Channel Selection Using N-Best Hypothesis for Multi-Microphone ASR

Martin Wolf and Climent Nadeu

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

10.21437/Interspeech.2013-266

Abstract

If speech is captured by several arbitrarily-located microphones in a room, the degree of distortion by noise and reverberation may vary strongly from one channel to another. Channel selection for automatic speech recognition aims to rank the signals according to their quality, and, in particular, to select the best one for further processing in the recognition system. To create this ranking, we propose here to use posterior probabilities estimated from the N-best hypothesis of each channel. When evaluated experimentally, this new channel selection technique outperforms the methods published so far. We also propose the combination of different channel selection techniques to further increase the recognition accuracy and to reduce the computational load without significant performance loss.

Index Terms: Automatic speech recognition, channel selection, acoustic likelihood, n-best hypothesis, multiple microphones

1. Introduction

In applications where distant-talking microphones are used for Automatic Speech Recognition (ASR), both additive noise and room reverberation are major factors of recognition rate degradation [1]. If multiple microphones are available, signal combination techniques (e.g. beamforming) are often used to improve the quality of the recorded speech. However, this combination may not be possible, or the quality of the combined signal may not be always better than the quality of the signal from the single best channel. Let's assume a scenario, where the microphones are arbitrarily located and have different characteristics: a meeting room where some microphones are hanging on the walls, others standing on the table, or they are built in the personal communication devices of the meeting participants. In such case, the combination of the signals becomes increasingly difficult and the ASR performance may actually gain from reducing the number of channels, or even from selecting just the best channel for recognition. Ideally, this would be the channel that leads to the lowest word error rate (WER) after recognition. Since WER is unknown during the recognition, a different measure, as correlated as possible with WER, is needed. Two kinds of measures have been proposed for channel selection (CS) in the literature: signal-based and decoder-based.

Signal-based methods operate in the front-end of the recognition system. They use some signal-processing measure to estimate the distortion from the speech signal or the channel characteristics. A typical example of such measure is the signal to noise ratio that was used for CS in [2] and [3]. In [4], the possibility to use information about the relative position and orientation of speaker and microphone for CS was evaluated. The function that describes the propagation of the sound in a room is called room impulse response (RIR). In [5], it was shown that certain parts of the RIR harm the speech recognition performance more than others. Based on that, a CS method that uses a measure extracted from the channel characteristics, similar to the well know direct-to-reverberation ratio, was presented in [6]. To avoid the requirement of a sufficiently accurate RIR estimation, we proposed another CS method in [4], where a time envelope based measure is extracted directly from the speech signal.

The signal-based methods just aim at finding the channel that carries a signal that is the most similar, in terms of distortion, to the type of signals observed in training (it would be the least distorted one, if clean speech was used to train the acoustic model). The decoder-based techniques, on the other hand, work in close cooperation with the decoder, so the measures they use should better reflect the decoder's preference than the signal-based ones. In [2] an approach based on feature normalization was introduced. In that approach, a feature normalization technique (e.g. mean and variance normalization or histogram equalization) is first applied to each channel, and both, the original and the compensated feature streams are recognized. The channel with the smallest difference between the recognized word sequences from the original and the compensated versions is selected. A CS method using a class separability measure was presented in [7]. The channel maximizing that separation measure extracted from speech feature vectors is selected for recognition. This approach can be implemented as stand-alone or decoder-based, and depending on that it is classified as either signal or decoder-based CS method. A straight forward decoder-based approach using the acoustic likelihood was presented in [8]. Likelihood, by itself, is not a good indicator of the signal quality if signals are coming from different channels. We addressed this issue in [9] by using a pairwise likelihood normalization across channels.

In this paper we use an alternative way to normalize the likelihoods. We have adopted a method commonly used in confidence measuring to estimate the likelihood normalization factor from the N-best hypothesis. The details are described in Section 2. Reverberant environments were used to evaluate the digit recognition performance of the new CS method and to compare it with the methods published so far. The experimental setup and results are presented in Section 3. In Section 4 we demonstrate that the combination of different CS methods allows further recognition rate improvements. We also propose a serial combination of signal and decoder-based CS methods, and show that it reduces the computational load without any significant loss in recognition performance.

2. CS based on the N-best hypothesis

In the conventional ASR systems the well known Bayes' rule is used to convert the prior probability of a word sequence $P(w)$

to the posterior probability of that sequence given the observed feature vector $P(w \mid O)$. In the multi-channel case the observation sequence is different for each channel, so the posterior probability for channel m can be expressed as

$$P(w_m \mid O_m) = \frac{p(O_m \mid w_m)P(w_m)}{p(O_m)}, \qquad (1)$$

with $p(O_m \mid w_m)$ being the acoustic likelihood in that channel. The posterior probability could be used as a CS measure, but it is not usually computed, because the normalization term $p(O_m)$ does not depend on the word sequence and can therefore be ignored. In a single-channel case this is not a problem, but unless normalized, the scores provided by the recognition system for different channels are not in the same scale and can not be directly compared.

The lack of normalization is also the key problem for confidence measuring, so many solutions may be found in that area [10, 11]. In this work, we adapt the N-best list approach and apply it as follows. It is a well known fact [12], that $p(O)$ may be computed as

$$p(O) = \sum_{w \in \Omega} p(O \mid w)P(w), \qquad (2)$$

where Ω is the set of all possible word sequences for O. Apparently, without any constraints this enumeration is not feasible, so some approximations are required. Let w^n be the n^{th} hypothesis in the N-best list. The $p(O)$ may then be approximated by the finite sum

$$p(O) \approx \sum_{n=1}^{N} p(O \mid w^n)P(w^n), \qquad (3)$$

as it was done for example in [13] or [14].

Finally, based on the above reasoning the CS measure in our N-best approach is computed as

$$C_m = \frac{p(O_m \mid w_m^1)^{1/\alpha}P(w_m^1)}{\sum_{n=1}^{N} p(O_m \mid w_m^n)^{1/\alpha}P(w_m^n)}, \qquad (4)$$

where n is the hypothesis index in the N-best list of channel m. The acoustic model likelihoods $p(O_m \mid w_m)$ usually have a very large dynamic range. An appropriate scaling factor α must be applied to them, otherwise the summations are often dominated by the largest value. The value of α can be estimated a-priori using a development corpus. Another option is to set it equal to the number of frames, as we did in this work. If the acoustic model likelihoods provided by the recognition system are in the log scale, setting α to this value is equivalent to divide the log likelihoods by the number of frames, which results in an average log-likelihood per frame.

3. Experiments

3.1. Experimental setup

The experiments were conducted with 2 databases. The first one was TIDigits, a well known database of connected digits in English [15], and the second was the Meeting Recorder Digit (MRD) corpus [16], a collection of connected digit strings recorded in a real meeting room at the International Computer Science Institute (ICSI) as a part of the ICSI Meeting corpus data collection [17].

In the first case, the original close-talking recordings from TIDigits were downsampled to 16 kHz and convolved with a set of real RIRs, which were measured in the UPC smart room [6]. In the tests, 6 omni-directional microphones installed on the walls of that room were used. The RIRs corresponding to seven different positions and four orientations of the speaker were recorded for each microphone. The reverberation time T_{60} is around 540 ms. Only the utterances from adult speakers were included in the experiments. The acoustic model for the test with this database was trained on clean speech, using the standard training set of TIDigits.

Table 1: WER using a single microphone for convolved TIDigits.

Microphone	1	2	3	4	5	6
WER	23.5	22.9	24.5	22.1	21.4	21.6

Table 2: WER using a single microphone for the MRD corpus.

Microphone	6	7	E	F
WER	14	16.4	17.1	15.6

In the second case, the MDR corpus was used to test the CS methods also with real signals. In this corpus the sequences of digits are read by the meeting participants and recorded in parallel using 4 distant-talking microphones that were placed on the table in the meeting room. There are 29 speakers who, in summary, read 2790 utterances over 22 sessions (not all speakers participated in all sessions). Both, native and non-native speakers were included in the tests. In this case, the acoustic model was again trained using TIDigits, but the utterances were first randomly convolved with the RIRs that were measured in the empty room where the MRD corpus was recorded. The acoustic model was trained on reverberant data in this case to check if CS can improve recognition performance also when it is used in combination with other robust ASR methods, which in this case is matched condition training.

A continuous density hidden Markov models (HTK toolkit) based system [18] was used, applying the setup commonly used for TIDigits. The 11 models for words (digits zero, oh, one, ..., nine) have 16 states, the silence model has 3 states, and a short pause model 1 state which is shared with the middle state of the silence model. There are 3 Gaussians per state for the words, and 6 for the silence model. Standard MFCC features were extracted from 20 mel-frequency bands. The feature vector consisted of 12 cepstral coefficients without the 0th coefficient, frame energy, delta and acceleration features. The size of the vector was therefore 39. Frame length was 25 ms and frame shift 10 ms. Finally, Mean and Variance Normalization (MVN) was applied.

The recognition WER of the baseline system, when trained and tested with clean TIDigits, is 0.6%. As expected, in the two distant-talking microphone environments, a significant ASR performance degradation may be observed. The WERs for every microphone, and for both the convolved TIDigits and the distant-talking recordings in the MRD corpus, are presented in Tables 1 and 2, respectively. In the UPC smart room the microphones are numbered from 1 to 6. In the MRD corpus the 4 microphones are labeled as 6, 7, E, and F. Each presented WER is calculated as an average over all positions, orientations and speakers. In other words, it shows the performance of the ASR system as if only that particular microphone was present in the room.

3.2. The size and structure of the N-best list

The number N of hypothesis included in the CS measure in (4) is a free parameter. In theory, the more hypothesis are used, the more precise is the approximation of $p(O)$ in (3). However, the objective in CS is not to estimate the posterior probability with the highest precision, but rather to minimize WER. Also, from a practical point of view, we do not want to generate large and computationally costly N-best lists.

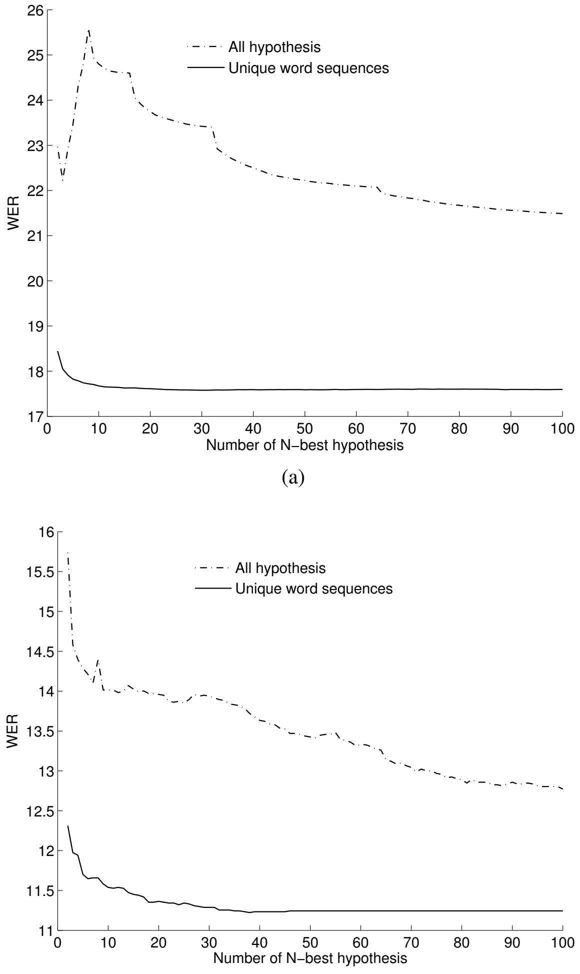
An important factor is what kind of hypothesis is included in the CS measure estimation. The N-best algorithm may generate a lot of hypothesis that have the same word sequence as the first one and differ only in the presence and position of the silence label (e.g. 'silence one two' is the same word sequence as 'one silence two'). The calculation of the CS measure from such a redundant list is not very efficient, because it means that we are aiming to maximize a ratio of likelihoods using almost the same word sequences.

In Figure 1 we show how increasing the number of hypothesis in the estimation of the N-best CS measure influences the recognition performance. There are two curves presented for each test set. One curve corresponds to the N-best measure extracted using all hypothesis in the N-best list, while the other corresponds to the case when the generation of the N-best list was constrained to give only unique hypothesis, i.e. when the redundancy of word sequences is avoided. As we can observe, the latter variant performs much better. Even if only 2 hypothesis (the first point of the graph) are used to extract the CS measure, the WER is lower than for any single channel in Tables 1 and 2. With increasing number of hypothesis, the WER decreases even more, until it saturates. Contrarily, the WER for the first variant with all the hypothesis decreases much slower and, for the case of convolved TIDigits, it even grows at the beginning.

3.3. Comparison with other CS methods

In Table 3, the performance of the new CS method based on a N-best list (NB) is compared with other CS techniques presented in the literature: envelope variance (EV) [4], normalized likelihood (NL) using the pairwise normalization [9], and feature normalization (FN) [2]. N-best lists with N=40 were used to extract the CS measure. The case of random CS (RND) is also included for comparison purposes. Note that the WER for random CS is equal to the average of WERs from the individual microphones in Tables 1 and 2. A single utterance was used to extract every CS measure, so a different channel can be selected for each utterance. For the FN-based method, we used non-normalized features (MFCC + energy $+\Delta$ + $\Delta\Delta$), and then we applied MVN to get the alternative stream. Normalized features were used for recognition to obtain the WERs, as was done for the other CS methods.

We observe that all techniques perform much better than random selection. The relative improvement with respect to the random case is shown in parenthesis. The NB technique outperforms all the other methods. Though class separability-based CS method [7] was also tested, it is not listed in the table. In the tests only a single utterance was used to extract the measure. Since the utterances in the databases are short, the separability measure could not be reliably estimated, and so the performance of this method was close to the random selection case.



(a)

(b)

Figure 1: CS performance in terms of WER for different numbers of N-best hypothesis, using all or unique word sequences for (a) the convolved TIDitits and (b) the MRD corpus.

4. Combination of CS methods

CS may benefit from combination of various methods in two ways. Firstly, since the CS measures are extracted from different domains, they may be complementary and their combination could increase the robustness of the CS system. We will refer to this case as a parallel combination. Secondly, decoder-based methods are computationally expensive. This may be a problem if number of channels is high and fast system response is required. Therefore, to reduce the number of channels, some computationally cheap signal-based CS method may be first applied in the front-end and in the next step more precise selection can be made on the reduced channel set using the decoder-based methods. This case will be referred as a serial combination and its block diagram is shown in Figure 2.

In the parallel combination all CS measures are extracted for all channels. Usually the measures are in different scales. In this work we applied a simple combination strategy, where we first rank the channels separately for each measure, and then select the channel with the highest ranking in average as

$$C = \arg\min_{m} \sum_{i} r_m(i), \tag{5}$$

Table 3: CS performance in terms of WER for convolved TIDigits and MRD corpus.

CS methods	Conv. TIDigits	MRD
RND	22.7	15.8
EV	19.2 (15.4%)	12.8 (19%)
NL	19.5 (14.1%)	12.1 (19.6%)
FN	19.6 (13.7%)	11.6 (26.6%)
NB	17.6 (22.5%)	11.2 (29.1%)

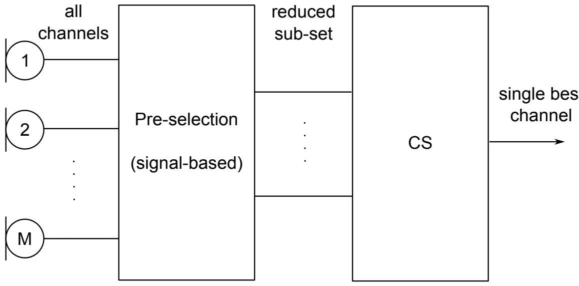


Figure 2: Block diagram of serial combination of signal and decoder-based methods.

where $r_m(i)$ is the ranking position of the channel m according to the measure i.

The performance of CS methods using parallel combination is shown in Table 4. We may see that the combination of all the presented CS methods except of NB, performs better than any method alone. However, it still does not reach the performance of the NB method. If we include also NB technique we observe a further WER reduction. The results indicate that combination of methods may help, but also demonstrates the superiority of NB method that single-handedly outperforms other combined CS methods.

In Figure 3, the performance of CS methods in serial combination is shown. Each point corresponds to the number of channels after the pre-selection step. There are 2 curves for each setup, one for the random pre-selection, and the other for the case when the channel is pre-selected using the EV method. A single channel is selected from the subset in the second step using parallel combination of all methods (NL + FN + EV + NB).

The first point of the graph shows the recognition performance if only one channel is pre-selected, either randomly, or using the EV-based CS method. The WERs are the same as in Table 3 for RND and EV case. The last point of the graph is the variant without pre-selection, when all available channels are passed to the second step. We can observe that if EV is used to pre-select the channels in the case of convolved TIDigits, the computation load may be reduced almost by half (only 3 channels are used after pre-selection), and the relative WER increase is only 0.6% compared to the case when all channels are used. On the other hand, if channels are pre-selected randomly, the relative WER increase is 10%. In the MRD case, we may in a similar way reduce the computational load by 25% without any loss in performance. If we removed one channel randomly, the relative WER increase would be more than 6%.

Table 4: CS performance in terms of WER using parallel combination.

CS methods	Conv. TIDigits	MRD
NL + FN + EV	17.5 (22.9%)	11.2 (29.1%)
NL + FN + EV + NB	17 (25.1%)	11 (30.4%)

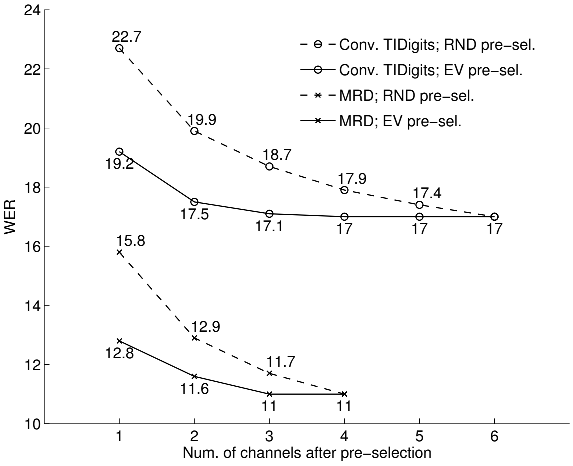


Figure 3: CS performance in terms of WER using serial combination of the measures.

5. Conclusions

In this paper we presented a new CS method that operates in the back-end of the ASR system. It is based on the acoustic likelihoods and uses N-best hypothesis to tackle the normalization problem present in the multi-channel case. When the new technique is experimentally compared to other CS methods presented in the literature so far, it consistently shows a better recognition performance (relative improvement up to almost 30% in comparison to the case of random CS was observed).

Another contribution of this work is the combination of different CS methods. We showed that simple combinations of signal and decoder-based methods lead to further WER reduction. Also, the combination may help to reduce the computational load by half with just a slight loss in terms of WER. In fact, the computationally cheap EV signal-based method is used at the front to reduce the number of input channels, so the more expensive decoding operation does not have to be made for all channels.

In the reported experiments we focused only on selection of the best channel, but indeed the CS techniques may be also used to select more than one channel. Presumably, further improvements can be achieved with more sophisticated combination schemes. In particular, the combination of word sequence hypothesis from different channels might lead to further WER reduction.

6. Acknowledgements

This work was supported by the Spanish project SARAI (reference number TEC2010-21040-C02-01).

7. References

[1] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Hoboken, NJ: Wiley, 2009.

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

[3] 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. of INTERSPEECH*, 2006.

[4] M. Wolf and C. Nadeu, "On the potential of channel selection for recognition of reverberated speech with multiple microphones," in *Proc. of INTERSPEECH*, Tokyo, Japan, 2010, pp. 80–83.

[5] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, "The harming part of room acoustics in automatic speech recognition," in *Proc. of INTERSPEECH*, 2007, pp. 1094–1097.

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

[7] M. Wölfel, "Channel selection by class separability measures for automatic transcriptions on distant microphones," in *Proc. of INTERSPEECH*, 2007, pp. 582–585.

[8] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multi-microphone," in *Proc. of ICASSP*, vol. 3, 2000, pp. 1747–1750.

[9] M. Wolf and C. Nadeu, "Pairwise likelihood normalization-based channel selection for multi-microphone ASR," in *IberSPEECH*, Madrid, Spain, Nov. 2012.

[10] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 288—298, 2001.

[11] J. Hui, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[12] R. O. Duda, P. E. Hart, D. G. Stork, and D. G, *Pattern Classification*. Wiley, 2001.

[13] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," in *Proc. ICASSP*, 1995, pp. 129—132.

[14] A. Stolcke, Y. König, and M. Weintraub, "Explicit word error minimization in n-best list rescoring," *Proc. of EUROSPEECH*, vol. 1, pp. 163—166, 1997.

[15] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. of ICASSP*, vol. 3, 1984, pp. 111–114.

[16] ICSI, "ICSI Meeting Recorder Digits corpus," http://www1.icsi.berkeley.edu/Speech/mr/mrdigits.html, 2003, [Online; accessed 24-January-2013].

[17] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macas-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI meeting project: Resources and research," in *Proc. of ICASSP 2004 Meeting Recognition Workshop*. Prentice Hall, 2004.

[18] S. Young and et. al., *The HTK Book (for HTK Version 3.4)*. Cambridge University Press, 2006.