using adaptation methods to fit the parameters of a model set in testing data.

Training corpus distortion. Reverberation and typical noise conditions in the apartment are expected to seriously degrade the performance of generic acoustic models for distant speech recognition. Training the models on apartment data can increase robustness but in most cases collection of a sufficient dataset for this purpose is not practical. Instead, the training corpus can be artificially distorted to reduce the mismatch with the real data [7]. According to this paradigm, clean training speech is convolved with acoustic impulse responses that have been measured in the apartment and can partially capture the reverberation properties of the rooms. To further simulate the real environment, additive noise can also be included.

In our work, we consider two variants of this distortion process that is applied on a generic speech corpus, namely the Greek, large vocabulary, continuous speech database “Logotypografia” [15][1] for the generation of two separate versions of the training corpus: a) $\mathbf{reverb}_1$ for which the room impulse response of a single source-and-microphone pair is used and white Gaussian noise is added at a certain gain, b) $\mathbf{reverb}_R$ for which one of 10 source-microphone impulse responses is randomly selected for each utterance and white Gaussian noise is randomly added at one of three gain levels. The first set corresponds to a simplified scenario that requires the measurement of a single impulse response in the apartment. On the other hand, the second set offers significantly higher diversity and is expected to be more representative of the real conditions. Additional details on the estimation of the impulse responses that are used and how the simulated data are generated can be found in [8].

Acoustic model adaptation. Environmental adaptation of the acoustic models to real data from the apartment can lead to additional performance improvements. In our current work, we only consider a supervised adaptation scenario, according to which few users (not necessarily the final users) of the system utter a predefined set of commands or other phrases in the apartment which can then be used for offline transformation of the acoustic models.

In the multichannel environment of the apartment, adaptation data, that comprise all the microphone recordings of the uttered commands, can be used in two ways, namely to determine a multichannel adaptation transformation that would be the same for all channels or a separate adaptation transformation per channel. The first scheme involves more data and can lead to a more robust transformation but the second scheme would allow a transformation to also capture more localized properties of the acoustic environment. In any case, conventional adaptation techniques are used to estimate the model transformations, i.e., global Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) adaptation [16].

2.2. Multichannel processing.

Channel combination via N-best list rescoring. Channel selection is a critical component of the proposed command spotting and recognition system. It is currently based on the SNR of the current speech segment to be recognized but more elaborate selection techniques can also be applied [13]. Recognition is only run on the microphone with the highest SNR. The corresponding results are then combined in three steps as described in the following:

1. Speech recognition on the microphone with the highest SNR returns an N-best list of possible hypotheses.
2. The N-best list is rescored for each of the selected microphones. Rescoring is achieved by forced-alignment of each of the hypotheses with the corresponding microphone recording using the adapted acoustic model for the specific microphone. The Viterbi algorithm is used for this alignment and the estimated log-likelihood of the best path is the hypothesis score.
3. The sum of all the microphone scores is estimated for each hypothesis and the list of hypotheses is resorted. The recognized command is the one with the highest combined score.

This algorithm was originally proposed for fusing heterogeneous speech recognition engines in [17] and to the best of our knowledge it has not been applied in the context of multichannel combination for distant speech recognition.

Beamforming is the typical way to fuse the microphones at the signal level. Speech signals are enhanced acoustically and their intelligibility is increased but in many cases there is no guarantee that the beamformed signal is directly exploitable from the recognizer due to the nonlinear distortions that enhancement methods can cause. In our experiments, we employ a minimum variance distortionless response (MVDR) beamformer, followed by a single channel Wiener post-filter with a minimum mean square error (MMSE) estimator [18] (MVDR-MMSE-est). The user (source) locations are estimated by an improved version of the method proposed in [19] based on previous work from our group in [20] [2]. Beamforming is only applied when the source is in the livingroom or kitchen, using the ceiling microphone arrays installed there, which satisfy the maximum inter-microphone distance requirements for the technique to be effective.

3. KEYWORD DETECTION AND SPEECH RECOGNITION

By using the techniques proposed in Sec. 2, we describe our keyword spotting and speech recognition systems, targeted for the apartment environment.

The problem of keyword-spotting (KWS) is well-studied [21] even in cases of large sets of keywords and noisy channels [22]. The classical keyword-filler approach for KWS implements whole-word Hidden Markov Models (HMMs) for keywords and one for general speech, i.e., garbage model. It is more flexible than the large vocabulary continuous speech recognition (LVCSR)-based KWS which acts more accurately in word- or phone-level lattices but requires large amounts of training data and a task-dependent language model [23]. Our implementation is in principle based on the approach introduced in [24]. It is designed to detect a predetermined set of small key-phrases. In the absence of training data, the keyword models are constructed from sub-word models pooled from the available LVCSR

[1] The corpus comprises 72 hours of speech (utterances of newspaper text) from 120 speakers using either a head-mounted or a desktop microphone.

[2] For comparison, we also experiment with the widely used delay-and-sum (DS) [3] and the MVDR-MMSE beamformers with known (ground truth) source locations: DS-gt and MVDR-MMSE-gt, respectively.

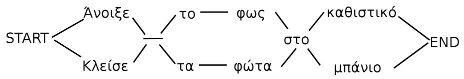

Fig. 2. An excerpt of the grammar for the recognition of 99 Greek commands for home automation. This subgrammar accepts the commands ”Switch on/off the light/lights in the livingroom/bathroom”.

system [25]. Sub-word models are properly concatenated to form whole-words which are then MLLR- and MAP- adapted to recordings of the key-phrases in the development set of the corpus coming from the apartment. Key-phrase spotting only runs for the best-SNR microphone. Note that the activation phrases that the keyword spotter aims to detect consist of one up to three words. The current implementation works with a simple grammar that forces the recognizer to recognize one key-phrase among possible garbage segments.

For speech recognition, the system is designed to recognize a set of 99 commands of various lengths. A finite-state-grammar has been built for the task, part of which is shown in Fig. 2. Speaker-independent, cross-word, triphone models are trained on the original and the two distorted versions of the training corpus. The corresponding model-sets are referred to by “clean”, “reverb$_1$” and “reverb$_R$” respectively. The models are then MLLR-adapted to the development set. Channel selection based on SNR and channel combination via N-best list rescoring determine the final command hypothesis.

4. EXPERIMENTS

We present experiments in the DIRHA simulated corpus. In our current work, we evaluate the keyword spotting and speech recognition components separately. So, potential keyword spotting errors, do not affect the speech recognition results.

The DIRHA corpus comprises simulated recordings of speech in a 4-room apartment that has been set up for the needs of the DIRHA project at the institute of Fondazione Bruno Kessler (FBK) in Trento, Italy, see Fig. 1. The corpus has been generated in two phases, namely the clean speech collection phase and the simulation phase. The goal has been to simulate realistic recordings in a home environment to the highest possible degree.

In the first phase, 20 speakers (10 male, 10 female) were recorded with a close-talking microphone at a rate of 48kHz in an unechoic studio uttering the following (among others) in Greek: a) 15 DIRHA key-phrases, e.g., “DIRHA activate”, b) 15 read DIRHA commands out of a set of approximately 250 commands, c) 15 spontaneous DIRHA commands, induced by showing a picture, d) 22 phonetically rich sentences, and e) 2 minutes of spontaneous, conversational speech, e.g., discussion about one’s favorite restaurant. The duration of the recorded speech material sums up to approximately 8 minutes per speaker [3].

In the second phase, this material was convolved with impulse responses from the DIRHA apartment to simulate real recordings of speech in the apartment. Real background noise and various random acoustic events representing typical sounds in a home environment (327 different events in total), e.g., appliance sounds, ringing, squeaking, were also properly added to generate the final simulated one-minute long recordings of acoustic events in the apartment. Each of these recordings can comprise multiple speech and acoustic events, occuring at random locations in the apartment and possibly overlapping. The corresponding recordings for all the 40 microphones are simulated. Overall, 150 simulations are generated, i.e., recordings for 150 mixtures of speech and noise events. Half of the data, involving half of the speakers, are held out as a development set and the rest form our test set.

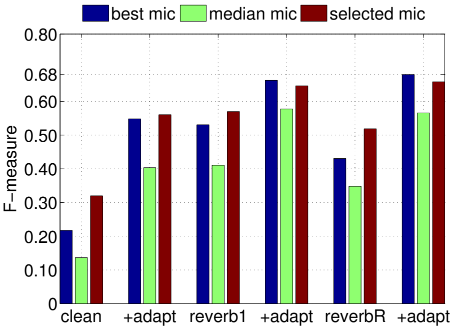

Fig. 3. F-measures for the key-phrase spotting system. The original models are just the word models created by concatenating the triphone models from the clean, reverb$_1$, and reverb$_R$ sets. The best and median microphones are found for all simulations and their performances are compared with a selected microphone having the best SNR per speech segment.

In this paper we focus on the recognition of the read commands that appear in these recordings and are preceded by one of the DIRHA key phrases. Our goal is to detect and recognize these commands in the one-minute long recordings. Key-phrase spotting is only activated for the speech segments identified by a voice activity detection module and speech recognition only runs for the speech segment following a keyword. Due to lack of space we are skipping the presentation of our multichannel VAD module[4]. To avoid any confusion, for the presented experiments we use the ground truth boundaries of speech segments in the recordings.

4.1. Key-phrase spotting experiments

The presented results for keyword spotting, in our case better described as key-phrase spotting, correspond to the 75 testing simulations of the DIRHA corpus. The detections must be accurate in terms of precision and recall in order to properly activate the ASR module when a command is going to be uttered.

The reported results are in terms of F-measure which is the harmonic mean of precision and recall. As it is depicted in Fig. 3, the adaptation step benefits all keyword and garbage model sets, especially the clean ones. Note that the synthetic word models correspond to the words contained in all system activation phrases. Also, the garbage model is trained on 10 hours of speech. Both model sets are trained in “Logotypografia” and adapted on the development data of each microphone in the “DIRHA” corpus. An important factor we noticed in the employed approach for keyword spotting is the duration of the garbage model. We experimented with the number of its HMM states to find that when we used 24 states, which is near the size of the average keyword model, with left-to-right transitions and 32 Gaussians per state, false alarms were significantly reduced.

In the multichannel environment of the apartment the keyword spotting module only runs on the audio coming from the microphone with the highest SNR. The results of channel selection are also depicted in Fig. 3. It is evident that the microphone with the best SNR per speech segment performs almost the same with the best microphone over all simulations yielding an F-measure of 0.67 in the best

[3] Examples of these recordings can be found at: http://cvsp.cs.ntua.gr/research/dirha

[4] It yields 4.2% frame detection error in the DIRHA test set.

	clean	reverb$_1$	reverb$_R$
DS-gt (unadapted)	0.22	0.66	0.68
DS-gt	0.74	0.81	0.79
SNR-best	0.54	0.62	0.62

Table 1. Key-phrase spotting F-measures using Delay-and-Sum beamforming with known source locations. Microphone selection results (with the adapted models) are also given for comparison.

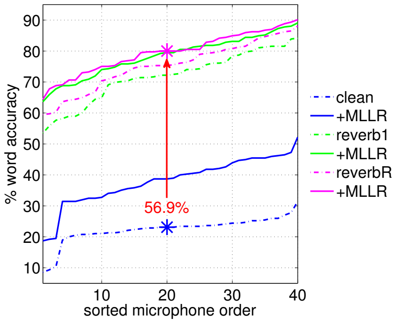

Fig. 4. Word accuracy for all the microphones. Results for all the models, before and after adaptation, are shown. Median performance is improved by almost 57% when the proposed robust modeling techniques are applied.

case of adapted reverb$_R$ models. Finally, the results for channel combination with beamforming (DS-gt) are presented in Table 1 for a subset of 51 simulations in which the activation phrase was uttered in the livingroom or kitchen. All the adapted models performed better in beamformed signals especially reverb$_1$ which achieved an F-measure of 0.81.

4.2. Speech recognition experiments

Model training and adaptation evaluation. For our current evaluations, speech recognition is run on all the speech segments containing only the spoken command, as ideally these would have been indicated by perfect voice activity detection and keyword spotting. Our engine is built using HTK [26]. More details on acoustic model training can be found in [25]. Separate microphone adaptation transformations are estimated using global MLLR. Preliminary experiments with regression-tree based MLLR did not yield any additional improvements possibly due to limited adaptation data. In Fig. 4 speech recognition performance is shown for all 40 microphones of the apartment, sorted by ASR performance, for all the acoustic models and their adapted versions. The word accuracies are shown in increasing order of magnitude. It can be seen that combined application of training using the properly distorted training corpus and adaptation leads to an improvement of median accuracy close to 57%. Adaptation alone can improve the performance of the "clean" models significantly but the performance of "reverb$_1$" and "reverb$_R$" models less so. Last but not least, the performance of the "reverb$_1$" and "reverb$_R$" appears to be more or less equalized after adaptation which indicates that large variability of the distortions of the training corpus may not be so critical as long as they capture the basic reverberation and noise properties of the real environment.

Channel selection, combination and beamforming evaluation. Results of SNR-based channel selection are shown in Table 2. The SNR-best microphone performs significantly better than the single best microphone. The latter is selected a posteriori based on its performance on all the test data, for all the cases of adapted speech models. By employing SNR as our channel selection criterion we pick the 3 SNR-best microphones for each simulation and combine them in the way described in Sec. 2.2. The second and third best-SNR microphones have SNRs that are typically not lower than 3dB from the top SNR. A 5-best list is generated for the highest-SNR microphone for each simulation and is then rescored by the other two microphones and resorted based on the combined score (mc-combined results). The resulting performance was better by 2.34% in absolute for the clean case compared with the SNR-best microphone performance as shown in Table 2.

	clean	reverb$_1$	reverb$_R$
best-mic	52.21	89.09	90.13
SNR-best	85.71	96.88	97.66
mc-combined	88.05	97.14	97.92

Table 2. Channel selection and channel combination word accuracy results for the adapted acoustic models. The best mic is selected a posteriori based on word accuracy for the entire test. The SNR-best microphone is the microphone with the highest SNR for each simulated command. The mc-combined results are based on the proposed multichannel combination approach.

	clean	reverb$_1$	reverb$_R$
DS-gt	78.41	96.52	98.49
MVDR-MMSE-gt	92.70	95.26	93.89
MVDR-MMSE-est	80.48	90.36	91.90
SNR-best	84.09	97.35	98.48
mc-combined	87.50	97.73	98.48

Table 3. Comparison with beamforming: word accuracy results in a subset for which beamforming can be effectively run. Simple delay and sum and MVDR-beamforming with MMSE postfiltering are tested. In all cases, MLLR-adapted models are used. Beamforming does not perform as well when the source location information is estimated and not known a priori (-est vs. -gt).

We also evaluated the proposed channel selection and combination scheme against the use of beamforming to fuse the information from multiple channels. Beamforming is applied for a subset of the test-set, i.e., for 51 simulations, for which the source is either in the kitchen or in the livingroom. Results for this subset using adapted models are shown in Table 3. Given the ground-truth source locations it appears that beamforming can perform very well. Also, it's worth noting that the MVDR-MMSE beamformer significantly outperforms the DS one for the clean models which have been trained in mismatched conditions. On the other hand, when source location information is estimated and not exact, the proposed channel selection and combination scheme performs better in all cases.

5. CONCLUSIONS

By combining robust modeling and multichannel processing, we presented our current approach for key-phrase spotting and command recognition for home automation. By employing SNR-based channel selection and the proposed N-best rescoring combination of multiple channels our system achieves word accuracy close to 88%, 97% and 98% for the adapted clean, reverb$_1$ and reverb$_R$ models respectively. For keyword spotting we reach an F-measure close to 0.70. In the future, we plan to explore alternative keyword modeling approaches and investigate the application of the proposed channel combination scheme for keyword spotting as well.

6. REFERENCES

[1] M. Chan, D. Estve, C. Escriba, and E. Campo, "A review of smart homes - Present state and future challenges," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 55–81, 2008.

[2] 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.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1, Springer, 2008.

[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 Process. Mag.*, vol. 29, no. 6, pp. 127–140, 2012.

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

[6] "DIRHA: Distant-speech Interaction for Robust Home Applications," [Online] Available: `http://dirha.fbk.eu/`.

[7] M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer, "Hidden Markov model training with contaminated speech material for distant-talking speech recognition," *Computer Speech and Language*, vol. 16, no. 2, pp. 205–223, 2002.

[8] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, "Impulse response estimation for robust speech recognition in a reverberant environment," in *Proc. European Signal Processing Conf.*, 2012, pp. 1668–1672.

[9] A. Sehr and W. Kellermann, "Towards robust distant-talking automatic speech recognition in reverberant environments," in *Speech and Audio Processing in Adverse Environments*, E. Hänsler and G. Schmidt, Eds., pp. 679–728. Springer Berlin Heidelberg, 2008.

[10] B. Lecouteux, M. Vacher, and F. Portet, "Distant speech recognition in a smart home: Comparison of several multisource ASRs in realistic conditions," in *Proc. Int. Conf. on Speech Communication and Technology*, 2011, pp. 2273–2276.

[11] M. Wölfel, C. Fügen, S. Ikbal, and J. W. McDonough, "Multi-source far-distance microphone selection and combination for automatic transcription of lectures," in *Proc. Int. Conf. on Speech Communication and Technology*, 2006, pp. 361–364.

[12] M. Wolf and C. Nadeu, "On the potential of channel selection for recognition of reverberated speech with multiple microphones," in *Proc. Int. Conf. on Speech Communication and Technology*, 2010, pp. 80 – 83.

[13] M. Wolf and C. Nadeu, "Channel selection using N-Best hypothesis for multi-microphone ASR," in *Proc. Int. Conf. on Speech Communication and Technology*, 2013.

[14] S.M. Chu, E. Marcheret, and G. Potamianos, "Automatic speech recognition and speech activity detection in the CHIL smart room," in *Machine Learning for Multimodal Interaction*, S. Renals and S. Bengio, Eds. Springer Berlin Heidelberg, 2006.

[15] 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. Int. Conf. on Speech Communication and Technology*, 2003.

[16] P. C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, 2001.

[17] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. M. Schwartz, and J. R. Rohlicek, "Integration of diverse recognition methodologies through reevaluation of N-Best sentence hypotheses.," in *Human Language Technology Conf.*, 1991, pp. 83–87.

[18] 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.

[19] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech and Audio Process.*, vol. 5, pp. 288–292, 1997.

[20] I. Rodomagoulakis, P. Giannoulis, Z.-I. Skordilis, P. Maragos, and G. Potamianos, "Experiments on far-field multichannel speech processing in smart homes," in *Proc. 18th Int. Conf. Digital Signal Processing*, 2013.

[21] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, no. 4, pp. 317–329, 2009.

[22] A. Mandal, J. Hout, Y.-C. Tam, V. Mitra, Y. Lei, J. Zheng, D. Vergyri, L. Ferrer, M. Graciarena, A. Kathol, and F. Horacio, "Strategies for high accuracy keyword detection in noisy channels," in *Proc. Int. Conf. on Speech Communication and Technology*, 2013.

[23] I.-F. Chen and C.-H Lee, "A resource-dependent approach to word modeling for keyword spotting," in *Proc. Int. Conf. on Speech Communication and Technology*, 2013.

[24] J. Wilpon, L. R. Rabiner, C.-H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 11, pp. 1870–1878, 1990.

[25] I. Rodomagoulakis, G. Potamianos, and P. Maragos, "Advances in large vocabulary continuous speech recognition in Greek: Modeling and nonlinear features," in *Proc. European Signal Processing Conf.*, 2013.

[26] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory, Cambridge, United Kingdom, 2002.