WORD BOUNDARY AGREEMENT TO COMBINE MULTI-MICROPHONE HYPOTHESES IN DISTANT SPEECH RECOGNITION

Cristina Guerrero and Maurizio Omologo

Fondazione Bruno Kessler-Irst
via Sommarive 18, 38123 Trento, Italy
{guerrero|omologo}@fbk.eu

ABSTRACT

In this paper we propose a technique for combining hypotheses generated in a multi-microphone setting, which exploits complementarity and collective agreement among ASR outputs of different channels. The technique draws upon the information encoded in the available set of word lattices. As a first step, we identify word boundaries in which a comprehensive inter-channel agreement is found; then, these boundaries are used to reduce the global hypothesis search space. Global word posterior probabilities are estimated for the candidate words associated to each of the bounded segments. As a result, a single combined confusion network is generated from the multiple lattices. This approach offers a novel perspective to state of the art solutions based on confusion network combination. Promising results were obtained from an experimental evaluation in a simulated domestic environment equipped with a distributed microphone network. The development and test sets were simulated using real impulse responses estimated for a large set of microphone-speaker position pairs.

Index Terms— Distant speech recognition, hypothesis combination, multi-microphone, confusion networks

1. INTRODUCTION

The potential of Automatic Speech Recognition (ASR) through distant non-intrusive sensors is undeniable. Several research activities are addressing the numerous challenges introduced by this kind of interaction under different conditions [1]. Of particular interest are application scenarios as the home automation and the support of physically impaired people. The “Distant-Speech Interaction for Robust Home Applications” (DIRHA) Project[1] represents one of the contexts in which this topic is being investigated. The distribution of multiple microphones within an enclosure is a common strategy followed to develop distant-speech interaction systems. In such a setup, different approaches can be applied, based on either the selection of one microphone, or a cluster, or the fusion of information extracted from all the available microphones. Channel selection can be performed in different possible ways [2, 3], limiting the posterior processing to a selected set of signals or components. On the other hand, in fusion methods the redundancy and complementarity of the whole available information can be exploited at signal, feature, model, or recognition hypothesis level. A traditional method of signal fusion is beamforming, which can be effective with certain microphone-network geometries. If the microphones are sparsely located from each other, spatial aliasing and other artifacts can strongly affect the resulting fusion. Another fusion method, characterized by a higher complexity than previously described approaches, is hypothesis level combination. In this case, one exploits all the available information captured by different sensors and processed at ASR level, without any constraint on the geometry of the microphone network.

In the past, Confusion Network Combination (CNC) [4] and Recognizer Output Voting Error Reduction (ROVER)[5] techniques were proposed, which can be effectively applied to distant speech recognition with largely spaced microphones [6, 7, 8]. These techniques aim to build a compact representation of the word hypothesis space, from which the most likely recognized sentence is derived.

In this work, an alternative method for hypothesis combination is investigated, which is characterized by a multi-microphone confusion network extraction based on the agreement of information shared by lattices derived from different microphone signals. The major difference with respect to CNC is the use of temporal information that is embedded in the ASR output.

This paper is organized as follows. Section 2 introduces the standard technique used in hypothesis combination. In Section 3, the proposed method is described. The experimental activities and results obtained on a simulated domestic environment are reported in Section 4, after which the conclusions and future work are discussed.

[1] The research leading to these results has partially received funding from the European Union’s 7th Framework Programme (FP7/2007-2013) under grant agreement n. 288121 DIRHA http://dirha.fbk.eu

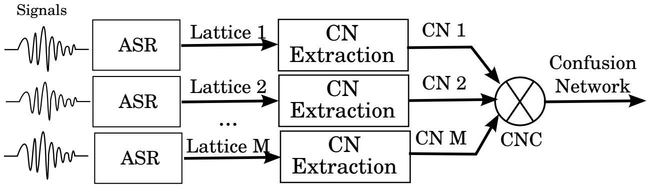


Fig. 1. Block diagram of CNC.

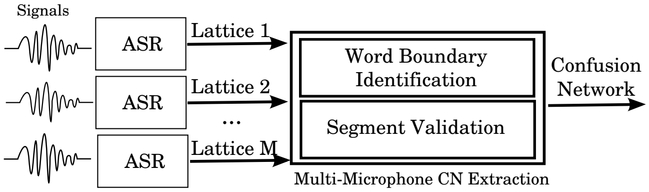


Fig. 2. Block diagram of the proposed method.

2. HYPOTHESIS COMBINATION

One of the first approaches proposed for hypothesis combination is ROVER [5], a voting procedure originally applied to the word sequences provided by a set of speech recognizers. This mechanism was later used for CNC[4], providing a remarkable improvement over ROVER, mainly because the combination was performed on a hypothesis space, rather than on individual word sequences. CNC is available in the SRILM toolkit [9].

The basic unit of CNC is the confusion network (CN) [10], which is a compact representation of a lattice. As the lattice, the CN is a directed graph, but follows a set of particular properties: a) the general network is formed by a sequence of confusion sets or bins, b) each confusion set is composed by one or more word candidates (or an instance of a SILENCE), c) each candidate in a confusion set has a posterior probability, d) the sum of the posterior probabilities of the candidates in a confusion set is equal to 1, e) the best hypothesis of the CN is extracted by selecting the word with the highest posterior probability at each confusion set. In the standard CN extraction procedure, the best path in the lattice is selected as the basis structure for the final CN. Then, an iterative alignment method optimizes the decision of assigning a word to a confusion set, or inserting it in a new one inside the final network.

In order to apply CNC, the lattices generated by the individual recognizers are transformed into CNs. Once the CNs have been extracted, a voting method is applied at a final stage to combine the different CNs into a single CN (see Figure 1). A weight can be assigned to the individual CNs before their combination, which can have a considerable impact on the resulting network. Another factor affecting this process, is the order in which the networks are combined.

The CNC technique has been explored in multi-microphone distant speech recognition tasks [6, 7, 8] and has not evidenced any significant improvement in comparison to signal based approaches (e.g., beamforming). Even though channel selection criteria have been explored for CNC in a multi-microphone scenario, an optimal solution for selecting or properly assigning weights to the multiple CNs has not been identified yet. In the multi-microphone framework addressed here, the issue of finding the optimal arrangement in the sequence of CNs becomes even more critical, since the number of possible permutations increases with the number of microphones. An exploration at the first step of CNC, namely at the extraction of a CN from a lattice, indicates that the best-path based alignments may provide substantially different resulting structures, even from lattices that encode the same utterance. A posterior integration of these different structures may introduce recognition errors.

3. INTER-MICROPHONE COMBINATION OF INFORMATION

A word lattice incorporates data already evaluated at acoustic and linguistic level, keeping the most relevant pieces of information as a hypothesis space. If a speech source is captured by M synchronized distant microphones, the corresponding signals will differ considerably from one another depending on the position and orientation of the sound source. However, it is reasonable to think that the word lattices generated from the ASRs decoding the different microphone signals, though diverse in structure, share some relevant information concerning the spoken utterance. In this work, we assume that the most relevant shared information are the word time boundaries and the occurrences of the words within these boundaries. The inter-microphone agreement of the information within these boundaries guides the construction of the global CN through a multi-lattice combination, as shown in Figure 2.

The information captured by multiple timed word lattices is then incorporated in a unique representation, thus performing in a coherent way, independently of the order of the lattices to be combined. No particular optimization of the sequence or weighting of the lattices is required. The technique relies on: a) the identification of word boundaries that will constitute possible confusion sets in the final network; b) the validation of the segments denoted by the identified boundaries. The latter step involves the search of potential word candidates within the selected boundaries, and the estimation of a posterior probability for each of the candidate words.

3.1. Inter-microphone Word Boundary Agreement

In this work, a straightforward approach is adopted to identify the potentially valid word boundaries, within which a word should be recognized. The approach collects the boundaries of all the links in all the available lattices, removing those links whose posterior probabilities are below an empirically identified threshold. A peak selection over the cumulative occurrence of the boundaries in all the lattices is performed (as shown in Figure 3a), for the identification of the boundaries to be used in the next stage of the processing.

3.2. Segment Validation

The previously identified temporal boundaries $B1, ..., BN$ are used to determine a set of search segments within the lattices. A temporal threshold Δ is set to define a range of frames around each boundary (see Figure 3b). This threshold is dynamically computed as a function of the segment length. For the generic i^{th} temporal boundary, all the links starting in the range from $(B_i - \Delta)$ to $(B_i + \Delta)$, and ending in the range from $(B_{i+1} - \Delta)$ to $(B_{i+1} + \Delta)$, determine a set of potential candidate words to include in the confusion set. A global posterior probability is estimated for each of these words by first computing an intra-microphone posterior score, inspired by the confidence measure used in [11]. For each microphone j and starting boundary index i, the intra-microphone posterior probability C assigned to the l^{th} word W_{lij} is computed as:

$$C\left([W_{lij}, B_i, B_{i+1}]\right) = \sum_{\substack{[w;\tau,t]: \\ [B_i-\Delta\leq\tau\leq B_i+\Delta], \\ [B_{i+1}-\Delta\leq t\leq B_{i+1}+\Delta]}} P\left([W_{lij}, \tau, t]\,|x_1^T(j)\right) \qquad (1)$$

where $P\left([W_{lij}, \tau, t]\,|x_1^T(j)\right)$ corresponds to the posterior probability of the link characterized by the word W_{lij} given the observation sequence $x_1^T(j)$ related to the lattice derived from the j^{th} microphone.

The resulting intra-microphone scores are then averaged over all the channels as follows:

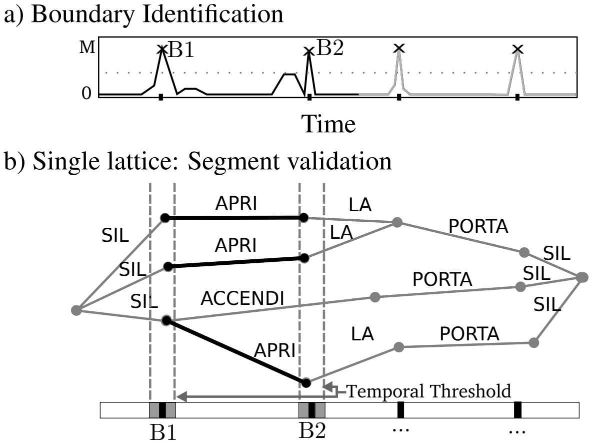
$$C\left([W_{li}, B_i, B_{i+1}]\right) = \frac{1}{M}\sum_j C\left([W_{lij}, B_i, B_{i+1}]\right) \qquad (2)$$

where W_{li} denotes the l^{th} word of the i^{th} confusion set.

The null word is assigned a posterior that is complementary to the sum of the posteriors of the other words hypothesized for the segment under analysis. If a null word dominates a segment, a new search is performed, setting as end-time the next available boundary. Otherwise, the search stops, assigning the estimated posteriors to each of the candidates in the segment. These candidates constitute a new confusion set in the final network.

Note that for each confusion set i, it holds the following relationship (including the null word):

$$\sum_l C\left([W_{li}, B_i, B_{i+1}]\right) = 1 \qquad (3)$$



a) Boundary Identification

b) Single lattice: Segment validation

Fig. 3. Example on a single lattice, segment B1-B2, of the selection of links for segment validation.

4. EXPERIMENTS

We have carried out experiments of recognition of read commands spoken in a domestic environment equipped with a distributed microphone network. The development and test sets were simulated using real impulse responses estimated for a large set of microphone-speaker position/orientation pairs. This data is part of the DIRHA corpora [12]. A single room (i.e., the living room), was selected for the experiments, but it must be noticed that the experimental work can be easily extended to other environments and set-ups. The room is characterized by an average T60 of 0.7 seconds. A subset of 5 microphones was employed, including ones on the wall (L1L, L2R, L3L, L4R), and on the ceiling (LA6) (see Figure 4).

The Hidden Markov Models toolkit (HTK) was used as speech recognition engine. The system adopted the conventional 39-dimensional MFCC feature vectors, including 12 MFCC and log energy, extracted every 10 ms (using a window of 25ms size), and augmented with their first and second derivatives. Cepstral mean normalization was applied. Monophone classes, 27 in total, were modeled with 3 states and 32 Gaussians per state. Contaminated acoustic models [13] were trained on speech simulated at a limited random set of source positions and orientations. The APASCI database[14] was used as training material. This database includes 164 speakers, each uttering 20 phonetically rich sentences. A bigram language model was trained over spontaneous and read-commands, collected also as part of the DIRHA project. The lexicon size of this task is 380 words. The language of the spoken utterances was Italian.

A common parameter explored in speech recognition is the beam pruning threshold, which excludes lower probability hypotheses at an earlier decoding level. A reduction of the beam pruning threshold makes the recognition less complex and faster (very useful in a real-time application) but sometimes introduces errors in the final hypothesis selection. In

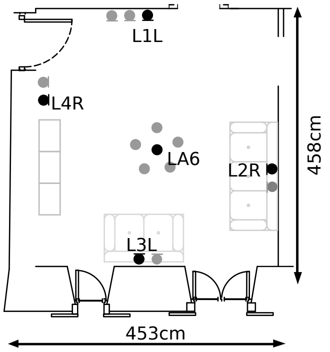


Fig. 4. DIRHA - Living room. Microphones selected for this study appear in the figure as black circles.

order to measure the validity of the technique under different lattices pruning conditions, different beam thresholds (80, 100, 120) were evaluated. Only the results with beam 80(b80) and 100(b100) are reported here, as results with beam 120 were very similar to those of beam 100.

The development set comprises 61 read-command phrases spoken at 43 position-orientations within the livingroom. All sentences of this set were manually segmented and labeled. The test set includes 2245 read-command phrases spoken at 74 position-orientations within the room, which amounts roughly to 1 hour 20 minutes of speech. The selection of source position/orientation was done randomly for each of the phrases spoken by a specific speaker. No overlap between development and test sets holds as for utterances spoken by a speaker at a specific location/orientation.

4.1. Results

In the following, the Multi-Microphone Confusion Network extraction technique proposed in this work is referred to as MMCN. The results of MMCN were compared to those of CNC. Additionally, ROVER results are reported. Table 1 shows the overall performance of the evaluated techniques on the development and test sets. The first five rows of the table report the results with Maximum A Posteriori (MAP) obtained from decoding each channel. For reference, the result extracted from the best channel for each utterance is indicated as ORACLE. Like CNC, ROVER is affected by the order in which the hypothesis are combined. The average(avg), minimum(min), and maximum(max) results are derived from the 120 possible permutations of microphones. CNC obtained a Word Error Rate (WER) ranging from 8.72 to 10.07 in the case of beam 80, and from 7.72 to 8.39 for beam 100. In the same set, it was observed that using manually derived reference word boundaries (MMCN Ref) produced an improvement over CNC for all the beam pruning cases. This procedure was not implemented in the test set, since manual annotations were not available. With the automatic identification of the word boundaries and validation of competing segments (MMCN Auto), MMCN performed better than CNC avg.

Table 1. WER results on Development(Dev) and Test sets

		Dev		Test	
	Mic	b80	b100	b80	b100
MAP	L1L	13.76	11.41	16.99	14.86
	L2R	16.11	11.07	16.21	14.32
	L3L	9.06	6.38	17.18	14.67
	L4R	14.09	12.08	19.43	17.42
	LA6	14.77	12.42	17.60	15.55
ORACLE		1.34	2.35	5.13	4.73
ROVER	avg	8.60	7.53	14.32	12.65
	min	7.72	7.38	14.16	12.46
	max	10.07	8.39	14.50	12.89
CNC	avg	9.25	8.12	13.71	12.17
	min	8.72	7.72	13.66	12.09
	max	10.07	8.39	13.79	12.22
MMCN	Ref	8.39	7.38	-	-
	Auto	9.06	7.72	14.39	13.07

Concerning the results on the test set, both CNC and MMCN outperform the standard MAP. MMCN Auto performed better than MAP over all channels. It is worth noting that although the resulting WER from MMCN is not improved in comparison to CNC, there is a clear advantage in using the proposed method in a multi-microphone setting. CNC presents a weakness since the optimal arrangement of CNs is unknown a priori. This order of CNs varies as a function of the source location and orientation, and the microphone setup. On the contrary, MMCN does not require such an optimization step.

A particular observation was the trend of a microphone to provide lower WER for the MAP decoding (L3L in the development set and L2R in the test set). This is caused by the set of source location/orientation coordinates used to generate the different sets.

5. CONCLUSIONS AND FUTURE WORK

In this paper we introduced a method of multi-lattice agreement for hypothesis combination. From an initial exploration on the development set, we observed that when adequate word boundaries are provided, the proposed method outperforms the state of the art. Next work will concern the development of more efficient algorithms, for the identification of boundaries and the validation of the segments. This work highlighted that there is room for significant improvement for the addressed techniques, as shown by the ORACLE performance. Future investigation will also regard different recognition tasks, the validation on real data, and the impact of additional microphones (available in the DIRHA framework). In a setting with a large number of sensors, applying CNC and ROVER would make the analysis problematic, given the increased number of required microphone permutations. As described before, this does not represent a problem for the proposed MMCN method.

6. REFERENCES

[1] M. Wolfel and J.W. McDonough, *Distant Speech Recognition*, John Wiley & Sons, 2009.

[2] Kenichi Kumatani, John McDonough, Jill Fain Lehman, and Bhiksha Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*. IEEE, 2011, pp. 1–6.

[3] Martin Wolf and Climent Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170–180, 2014.

[4] Gunnar Evermann and P.C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proceedings of Speech Transcription Workshop*. Baltimore, 2000, vol. 27.

[5] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Workshop on Automatic Speech Recognition and Understanding*, dec 1997, pp. 347 –354.

[6] Andreas Stolcke, "Making the most from multiple microphones in meeting recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4992–4995.

[7] Matthias Wölfel, Christian Fügen, Shajith Ikbal, and John W McDonough, "Multi-source far-distance microphone selection and combination for automatic transcription of lectures," in *Proceedings ICSLP*, 2006, pp. 361–364.

[8] Michele Cossalter, Priya Sundararajan, and Ian R Lane, "Ad-hoc meeting transcription on clusters of mobile devices," in *INTERSPEECH*, 2011, pp. 2881–2884.

[9] Andreas Stolcke, "SRILM – An extensible language modeling toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 901–904.

[10] Lidia Mangu, Eric Brill, and Andreas Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.

[11] Daniele Falavigna, Roberto Gretter, and Giuseppe Riccardi, "Acoustic and word lattice based algorithms for confidence scores," in *INTERSPEECH*, 2002.

[12] Luca Cristoforetti, Mirco Ravanelli, Maurizio Omologo, Alessandro Sosi, Martin Hagmueller, and Petros Maragos, "The DIRHA simulated corpus," *LREC*, 2014.

[13] Marco Matassoni, Maurizio Omologo, Diego Giuliani, and Piergiorgio Svaizer, "Hidden Markov Model training with contaminated speech material for distant-talking speech recognition," *Computer Speech & Language*, vol. 16, no. 2, pp. 205–223, 2002.

[14] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus," *Proceedings ICSLP*, September 1994.