On the Potential of Channel Selection for Recognition of Reverberated Speech with Multiple Microphones

Martin Wolf and Climent Nadeu

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
martin.wolf@upc.edu, climent.nadeu@upc.edu

Abstract

The performance of ASR systems in a room environment with distant microphones is strongly affected by reverberation. As the degree of signal distortion varies among acoustic channels (i.e. microphones), the recognition accuracy can benefit from a proper channel selection. In this paper, we experimentally show that there exists a large margin for WER reduction by channel selection, and discuss several possible methods which do not require any a-priori classification. Moreover, by using a LVCSR task, a significant WER reduction is shown with a simple technique which uses a measure computed from the sub-band time envelope of the various microphone signals.

Index Terms: automatic speech recognition, microphone selection, reverberation, room impulse response

1. Introduction

In reverberant environments, the acoustic waves reflected by the walls and the objects in the room arrive to the microphone attenuated and with different delays. Modified and delayed copies of the original speech signal are summed up in the receiver introducing an undesirable interference, which can be modeled as a convolution of the room impulse response (RIR) with the original speech signal.

The quality of the acquired speech strongly depends on the characteristics of the RIR that describes the acoustic channel between the source and the microphone. Consequently, for a given speech source, the degree of distortion depends on the position of the microphone, so that signals coming from some microphones or channels may be more suitable for ASR, and also for further processing towards their de-reverberation, than others.

The idea of selecting the best channel (or microphone) in terms of recognition accuracy is not new in ASR. Most of the previous works on channel selection (CS) include some kind of classification. In [1], the authors directly use the likelihood at the decoder output to select the channel. In [2], the effect of a feature compensation technique (e.g. mean and variance normalization) on the decoder output is used; the channel with smallest likelihood difference between the compensated and the uncompensated features is selected. Class separability was used for channel selection in [3]. The channel with maximum ratio between inter and intra class separability is chosen as potentially the least distorted one.

However, we can consider an alternative way of approaching CS which does not imply any classification or previous training and may be better suited to real-time processing. The decision is done before entering the recognition system, and it is based on measures or features extracted from the signals corresponding to the various channels. Several possibilities of doing CS based on such approach are discussed in Section 2 of this paper. The experimental setup, involving a large vocabulary continuous speech recognition (LVCSR) task is presented in Section 3. The potential of CS for WER reduction is experimentally shown in Section 4. In Section 5, a new CS method based on measures of degradation of the speech time envelope is proposed, and it is tested in both calibrated and non-calibrated scenarios. Preliminary results show an encouraging relative improvement of more than 28% compared to the case of random CS.

2. Real-time channel selection

The objective in real-time CS is to design an algorithm allowing picking up a microphone which will presumably lead to the highest recognition accuracy at the end. There are two basic questions: (1) what should be the selection decision based on, and (2) how can it be extracted or measured effectively.

2.1. Selection based on RIR related measures

In [4], relations between the different parts of the RIR and the word error rate (WER) of an ASR system were investigated. Authors showed that there are certain components of the RIR that harm the speech recognition more than others. Assuming there is a feature or measure extracted from the RIR which indicates the degree of degradation of the WER, CS could be done from that measure before recognition, provided that the RIR can be estimated for each microphone.

Recently, we presented in [5] a methodology to identify relevant measures for CS, assuming an exact knowledge of the RIR. To find out the candidates, a set of ASR experiments was conducted to measure correlations between different RIR features and the WER.

There are two main problems with this approach. The first one is the estimation of the RIR, or the direct estimation of measures derived from it. As RIRs may change while speech is produced, that estimation should be made online and directly from speech, what may be too demanding in quickly changing environments. A second drawback is the fact that the measure depends only on the RIR so it does not take into account the speech content. Actually, a given RIR affects different utterances in a distinct way, so the minimum WER channels of those utterances may be different. With this method, assuming a static speaker and invariant room conditions, the same microphone is chosen for all the utterances.

2.2. Selection based on position and orientation

It is well-known that if multi-conditionally trained acoustic models are used, the closer the distance of the microphone to the speaker the better the recognition rate (see e.g. [6]). Direct orientation of the speaker towards the microphone is also desirable if we aim to lower the WER.

This fact is suggesting the use in CS of the existing techniques to estimate the position and the head orientation of the speaker. The most straightforward selection criterion would be to pick up a microphone which is the closest and the most directly oriented to the speaker's mouth.

Although by following this approach CS can be easily implemented using only a simple set of rules, there are again some problems associated to it. It relies on a different technology that may not always provide accurate measures and, similarly to the previous approach, one microphone is selected independently of the speech content, so leading to suboptimal decisions. Finally, knowledge about the positions of the microphones is needed, putting additional demands on the system deployment.

2.3. Selection based on signal distortion

Besides the difficulties in the extraction of accurate measures, the above presented CS approaches suffer from a lack of sensitivity regarding the speech content. The third category we consider here is based on measuring the distortion directly from the signal. The effect of reverberation on the speech signal may be observed in several ways. With respect to the time span considered, we can extract measures at the level of the pitch period, at the frame level, or at longer segments. We will consider the last one in this work by looking at the speech time envelope. In fact, the low-pass character of the RIR causes smearing and blurring of the speech envelope. In [7], the effects in the modulation spectrum domain were shown. The same or similar principles may be also applied to the problem of CS.

This approach, where the reverberation effects are measured directly on the signal, seems to be the most appropriate for the following reasons: ideally, a selection can be made for each specific speech segment; no additional technologies are required; and the processing may be well tightened with the feature extraction. The direct effect of reverberation on the speech signal may be observed more easily and with less delay than if it is done indirectly through the RIR or using a high level classifier.

3. Experimental setup

We defined a set of ASR tests where close-talk microphone recordings of continuous speech were convolved with several RIRs measured in the UPC smart room. Convolution was made on an utterance basis, so the RIR does not change along the utterance.

The RIR measurements were made using a sweep excitation signal with logarithmically increased frequency. The signal was emitted from a loudspeaker held on the chest of a person. Seven different positions in the room and four directions of reproduction (orientation of the speaker) were defined. The setup may be seen in Figure 1. In the experiments, we used 6 microphones placed on the walls 2.4m above the ground. 7 positions, 4 directions and 6 microphones give a total number of 168 RIRs.

When the RIR were measured, microphones in the room were (almost) calibrated. Calibration is important to compensate for different attenuations in the electrical path among microphones (different wire length, varying volume set on preamplifier, etc). In the following explanation this case will be simply referenced as calibrated. In the experiments we also used a second set of RIRs, where RIRs from calibrated measurements were de-calibrated attenuating each RIR by a random factor in the range 0-15dB.

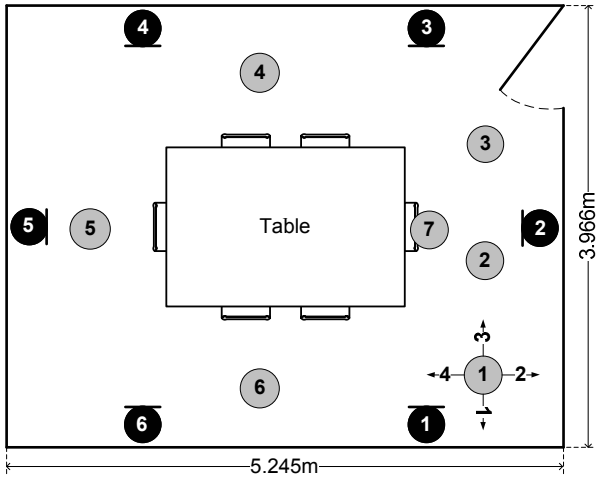

Figure 1: *UPC smart room - experimental arrangement showing 7 positions of the simulated speaker (grey) and 6 microphones (black).*

3.1. ASR system and databases

Experiments were made with the open-source ASR system from RWTH Aachen University [8] using two Catalan speech databases. The Speecon database is made of real world speech signals recorded in rooms and outdoors environments, using four microphones (one close-talk and three distant microphones). The other Catalan database, Freespeech, was build to develop 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, including all four microphones from Speecon. In the testing, only a subset (nearly 1.5 hour, 294 utterances) of the dictation database was used. Those close-talk microphone recordings were convolved with the RIRs before testing.

Note that the acoustic models were trained in a multi-conditional way, but not specifically for the UPC smart room, since it was not used for recording the databases. The specific settings of the ASR system are like those described in [9].

4. Margin of WER reduction by CS

Using the system and scenario described in previous section, the potential of microphone selection may be easily demonstrated. Recognition results using different CS configurations are compared (e.g. randomly selected microphones, ideal selection etc.) here. In these demonstrations the decision was made knowing priory the recognition result, therefore it is just theoretical. Results using real measurements will be presented later. The main purpose of this section is to show what is possible to achieve with CS. Results are presented for both calibrated (C) and non-calibrated (NC) sets.

Results in Table 1 come from the scenario where always the same microphone is chosen independently of position and orientation. It may be seen that WER is more or less equal for all microphones and in average more than 21%. The results are similar in both, C and NC sets. This indicates that attenuation of the RIR introduced in NC case was not harmful to the signal and did not have a negative impact on the recognition performance.

Mic. #	1.	2.	3.	4.	5.	6.
WER_C	20.8	20.3	20.8	23.0	21.4	22.0
WER_{NC}	20.7	19.9	20.6	22.6	21.3	21.6

Table 1: *Always the same microphone*

Table 2 shows the recognition results from experiments when only the best microphone is chosen for each utterance (the first column) or only the second best, etc., ending with the case where the worst microphone would always be selected. Assuming we are able to perfectly decide what microphone is the best for each utterance, the WER in the experiment would be less than 11%. This is the best result we can achieve for this experimental setup selecting one microphone per utterance. Again the WER is similar for C and NC scenario. This suggests that CS may be done equally well, even if the microphones are not calibrated.

Pref.	1.	2.	3.	4.	5.	6.
WER_C	10.8	15.1	18.8	22.6	27.0	34.1
WER_{NC}	10.6	14.9	18.5	22.2	26.7	33.7

Table 2: *From the best possible to the worst CS*

To complete the picture, if a microphone is randomly selected from the set of 6 microphones for each utterance, corresponding WER is equal to 21.4% in the C case and to 21.1% in the NC one. Assuming there are different speaker positions involved in the experiment, selecting a microphone randomly for each utterance is similar to selecting all the time the same microphone and keeping it unchanged during recognition.

There are two important conclusions. The best possible WER that might be achieved in our configuration selecting a microphone per utterance is 10.8% and 10.6% (perfect CS). If we do not make any selection or the microphone is selected randomly, the WER is around 21.4%. With perfect CS we might achieve around 50% relative improvement with regard to the case with no selection.

Results in this section show what can be hypothetically achieved using prior knowledge of the WER. In the next section, some methods are presented where no prior knowledge of the WER is available.

5. Implementation examples and results

In Section 2 we discussed several approaches to CS. Some implementations of them are presented here including further explanations and recognition results.

5.1. RIR energy related measures

RIR can be split into 3 parts: direct sound and early reflections, late reflections and very late reflections. 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 [4].

We investigated relations among WER and different measures in [5] based on RIR energy and experimentally identified several candidates for the features. Among them, a measure of energy of the late reflections (50ms and 190ms) normalized by the energy of the whole RIR showed he highest correlation index with the WER (equal to 0.78632). Exact intervals of late reflections were identified empirically doing a grid search over different combinations of starting and ending times with a 10ms step.

This observation may be interpreted as the lower the energy of late reflections normalized by the global energy, the 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. Using this method we achieved 15.8% WER for the calibrated case. This result was obtained assuming a known RIR.

5.2. Position and orientation

In our room, depicted in Figure 1, the exact positions of microphones and speakers are known and remain unchanged along the utterance. It is reasonable to assume that close distance and more direct orientation of the speaker to the microphone indicate better channel. If we use only the closest distance and ignore the orientation, the average WER is nearly 20% for both C and NC sets. This is partly because it might happen that a microphone was selected as preferred one when speaker actually had it behind his back. If we take only the orientation into account, average WER is little above 16% in both scenarios. Our acoustic models were trained using speech uttered directly towards the microphone. This of course contributes to better performance of orientation based CS. More sophisticated training using microphones recording from different directions may bring further improvements. Distance and orientation measures may also be efficiently combined.

5.3. Distortion of the speech signal

In [10], the authors investigated a technique for CS which selects the channel with maximum signal-to-noise ratio (SNR), despite the fact that the SNR measure is related to the additive noise, not to the convolutive noise involved in reverberation. The noise power was estimated in silence portions, but no information about how the silence intervals were determined is provided. To avoid the dependency on the particular speech activity detection (SAD) system, in our implementation we estimate the noise power using a noise recording that was made for each microphone when the RIRs were measured.

In our computation of the SNR, the "signal" is clean speech convolved with the RIR. Noise in the SNR expression is speech convolved with the measured noise and further convolved with the inverse of the excitation sweep signal (used for RIR measurement). This may not be done in a real situation, but we only want SNR measure as a reference.

Other measures are proposed here which are extracted from the signal envelope. We define the envelope as a time sequence of the frame energies. Speech utterances are framed, multiplied by a 30ms long Hamming window, and the energy is calculated for each frame. That calculation is done either in the full band or in the subbands. In our experiments we used 20 mel-scaled filters for subband analysis. That number was not optimized. After computing the frame energies, their dynamic range is compressed powering each sample by 1/3.

Along with the SNR, we have tested, as reference, the energy, which is measured as the average of the frame energies along the utterance before the dynamic range compression. In both cases, the selected channel is the one that shows the maximum value. Moreover, two other measures are proposed trying to avoid the drawbacks of both the energy and the SNR.

Due to reverberation, the low energy spectral valleys of close-talk speech are filled with energy coming from the preceding peaks, smearing the time envelope. The amount of smearing may be observed either in the variance of the envelope or in the modulation spectrum domain [6].

In order to compensate for a lack of calibration we use the variance of the envelope normalized by the average of the envelope values. The decision rule to select the channel is to look for the maximum of the normalized variance.

The measure extracted from the modulation spectrum for a given band is calculated from the absolute value of the Fourier transform of the time envelope in that band. We integrate the area between 0.25Hz – 16Hz, and normalize it with the average energy as it is done with the variance. Again, the decision rule is based on selecting the channel with the maximum of that normalized modulation spectrum area (MSA).

When working in subbands there is a measure extracted for each subband. However, only one microphone is selected for the whole utterance. This was made in the following way. For each subband, measures were normalized to be within the interval [0, 1], dividing by the maximum in that subband. The microphone giving the best average measure over all subbands was selected for the utterance.

Four measures were evaluated: (1) energy of the signal, (2) SNR, (3) normalized variance of the envelope; and (4) normalized MSA. Recognition results are shown in Table 3 in terms of WER. The analysis in subbands shows clear advantage over the full band processing. The reason for that is the frequency dependent behavior of the RIR.

The simple energy measure works fine for the calibrated case. However, it can not compete with other measures for the NC case, due to the fact that it is not normalized.

The good behaviour of the SNR-based method for the NC case may be explained in the following way. As the additive noise in the room may be assumed homogenous among microphones, the measured noise energy conveys the amplification factor of each electrical channel, so the used SNR measure actually is a way of compensating the lack of calibration. In this way, the SNR measure avoids the need of calibration, though it requires the estimation of the noise energy. In our case it is measured directly in the room, but more practical implementation would require a SAD system.

The normalized variance and the normalized MSA perform similarly to SNR without requiring estimation of the noise energy. The CS method based on normalized MSA performs slightly better in both C and NC scenarios.

	Calibrated	
	full band	20 - mel
Energy	18.3	16
SNR	18.3	15.8
Norm. VAR	19.7	16
Norm. MSA	18.8	**15.4**

	Non-calibrated	
	full band	20 - mel
Energy	20.7	20.6
SNR	18	15.5
Norm. VAR	19.4	15.8
Norm. MSA	18.6	**15.2**

Table 3: Speech recognition results for the signal distortion based methods

As mentioned before, all the measures in the experiments were extracted from whole utterances. Additional experiments showed that similar results can be achieved using speech segments which are only 1s long.

6. Conclusions

In this paper we have tried to show that the recognition accuracy can benefit largely from a proper selection of the acoustic channel, i.e. microphone, when ASR is carried out through the use of multiple distant microphones in a room. The potential of CS is illustrated with a LVCSR task and a state-of-the-art ASR system, showing that, in the ideal case, CS achieves about 50% relative improvement in comparison to the case of random or no selection. Several alternative approaches for selecting the channel before recognition without requiring classification are discussed. Measures of signal distortion based on either the variance of the time envelope or the modulation spectrum, which are not affected by a lack of calibration and can be estimated in a short interval, are used in this initial work. In spite of their simplicity, around 28% relative improvement in terms of WER is achieved with them, comparing to no or random selection, what is more than half of the existing margin for improvement.

7. Acknowledgements

The authors want to thank Henrik Schulz for providing the ASR models and setups. This work has been supported by the Spanish project SAPIRE (TEC2007-65470), and the Catalan project TECNOPARLA.

8. 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.

[2] Obuchi, Y., “Multiple-microphone robust speech recognition using decoder-based channel selection”, Workshop on Statistical and Perceptual Audio Processing, Jeju, Korea, 2004.

[3] Wölfel, M., “Channel selection by class separability measures for automatic transcriptions on distant microphones”, Proc. of INTERSPEECH, 2007.

[4] Petrick R., Lohde K., Wolff M. and Hoffmann R., “The harming part of room acoustics in automatic speech recognition”, Proc. of INTERSPEECH, 2007.

[5] Wolf M. and Nadeu C., “Towards microphone selection based on room impulse response energy-related measures”, Proc. of I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages, Porto Salvo, Portugal, 2009, pp. 61-64.

[6] Pujol P., Padrell J., Nadeu C., and Macho D., “Speech recognition experiments with the SPEECON database using several robust front-ends”, Proc. of International Conf. on Spoken Language Processing (ICSLP) 2004, pp. 2105-2108.

[7] Houtgast, T. and Steeneken, H. J. M., “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria”, The Journal of the Acoustical Society of America, vol. 77, Mar. 1985, pp. 1069-1077.

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

[9] Schulz, H., Fonollosa J. A. R. and Rybach D., “Transcription of Catalan broadcast conversation”, Text, Speech and Dialogue, vol. 5729/2009, pp. 154-161, Springer Berlin / Heidelberg.

[10] Wölfel, M., Fügen, C., Ikbal S., Mcdonough, J. W., “Multi-source far-distance microphone selection and combination for automatic transcription of lectures”, Proc. of INTERSPEECH, 2006.