Room Localization for Distant Speech Recognition

Juan A. Morales-Cordovilla, Hannes Pessentheiner, Martin Hagmüller, Gernot Kubin

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
{moralescordovilla,hannes.pessentheiner,hagmueller,gernot.kubin}@tugraz.at

Abstract

The problem of room localization is to determine where, in a multi-room environment, a person is producing a speech utterance. In our work, we are exploiting the information gained from a network of microphones installed all over a house, where the lack of calibration of the microphone energies creates an additional challenge. This paper compares room localizers based on different features (such as energy and cross-correlation between microphones) and classifiers (such as neural networks and discriminative analysis). In order to evaluate the different room localizers in terms of word accuracy this paper also presents a complete distant speech recognition system which tries to take advantage of synergy between the different components without using any oracle information. Finally, the system is analyzed in terms of computational and time resources.

Index Terms: Distant speech recognition; microphone network; VAD; room localization; machine learning classification; enhancement; reverberant and noisy environment.

1. Introduction

Different challenges such as the recent REVERB [8] and projects such as CHIL, CHiME [2] and the current Distant-speech Interaction for Robust Home Applications (DIRHA, [3]) have been introduced to address the problems which appear in a home automation system controlled by a distant speech recognizer. These problems can be: the wake-up of the system (distinction between human-human conversations and human-system commands), the degradation of the speech signal due to the background noise or reverberation, and the localization of the command emitted by the speaker. This last problem is not only interesting because of the enhancement of the signal (by means of beamforming, etc.) but also because it can help the dialog system to distinguish, in an ambiguous situation, the device which the speaker wants to control. In this paper we focus on the determination in which room a person is producing a speech utterance. This can be solved in different ways such as using the WLAN signal emitted by a device [7] or with video cameras [14]. Some literature has tried to estimate the speaker position inside of a room using a microphone array [5] or a microphone network [1, 6]. The innovation of this paper is to localize the room using a microphone network. In addition this paper also presents a complete distant speech recognition system which exhibits synergy between the different components without using any oracle information.

The paper is structured as follows: Section 2 describes the database and the proposed system. Different room localization approaches are presented in Sec. 3. Sec. 4 analyses the results over the full system and gives some qualitative information on the computational cost of the system. Finally, Sec. 5 summarizes the most important ideas presented in this paper together with some future work.

This work has been supported by the European project DIRHA FP7-ICT-2011-7-288121 and the Austrian Marshall Plan Foundation.

2. Database and system

2.1. Database

The DIRHA-GRID database [9, 3] used to evaluate our algorithms has been proposed by the organizers of the Interspeech2014 special session: Multichannel Processing for Distant Speech Recognition. This database has 3 test sets: Dev1, Test1 and Test2 sampled at 16 kHz. Only the Dev1 (development set) has oracle spatial information and time boundaries of the utterances. Each audio file of around 1 minute of duration contains 40 microphone channels distributed in the 5 rooms of the ITEA apartment [3]. Each file has around 6 embedded utterances emitted randomly at different positions of the apartment in a noisy and highly reverberant (t_{60}=0.8 sec) ambient. The speaker utterances correspond to the GRID corpus [2] where they have a very restricted grammar (verb colour prep letter number coda), one utterance example is "bin blue at f two now". The utterances fulfil the non-overlapping condition, i.e., cross the time and rooms only one utterance is spoken. This condition simplifies the problem and is very realistic if we are interested in localizing the commands of one user. The organizers provide a training set with 17000 GRID clean utterances without any reverberation to train the automatic automatic speech recognition (ASR) system.

2.2. Description of the distant speech recognition system

The proposed distant speech recognition system is depicted in Fig.1. For each room we have a single-channel voice activity detector (VAD). These detectors use the following microphones [9]: KA6, LA6, R1C, C1R and B2C for 1 (kitchen), 2 (living-room), 3 (bathroom), 4 (corridor) and 5 (bedroom) rooms, respectively. The VAD decision is passed through a dilation morphological filter [4] of 0.5 sec of half length to obtain the segmented utterance. This point forward, we assume that the VAD provides utterances and not speech segments. After that, we localize the room of the utterance, estimate its position inside the room (using the pentagonal microphone arrays) and apply beamforming. Note that only if we detect the utterance in the living-room or in the kitchen we do beamforming with the corresponding pentagonal arrays, otherwise we use the reference single-channel signal of the room. Finally we enhance the utterance and recognize it. Other block configurations would be possible, such as to execute in this order: VAD, position estimation and beamforming for each room and later room localization. This last configuration would provide a very slow response until the utterance is completely segmented and localized in the room due to the expensive computational cost of our position estimation (see Sec. 4.2). In conclusion, our system configu-

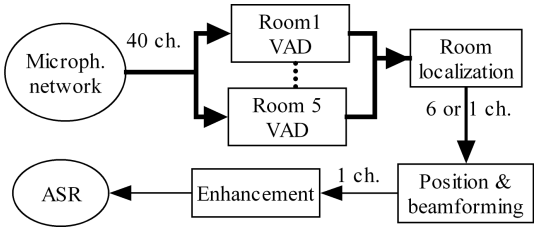

Figure 1: Block diagram of the proposed system for distant speech recognition which consists of a 40-elements microphone network, VAD for each room, room localization of the utterance, estimation of its position in the room, beamforming, enhancement of the single-channel signal and ASR.

ration is chosen basically because it segments and localizes the room relatively fast (Sec. 4.2).

2.3. ASR baseline

Our ASR is similar to the one proposed by the organizers: the same number of states per words and 7 Gaussians per state. The grammar for one utterance starts and ends with silence. The feature vector consists of a standard mel frequency cepstral coefficients (MFCCs) using these parameters: frame shift and length of 10 and 32 ms, 1024 frequency bins, 26 mel channels and 13 cepstral coefficients with cepstral mean normalization. Delta and delta-delta features are also appended, obtaining a final feature vector with 39 components. To compare our ASR with the one provided by the organizers we are going to recognize without any enhancement the signal provided by the LA5 microphone assuming oracle segmentation. First, we train an initial HMM using the training set and later we adapt it (doing 4 more EM iterations) using the 436 utterances of the Dev1 set of the LA5 microphone. The word accuracy (WAcc) on the Dev1 is 89.07 % (we concatenate all the transcriptions and compare with the true transcription). Note that in the following we will drop the % of the WAcc result. This result is different from the organizers (61.39) because their grammar recognize all the 1 min. signal but we recognize segment by segment. If we assume true room localization and we use the reference microphone of each room (Sec. 2.2) we improve our baseline to 92.89.

3. Room localization

In this section we compare different features on different classifiers for room localization assuming true VAD.

3.1. Maximum energy selection approach

When a speaker pronounces an utterance this can be detected not only by the VAD of the current room but also by the VADs of the surrounding rooms due to the omni-directional sound propagation through the open doors between the rooms. In Fig. 2 we can observe this phenomenon when we see vertically how most of the utterances appear repeatedly in all rows. If we consider the non-overlapping condition (Sec. 2.1), we can estimate the true room localization by means of a maximun selection along the rooms of the averaged noisy energy of the VAD segmented utterances. For one utterance this is computed as follows:

$$E_y^{utt} = \frac{\sum_{i=1}^{N} E_y(i)}{N} \quad (1)$$

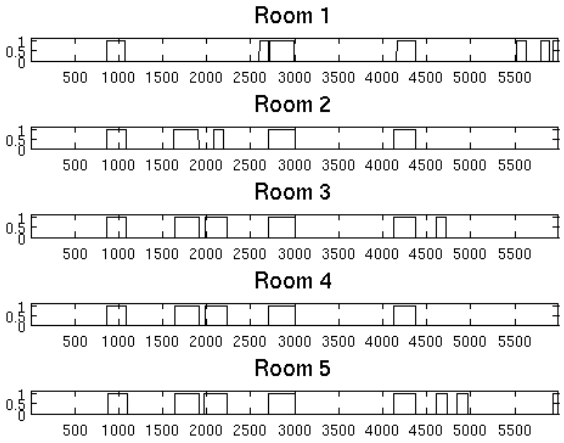

Figure 2: VAD estimations by the deep belief network (DBN) for the five rooms of the apartment for the signal sim2 of Dev1. The abscissa represents the frame number.

where N is the number of frames of the utterance. $E_y(i)$ is the noisy energy at frame i, computed as $E_y(i) = \sqrt{\sum_{i=1}^{L} y(t)^2/L}$ where $y(t)$ is the noisy signal of the room reference microphone at frame t and L is the frame length. Note that we will employ the notation of y, x and n for noisy, clean and noise signals respectively. To separate the room localizer from the rest of the system blocks we are going to test the localizer over the Dev1 assuming perfect VAD (i.e., the utterance is detected in all rooms, even if its SNR at the room is very low). For true room localization and with the above mentioned maximum energy selection the WAcc are: 92.89 (100) and 78.98 (56). In parenthesis we provide the F-Score comparing room by room the true with the estimated VAD after room localization. The F-Score is the harmonic mean between the frame precision and recall. We can see that the energy approach yields a poor performance mainly in the F-Score. This can confuse a lot the dialog system which can receive a correct command (such as "open the window!") but may associate it with the wrong room (Sec. 1).

3.2. Classification approach

The energy approach can produce localization errors because either the calibration of the microphone gains are not correct or the reverberation characteristics of the rooms enhance differently the speech. To mitigate this problem, we can train a classifier which learns the true utterance room localization depending on the energies observed in the five rooms. The input is a vector with the energies (Eq. 1) and the output is a vector with the five probabilities of observing the utterance in every room. For training purpose, we use Dev1 and we can directly take the true room localization as the probability vector, putting a 1 at the element corresponding to the room of the utterance and 0 otherwise. Due to the non-overlapping condition the sum of the probability elements is always 1. The Dev1 provides around 500 unique vectors of energies and probabilities. We use four standard classifiers from supervised machine learning: a multilayer perceptron neural network (NN, with a topology of 2 hidden layers of 10 and 5 neurons), linear and quadratic dis-

Table 1: Averaged word accuracies (WAcc, %) over the (Dev1,Test1,Test2) sets of the DIRHA-GRID database for different room localizations and configurations of the proposed system. * means not available result because not oracle information.

VAD	Pos., beamf. & enhanc.	Room localization						
		True	Energy	SNR	NN	LDA	QDA	SVM
True	No	93 (93,*,*)	79 (79,*,*)	83 (83,*,*)	80 (80,*,*)	90 (90,*,*)	89 (89,*,*)	87 (87,*,*)
True	Yes	96 (96,*,*)	82 (82,*,*)	85 (85,*,*)	85 (85,*,*)	92 (92,*,*)	92 (92,*,*)	90 (90,*,*)
DBN	No	93 (93,*,*)	52 (60,46,49)	45 (55,38,43)	55 (67,48,49)	60 (74,51,55)	57 (71,50,51)	58 (73,51,52)
DBN	Yes	96 (96,*,*)	52 (60,44,52)	46 (53,40,44)	55 (66,49,52)	**61 (72,54,57)**	59 (70,55,52)	60 (72,54,54)

criminant analysis (LDA and QDA) and support vector machine (SVM; which due to its binary decision there is one per room). All of them are taken from Matlab2012. The new room localizers use the same maximum selection as before but now the feature for comparison is not the energy but the classification probability averaged over the whole utterance. For NN, LDA, QDA and SVM the results are: 81.96 (65), 80.62 (66), 79.59 (60) and 81.77 (68). We see that using the probabilities as the comparison features improves the previous energy result (78.98 (56)), especially for the NN classifier.

3.3. SNR approach

Another possibility for the maximum selection is to use the SNR of the utterance as the comparison feature instead of the energy or the probabilities. Here, we estimate the averaged noise energy of the utterance (E_n^{utt}) by averaging the 20 frames of noisy energy $E_y(i)$ placed before and after the utterance. Then, the clean energy is $E_x(i) = E_y(i) - E_n^{utt}$ (we set to 0 if it is negative). The SNR of the averaged utterance energies is computed as follows:

$$SNR^{utt} = 20 \log_{10} \frac{E_x^{utt}}{E_n^{utt}} \quad (2)$$

where E_x^{utt} is the averaged clean energy of the utterance. When we use the SNR as comparison feature the WAcc (F-Score) is: 82.68 (48). We see that we improve regarding the classification approaches but the F-Score is very low. This was expected because for ASR a high SNR is important but not for the dialog system (Sec. 1). However, we could use one method with the highest WAcc as the selection method for the ASR input and another method with the highest F-Score as the selection method to inform the dialog manager about the room localization.

3.4. High-SNR classification approach

The problem with the previous classification approach is that if, in a room, the energy of the speech signal is dominated by another source, then the energy of this other source can confuse the classifier in the room estimation. So in case of low SNR it would be better not to include this energy information in the classification even if the VAD detected this utterance in the room. We can compute the instantaneous SNR at frame i as $SNR(i) = 20 \log_{10} E_x(i)/E_n^{utt}$. If the frame percentage of $SNR(i)$ higher than 0dB is lower than 15% (empirically chosen value) in a utterance then we set to 0 the corresponding element of the energy vector passed to the classifier. If we train and test the previous classifiers applying this high-SNR classification approach, for NN, LDA, QDA and SVM the results are: 82.00 (58), 88.11 (73), 87.81 (77), 86.93 (72) and 82.76 (48). We can see that the results improve and that now the best result is obtained by the LDA and not by the NN.

3.5. Coherence classification approach

Until now, we have employed only the energy as the input feature to the classifier. Other features which make a difference between the rooms can also be used and even combined with the energy to improve the localization. Here we propose to employ the cross-correlation between close microphone pairs because if the utterance is spoken in the same room as the pair, then the maximum of the cross-correlation will be higher than the maximum of another pair in other room. This is not only because of the attenuation of the sound when it travels through the rooms. This is because the microphone pairs of the true room receive the direct sound (with less reflections than the others) and a more coherent signal. If $y1$ and $y2$ are the noisy signals of a microphone pair p we define its coherence at frame i as:

$$c_p(i) = \max_k r_{y1,y2}(k,i) \quad (3)$$

where $r_{y1,y2}(k,i)$ is the cross-correlation and k is the lag. If in a room r we have a set of microphones the coherence of the room is defined as $c(i) = \max_p c_p(i)$. In order to guarantee a high coherence in all positions of a room, we use pairs on every possible room side: (K1R,K1L), (K2R,K2L), ... for room 1, (L1C,L1L), (L2R,L2L), ... for room 2, etc. The average of $c(i)$ provide the utterance coherence c^{utt}. We train and test the previous classifiers with the new input vector of 10 elements. This vector has the previous five high-SNR energies together with the c^{utt} of the five rooms. For NN, LDA, QDA and SVM the results are: 79.85 (50), 89.91 (81), 89.30 (82) and 86.89 (76). We can see that the addition of this new information improves the results for all the classifiers specially for the F-Score. Only for the NN, it is not the case. One reason can be that the topology of the NN needs to be better tuned because we have increased the dimension of the input vector. From here on, we will use this last presented feature, called high-SNR-energy+coherence, as the input feature to our classifiers.

4. Analysis of the full system

4.1. Experimental results

Now we are going to analyse the performance of the full system (Fig. 1). The first row of Table 1 summarizes the WAcc (averaged over Dev1, Test1 and Test2 sets) obtained by the most important seven room localizers presented along the paper: true VAD+room localization, maximum energy selection, maximum SNR selection, and the high-SNR-energy+coherence feature on the four classifiers. The results for Test1 and Test2, when we use true VAD or true room localization, are marked with * because the oracle information of them is not available. The second row in contrast to the first one applies position estimation, beamforming and enhancement. In [10] we study different possibilities for these blocks in a similar database and we draw the conclusion that the best performance is obtained when we use

the PoPi position [5], convex-optimized beamforming [11, 12] and vector Taylor series enhancement (VTS) [13]. Here, the VTS uses 128 Gaussians trained on a clean version of the Dev1 (the organizers provided us the impulse responses). The noise spectrogram is estimated with the first last 20 frames (FLFr) as in Sec. 3.3. Note that in order to avoid mismatch, we apply beamforming and enhancement to Dev1 and later we retrain our HMM (Sec. 2.3). The comparison of the first two rows clearly show that the addition of these enhancement blocks improve the results for all the room localizations. The third and the fourth rows repeat the same experiments but wit a real (not oracle) VAD based on deep belief networks (DBN) [15, 10]. We see that now we have less improvement with the addition of the enhancements. This is explained because when the utterance boundaries are not correct the VTS uses part of the speech to estimate the noise and it degradates the clean signal estimation. The result in bold (61), when we use the LDA room localizer, shows the best proposed result without using any oracle information.

4.2. Computational cost and time resources

We give some qualitative information about the computational cost as the organizers suggest. In Sec. 2.2 we claim that the first stage of our system provides a relatively fast response in the segmentation and room localization. In fact, when we use a Matlab2012 implementation on a x86-64bit AMD with 1400 MHz, this first stage has an averaged real time of x1.08 (i.e., it takes 65 seconds in processing the 60 seconds of the 40 channels). Most of the processing time of this stage is consumed by the feature extractor for the DBN-VAD (x0.90). The algorithms for DBN-VAD inference and room localization are very fast (x0.18). Add that they only require a maximum delay of 0.5 second because of the morphological filter (Sec. 2.2). The rest of the blocks have a real time of: x1.09, x0.05, x1.70, x0.15 for the position estimation, beamforming, enhancement and ASR, respectively.

5. Conclusion and future work

This paper has focused on the room localization problem and has shown that the maximum selection approach, of the VAD energies, produces a low F-Score. In order to improve this result, we have proposed different features on several machine learning classifiers. We have also presented a distant speech recognition system and shown that the different blocks, together with the room localizers, create synergy without using any oracle information. Finally, we have provided some complexity information of the system and have drawn the conclusion that the segmentation and room localization can almost be implemented in real time. The best WAcc performance (61%), achieved by the LDA classifier with high-SNR-energy+coherence as input feature, is still low to be implemented in a real system. This is mainly due to the errors of the VAD. As future work, we plan a further exploration of the relation between the VAD and the room localization, using multi-microphone VAD.

6. References

[1] A. Brutti. *Distributed Microphone Networks for sound source localization in smart rooms*. PhD thesis, ITC-irst Centro per la Ricerca Scientica e Tecnologica, 2007.

[2] H. Christensen, J. Barker, N. Ma, and P. Green. The CHiME corpus: A resource and a challenge for computational hearing in multisource environments. In *Interspeech*, pages 1918–1921, 2010.

[3] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, and P. Maragos. The DIRHA simulated corpus. In *LREC*, 2014.

[4] E. R. Dougherty. *An Introduction to Morphological Image Processing*. SPIE-International Society for Optical Engine, 1992.

[5] T. Habib and H. Romsdorfer. Auditory inspired methods for localization of multiple concurrent speakers. *Computer Speech & Language*, 27(3):634–659, 2012.

[6] F. Hummes, J. Qi, and T. Fingscheidt. Robust acoustic speaker localization with distributed microphones. In *EUSIPCO*, pages 240–244, 2011.

[7] Jozef Ivaneck, Stephan Mehlhase, and Margot Mieskes. An intelligent house control using speech recognition with integrated localization,. In *Ambient Assisted Living*, pages 51–62, 2011.

[8] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakataniand, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *IEEE WASPAA*, 2013.

[9] M. Matassoni, R. Astudillo, A. Natsamanis, and M. Ravanelli. The dirha-grid corpus: baseline and tools for multi-room distant speech recognition using distributed microphones. In *Interspeech*, 2014.

[10] J. A. Morales-Cordovilla, M. Hagmüller, H.Pessentheiner, and G. Kubin. Distant speech recognition in reverberant noisy conditions employing a microphone array. In *EUSIPCO*, 2014.

[11] J. A. Morales-Cordovilla, H. Pessentheiner, M. Hagmüller, P. Mowlaee, F. Pernkopf, and G. Kubin. A German distant speech recognizer based on 3D beamforming and harmonic missing data mask. In *AIA-DAGA*, pages 2049–2045, 2013.

[12] H. Pessentheiner, S. Petrik, and H. Romsdorfer. Beamforming using uniform circular arrays for distant speech recognition in reverberant environments and double-talk scenarios. In *Interspeech*, 2012.

[13] J. C. Segura, A. de la Torre, M. C. Benitez, and A. M. Peinado. Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks. In *Eurospeech*, 2001.

[14] M. Volkhardt, C. Weinrich, C. Schroeter, and H. Gross. A concept for detection and tracking of people in smart home environments with a mobile robot. In *2nd CompanionAble Workshop co-located with the 3rd European Conference on Ambient Intelligence*, 2009.

[15] X. L. Zhang and J. Wu. Deep belief networks based voice activity detection. *IEEE Trans. ASLP*, 21(4):697–710, 2013.