Towards Microphone Selection Based on Room Impulse Response Energy-Related Measures

Martin Wolf, Climent Nadeu

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
{mwolf, climent}@gps.tsc.upc.edu

Abstract

In a room where several distant microphones are capturing signals in parallel, the quality of the recorded speech signals strongly depends on the characteristics of the room impulse responses that describe the wave propagation between each source and each microphone. In this paper we present an initial attempt to investigate the possibility of selecting the microphone that offers the best quality of speech. As we want to apply it to an automatic speech recognition system, we aim to select the microphone according to some optimization criterion that has been inferred from the recognition rate in a prior learning process. Several energy-related measures that carry relevant information of the room impulse response are being considered. They should be estimated directly from the speech signal, possibly in real time, but avoiding the need to estimate the whole room impulse response. In this paper, we present the rationale behind the intended investigation, and offer preliminary experiments for a large vocabulary continuous speech recognition task which show how microphone selection using an ideal relative energy measure can largely improve the recognition rate.
Index Terms: microphone selection, reverberation, room impulse response, ASR

1. Introduction

Speech recognition in a room using distant microphones is a challenging task mainly due to background noise and reverberation. Acoustic signal is reflected from the walls and objects and arrives to the microphone attenuated and with different delays. In reverberant environments degraded copies of the original signal sum up in the receiver introducing interference even after the original sound disappears. This is usually modeled by convolution of the room impulse response (RIR) with the original speech signal

$$y(t) = x(t) * h(t) \quad (1)$$

where y(t), x(t) and h(t) are the recorded signal, the original speech signal and the room impulse response respectively.

Many techniques have been developed and successfully tested to cope with noises with short temporal effects (telephone channel effect, additive noises), but few have been so far reducing the long lasting effect of room reverberation. In conventional Automatic Speech Recognition (ASR) systems, short time spectra is used to derive the features for recognition and effect of reverberation may be observed as temporal smearing of the short term spectra. Size of the analyzed window is very short (tens of milliseconds) in comparison to the usual length of RIR, therefore techniques developed for the reduction of linear distortion in the short-term spectra usually fail.

In this work, we focus on room scenarios where there are multiple microphones capturing signal in parallel. Quality of recorded speech in each of them differs. We cope with the disturbing effects of the room by choosing the best signal available (in terms of speech recognition rate) at given time and under given conditions. The decision should be made before the feature extraction takes place without a need of any feedback from the ASR system. Method is suitable for scenarios where microphone array processing is not possible or desired since no assumptions about the position of microphones are made. In order to increase space diversity microphones should be distributed around the room, rather then concentrated in one place.

In [1], space-diversity speech recognition technique using distributed multi-microphone in room was investigated as a new of speech recognition. Authors propose microphone selection method based on maximum likelihood. There are several distant speech models trained for the room. In the first pass, speech is independently recognized for each microphone, and the model giving maximum likelihood is selected. In the second pass, the microphone with maximum likelihood is chosen. In this way the most reliable model and acoustic channel are selected. Disadvantage of this approach is that several acoustic models need to be trained under different conditions and evaluated in parallel. In our approach we assume only one speech model because microphone selection is made before recognition and is based solely on the measures extracted from the speech signal.

In previous study [2], relation among different parts of RIR and Word Error Rate (WER) were investigated. Results show there are certain components of RIR that harm the recognition more than others. If we identify and measure these components, it should be possible to say how is the signal in each of microphones affected by the conditions in the room. The least harmed signal will presumably lead to lower WER.

Estimation of RIR is a costly and difficult process, especially, if positions of speaker or conditions in the room are changing over time. Therefore for purposes of ASR it is more desirable to avoid such methods that require exact RIR measurement.

In the remaining part of the paper we describe the methodology, show preliminary results and outline future directions. The work is preliminary.

2. Experimental setup

There are two basic questions: what are the parameters of RIR that should be taken into account when making decision, and how to measure or estimate them? To answer this we define a set of experiments where close talk microphone recordings, without influence of the room reverberation, are artificially con-

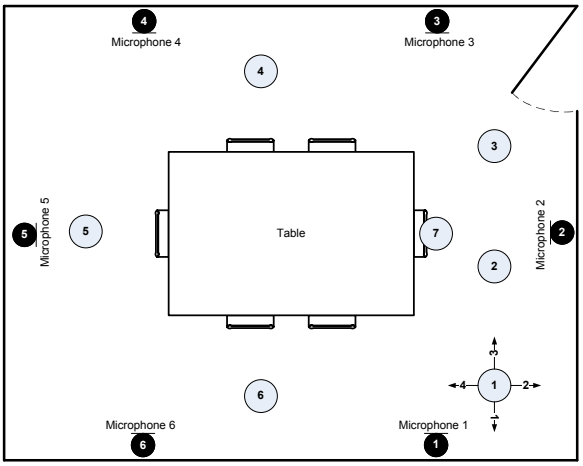

Figure 1: *UPC smart room – experimental arrangement*

volved with the RIR measured in the UPC smart room [3].

RIR measurements were made using a sweep excitation signal with logarithmically increased frequency. Signal was reproduced from the speaker held on the chest of a person. Seven different positions in the room and four directions of reproduction (orientation of the speaker) were defined, emulating the scenario where a person is giving a talk and moves along the room. Setup may be seen in Figure 1. In the experiment we used 6 microphones placed on the walls 2.4m above the ground. Seven positions, 6 microphones and 4 directions give a total number of 168 RIRs that were used in the experiment.

2.1. ASR system and databases

Experiment was made with the RWTH Aachen university speech recognition system [4] using Catalan Speecon and FreeSpeech databases. The Speecon database is made of real world speech signals recorded in room and outside environments using four microphones (one close-talk and three distant microphones). The Catalan FreeSpeech database was build for an automatic dictation system and consists of close-talk recordings of large vocabulary continuous speech.

For training, approximately 121 hours of recordings data from both databases were selected. In the testing phase only a subset (duration approximately 1.5 hour) of FreeSpeech database was used. Note that the acoustic models are trained in a multi-conditional way and they were not trained specifically for the UPC smart room.

Speech signal was framed applying 25ms long Hamming window with 10ms overlap. Basic speech feature vector consists of 16 Mel frequency cepstral coefficients (MFCC) extended by a voicedness feature [5]. Mean and variance normalization was applied on the cepstral coefficients and fast Vocal tract length normalization (VTLN) to the bank of filters. The temporal context is preserved by concatenating the features of 9 consecutive frames. Prior to the acoustic model training, linear discriminant analysis (LDA) was applied in order to reduce the dimensionality and increase the class separability. Acoustic modeling was using Hidden Markov models and emission probabilities were modeled with continuous Gaussian mixtures sharing one common diagonal covariance matrix.

3. Microphone selection based on energy related measures of the room impulse response

As the speech recognition accuracy varies strongly across the various microphones in the room, our objective is to design a way to select the microphone that offers the highest average accuracy. For that purpose, we want to rely the decision on measures or parameters associated to the RIR that would indicate the degree of harming caused by the reverberation to the signal and, consequently, to the recognition performance. If we were able to compute those measures from the speech signals associated to the various microphones, we would be able to choose the best microphone before entering the recognition system.

To find out candidates for RIR measures that are useful to that purpose, we designed a process as outlined in block diagram in Figure 2. First we trained the acoustic models of a speech recognition system using general databases (the Catalan Speecon and FreeSpeech). Then, we used the trained system to recognize speech signals from FreeSpeech that were convolved with a set of RIRs measured in our UPC smart room.

Let's denote by WER_i the obtained WER corresponding to the i-th RIR. Note that the exact h(n) is known for each microphone. Now we can choose a particular measure M_j, compute its values M_{ji} from every RIR_i and compare (correlate) those values M_{ji} with the corresponding values of WER_i. In this way, we can see the relation between each of the defined RIR measures M_j and the speech recognition rate, and choose the most relevant one(s). Then, such measure(s) can be used for selecting the best microphone before entering the recognition system.

Once relevant measures are identified, question is how to estimate them in the real scenario where the RIR is not known in advance. This problem is still open.

3.1. Energy-based features

RIR can be split into 3 parts: direct sound and early reflections, late reflections and very late reflections. In [2], it was experimentally shown that early reflections are not harming the speech recognition. On the other hand, the middle part (late reflections between approximately 70ms and 2/3 of reverberation time T_{60}) is the harming one.

We investigated relations among WER and different measures M_j based on RIR energy and experimentally identified several candidates for the features:

1. Energy of the whole RIR
2. Energy of direct wave and early reflections (approx. 0-70ms) normalized by energy of whole RIR
3. Energy of late reflections normalized by energy of whole RIR (M_3)
4. Ratio between energies of early and late reflections

Among them, measure M_3 calculated as energy of the late reflections (50ms and 190ms) normalized by the energy of the whole RIR

$$M_{3i} = \frac{\sum_{t=50ms}^{190ms} h_i^2(t)}{\sum_{t=0ms}^{T_i} h_i^2(t)} \qquad (2)$$

showed the highest correlation index between the parameter and the WER (equal to 0.78632). Exact intervals of late reflections

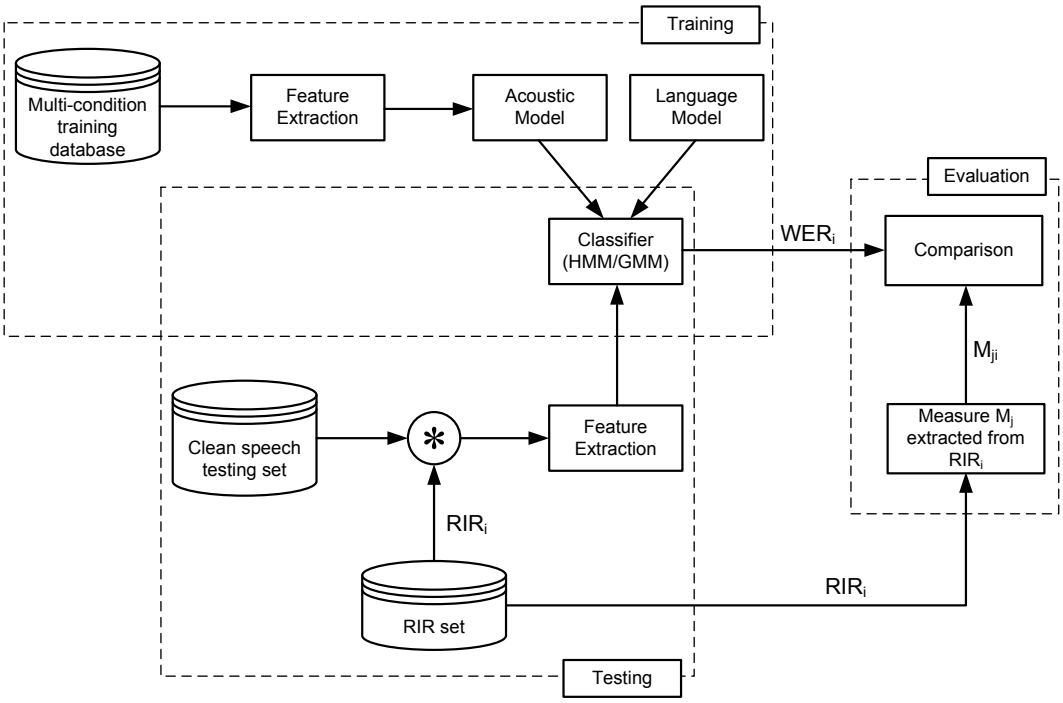

Figure 2: *Block diagram of evaluation of different RIR features*

were identified empirically doing a grid search over different combinations of starting and ending times with the step 10ms. Index i in Eq. 2 denotes the measure taken from RIR$_i$ and T$_i$ is the duration of given RIR.

This observation may be interpreted as lower the energy of late reflections normalized by global energy, lower the WER. It means that the microphone where this quotient of energies is the lowest will be chosen as the most suitable for recognition.

4. Preliminary results and discussion

As a proof of concept, we made an experiment where we compared recognition results when microphone was selected prior to recognition, using only the measure described above (energy of late reflections normalized by global energy) with the case, where the best result was selected from all microphones after recognition (reference).

Results are shown in Table 1. The first column denotes p – position and d - direction of the speaker in the room (Figure 1). All 6 microphones were included in this experiment and for each position and orientation, the most suitable one was chosen. Numbers of selected microphones are in the column 2 and 3. The "Reference" column contains number of microphone that gave the lowest WER after recognition for given position and direction. Column 3 shows what would be our choice if we measured the energy of late reverberation normalized by global energy and made microphone selection before any recognition takes place. Last two columns are showing WERs corresponding to each microphone.

Average WER from all microphones in experiment was 21.37%. This corresponds to the case when microphones would be selected randomly. Next, it may be observed from 28 cases (7 points and 4 orientations) same microphone was chosen 20 times what is more than 71% of cases. Average word error rate when the best microphone is selected after recognition was 14.7% (ideal case) while in case of prior selection average result was only 1.1% worse. This indicates that even if the most appropriate microphone is not chosen, the chosen one is only slightly worse. We further see improvement of 5.6% using our method comparing to random selection (21.37%).

5. Conclusion

In this work we investigate the possibility to use energy based measures from the RIR to make a microphone selection for improving robustness of ASR. We defined a methodology and prepared a setup to search for relevant properties of RIR that may be extracted from the speech signal in each microphone prior to recognition and indicate the input that would presumably lead to the increased recognition rate.

So far we identified and verified one measure: energy of late reflections normalized by energy of whole RIR, and showed that, based only on this single criteria, it is possible to achieve results that are only 1.1% worse in average than the case where microphone would be selected evaluating each input against the speech model separately.

There are several remaining problems to solve. Complementary measures are needed to further improve the selection process. Once more measures are available, they will probably need to be integrated in an efficient way by means of a cost function. Nevertheless, the most important remaining task is to find a method to extract those parameters online from the speech signal.

	Selected microphone		WER [%]	
Position Direction	Reference	$\frac{E_{late}}{E}$	Reference	$\frac{E_{late}}{E}$
p1_d1	1	1	11	11
p1_d2	1	1	13.1	13.1
p1_d3	2	3	16.6	16.9
p1_d4	6	5	14.4	19.3
p2_d1	1	1	12.5	12.5
p2_d2	5	5	15.3	15.3
p2_d3	3	3	13.5	13.5
p2_d4	5	5	12.6	12.6
p3_d1	2	1	15.6	16.8
p3_d2	5	5	12.4	12.4
p3_d3	3	3	12.2	12.2
p3_d4	4	5	16.3	22.1
p4_d1	6	6	18.8	18.8
p4_d2	3	3	10	10
p4_d3	3	3	19.3	19.3
p4_d4	4	5	15.9	18.7
p5_d1	6	6	14.4	14.4
p5_d2	2	2	17.1	17.1
p5_d3	4	4	18.4	18.4
p5_d4	2	5	14.7	21.4
p6_d1	6	5	15.6	19.9
p6_d2	1	1	12.9	12.9
p6_d3	3	4	17.8	22.8
p6_d4	6	6	9.7	9.7
p7_d1	1	1	14.2	14.2
p7_d2	2	2	15.6	15.6
p7_d3	3	3	13.1	13.1
p7_d4	5	5	17.7	17.7
average WER			14.7	15.8

Table 1: *Results - selection based on late reflections normalized by global energy*

6. Acknowledgements

This work has been supported by the project SAPIRE (TEC2007-65470), founded by the Government of Spain, and by the project TECNOPARLA, founded by the Government of Catalonia. Authors would also like to thank to Henrik Schulz for providing help with the acoustic models and the ASR system setup.

7. References

[1] Shimizu Y., Kajita S., Takeda K. and Itakura F., "Speech recognition based on space diversity using distributed multi-microphone," Proc. of ICASSP, 2000, pp. 1747-1750.

[2] Petrick R., Lohde K., Wolff M. and Hoffmann R., "The Harming Part of Room Acoustics in Automatic Speech Recognition", Proc. of INTERSPEECH, 2007, pp. 1094-1097.

[3] Neumann J., Casas J.R., Macho D. and Hidalgo J.R., "Integration of audiovisual sensors and technologies in a smart room," Personal Ubiquitous Comput., vol. 13, 2007, pp. 15-23.

[4] RWTH ASR - The RWTH Aachen University Speech Recognition System, Online: http://www-i6.informatik.rwth-aachen.de/rwth-asr/

[5] Zolnay A. A.,Uter R.S. and Ney H., "Robust Speech Recognition Using a Voiced-Unvoiced Feature," International Conference on Spoken Language Processing, vol. 2, 2002, pp. 1065-1068.