EXPLOITING INTER-MICROPHONE AGREEMENT FOR HYPOTHESIS COMBINATION IN DISTANT SPEECH RECOGNITION

Cristina Guerrero and Maurizio Omologo*

Fondazione Bruno Kessler-Irst
Via Sommarive 18, 38123 Trento, Italy

ABSTRACT

A multi-microphone hypothesis combination approach, suitable for the distant-talking scenario, is presented in this paper. The method is based on the inter-microphone agreement of information, extracted at speech recognition level. Particularly, temporal information is exploited to organize the clusters that shape the resulting confusion network, and to reduce the global hypothesis search space. As a result, a single combined confusion network is generated from multiple lattices. The approach offers a novel perspective to solutions based on confusion network combination. The method was evaluated in a simulated domestic environment equipped with largely spaced microphones. The experimental evidence suggests that results, comparable or, in some cases, better than the state of the art, can be achieved under optimal configurations with the proposed method.

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

1. INTRODUCTION

Our voice, as a means to interact with an automatic system, is a powerful instrument. Over the last decade, it has been observed an increasing introduction of Automatic Speech Recognition (ASR), in several services and application fields. Moreover, with the potential of improving the quality of life of physically impaired people, smart voice-operated domestic spaces equipped with sensors and remotely operable devices have been envisioned. As example, a related recent work is the "Distant-Speech Interaction for Robust Home Applications" (DIRHA) Project[1], in which a non-intrusive interaction between a motor impared user and an automated house is explored. For this purpose, the use of multiple distributed microphones in a space is a commonly adopted strategy. This strategy empowers signal and speech processing methods, exploiting the complementarity and redundancy of information. In such a setup, approaches based either on the selection or on the fusion of information are addressed.

Under the first category of approaches, we can mention Channel Selection [1, 2], which aims at limiting the posterior processing to a subset of signals or components. For the second category, efficient fusion methods at signal, feature, model, or hypothesis level are under study. Feature and model combination techniques, generally subject to restrictive assumptions, require an additional step for the integration of the multiple channel processing outputs. When a single output is the target, signal fusion methods are perhaps the most popular lines of fusion research. Beamforming is an example of these methods, in which a specific geometrical configuration of a microphone array allows spatial filtering of sounds based on the relative locations of the sources. Under loosely specified distant microphone configurations, its application is not advisable, since spatial aliasing and other artifacts would strongly affect the resulting fusion. Hypothesis combination, though higher in complexity than the previously described approach, supports the exploitation of the information captured by different sensors, without being limited to specific characteristics of a microphone network.

In the context of distant speech recognition, a well-known hypothesis combination method is Confusion Network Combination [3]. The method mixes individual Confusion Networks (CNs) [4] into a single clustered structure. The CNs are extracted from the lattices produced by decoding different microphone signals. This framework was motivated by Minimum Bayes Risk Decoding, which addresses the mismatch between the Bayes decision rule and the evaluation criterion in standard statistical speech recognition. As a result, a better recognition performance is achieved.

In this work, a domestic environment, with largely spaced and distributed microphones, is considered as research ground. In this context, other works are reported in the literature, as [5] which compared beamforming and other fusion techniques. Our paper extends a preliminary investigation [6] conducted with a reduced number of microphones. The goal is to provide a different approach for hypothesis combination, more suited for multi-microphone settings, and comparable to the state of the art. The basis of the proposed method of multi-microphone CN extraction, referred in the following as MMCN, is the agreement of temporal features among the lattices to be combined.

[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 (see http://dirha.fbk.eu).

* The author is a PhD student at the ICT School - University of Trento

The remainder of this paper is organized as follows. Section 2 introduces the standard hypothesis space combination approach. In Section 3, MMCN method is described. Section 4 presents the experimental setup. The discussion of the results and findings are detailed in Section 4. The conclusions and future work are reported in Section 5.

2. MINIMUM BAYES RISK IN HYPOTHESIS SPACE COMBINATION

The goal of standard ASR decoders is to identify the sequence (W) of words or symbols which maximizes the sequence posterior probability ($P(W|X)$), where X corresponds to the acoustic observation sequence. This is enclosed under the Maximum A Posteriori rule:

$$W^* = \arg\max_W P(W|X) \quad (1)$$

$$= \arg\max_W P(W)P(X|W) \quad (2)$$

where $P(W)$ is the score given by the language model, and $P(X|W)$ is the acoustic model probability. The recognition output, however, is evaluated on a word level basis. A mismatch holds between the decoding and the evaluation, which motivated the study of approaches addressing this issue. Minimum Bayes Risk (MBR) Decoding [7] comprises the diverse work explored in this respect. Its target is the extraction of a W that minimizes the expected Word Error Rate.

Consensus Decoding [4] was proposed as a MBR decoding approach. The method takes, as input, the lattice generated by a decoder, and produces a very compact clustered version of it, denominated a confusion network (CN). The extracted network is the result of a series of iterative merging steps. In the CN, the final hypothesis is extracted by a simple selection of the unit (a word or a silence) with the highest posterior probability at each cluster. This technique has been compared to the NIST Recognizer Output Voting Error Reduction (ROVER) [8] system. CN provides a more comprehensive result, since the combination is performed on the hypothesis space level and not on individual hypotheses.

Soon afterwards, the combination of these compact networks was explored for multi-decoder hypothesis combination. The approach known as Confusion Network Combination (CNC), initially explored on a single signal, has been used in recent years to study its performance in a multi-microphone setting [9, 10]. These works explored Channel Selection and CN weighting complementary techniques, as an attempt to enhance the performance of CNC in such a scenario. Nevertheless, they achieved discouraging results versus signal fusion techniques.

3. BUILDING A CONFUSION NETWORK OVER MULTIPLE LATTICES

The focus of this work is a reconsideration of the CN extraction from lattices generated from multiple microphone signals. Instead of relying on individually compacted versions

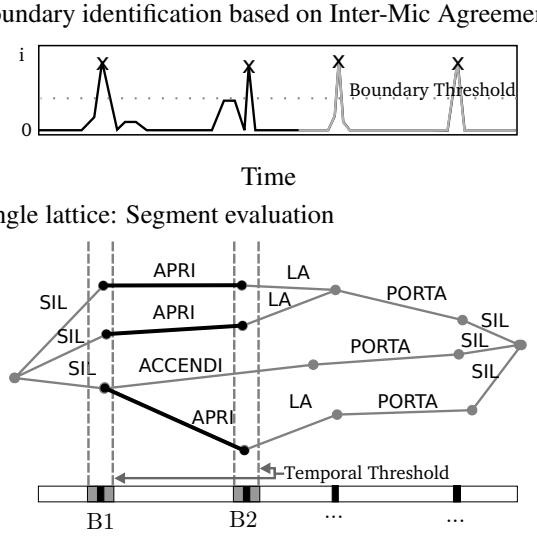

Fig. 1. An example, for a single lattice, and a single segment (B1-B2), of the selection of links in the boundary-based segment validation procedure.

(CNs) of the lattices, a method for directly compacting the multiple lattices is explored here. In our recent work [6], a multi-microphone agreement method was conceived to establish the key elements of the final confusion network.

The structure of a confusion network is given by a sequence of a limited number of clusters or confusion sets. A confusion set is characterized by a number of candidate words (or the silence unit), each one with its posterior probability. The sequentiality of the sets defines the possible sequence of words in the resulting hypothesis space. One can say that each of the confusion sets represents a segment in time, and that the number of candidates in a set represents the degree of confusion in the corresponding hypothesis sub-space.

Both, ordering and temporal information are encoded in the lattices. The investigated method, MMCN, assumes that there is an agreement, among the different lattices, of the temporal information associated to relevant words. It must be noticed that this technique operates on time synchronized lattices.

Additionally, acoustic and linguistic information encoded in the lattices is unified in the word posterior probability, computed over each link l in the lattice. The posterior is the ratio of the probability of all complete paths that pass through the link l over the probability of all complete paths in the lattice Λ, i.e.,

$$P(l|\Lambda) = \frac{\sum_{C\in\Lambda, l\supset C} P(C|\Lambda)}{\sum_{C\in\Lambda} P(C|\Lambda)} \quad (3)$$

where $C \in \Lambda$ means that C is a complete path in the lattice Λ, and $l \supset C$ indicates that the complete path C passes through the link l. The word posterior probability can be computed using a forward-backward algorithm.

The technique exploits the information within the lattices

to derive features, such as word time boundaries, for extracting a coherent hypothesis space represented as a CN. The coherence of the information is used both for identifying potential segments that will form the resulting CN, and for validating these segments and its candidate words.

3.1. Inter-Microphone Boundary Agreement

In the standard CN extraction, links with a very low posterior probability are discarded at an initial stage to avoid detrimental effects on the final alignment. We keep the entire structure of the lattice for the extraction of the final CN, but at the initial step of the process we also follow a similar pruning strategy. With this procedure, we expect coherent temporal information of the relevant words is kept. A cumulative sequence of the link boundaries of all microphones is computed, and then the predominant boundaries are selected. These boundaries are processed in the next stage.

3.2. Intra/Inter Microphone Score Estimation

The previously estimated boundaries mark segments of interest, to be accepted or discarded based on the analysis of the word candidates and their global posterior scores. The starting point of this step concerns the selection of l word candidates that are present within the segment under analysis. A certain tolerance threshold Δ around the boundaries is allowed. Then, for each pair of boundaries (B_i, B_{i+1}), the posterior probabilities for the final network are estimated in two steps. First, for each microphone j, an intra-microphone computation of the posterior scores C is achieved through an accumulation of posteriors, on a per word basis:

$$C\left([W_{lij}, B_i, B_{i+1}]\right) = \sum_{\substack{[w;\tau,t]: \\ [B_i-\Delta\leq\tau\leq B_i+\Delta], \\ [B_{i+1}-\Delta\leq t\leq B_{i+1}+\Delta]}} P\left([W_{lij}, \tau, t]\,|x_1^T(j)\right) \quad (4)$$

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. Then, the global posterior is computed as the average of the estimated probabilities for all the microphones.

A posterior probability is estimated also for the Null or Silence, as the complement of the sum of all candidates posterior probabilities. The rejection or acceptance of a segment is subject to the dominance of the Null, determined by a threshold. If a segment is rejected, the search is expanded to a new segment whose starting boundary is the same as the rejected one, and its end boundary is the next available boundary.

Figure 1 depicts an example of the selection of the links within a segment under analysis (the segment starting at B1 and ending at B2). The selected links are shown in bold. For each of the words in these links, posterior probabilities are then going to be estimated from the scores present in the multiple lattices.

As highlighted in [6], no alignment is involved in MMCN. For this reason, the application of this method is independent of the order in which lattices are combined. This property differentiates MMCN from CNC and ROVER, and is particularly important when a large number of sensors is available.

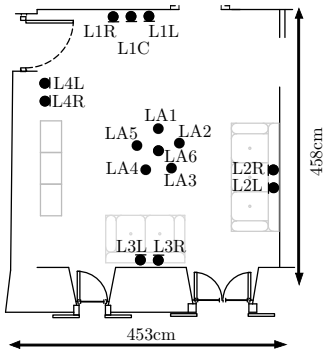

Fig. 2. DIRHA - Living-room. The microphones used in the experiments appear in the figure as black circles.

4. EXPERIMENTAL SETUP

The evaluation task is distant speech recognition in a multi-microphone setting. The impact of varying the number of microphones in the hypothesis combination was measured. The hypothesis from the resulting network was evaluated using the standard error metrics for ASR, the word error rate (WER).

4.1. Corpora

The language of the data used in the experiments was Italian. Datasets were simulated using real impulse responses estimated for a large set of microphone-speaker position pairs. This data is part of the DIRHA corpus, whose details can be found in [11] or on the referred project website. The selection of source position/orientation was performed randomly. No overlap between development and test sets holds as for utterances spoken by a speaker at a certain location. The development set (devset) was composed of 61 phrases, spoken by 27 speakers at 43 position/orientations. In the test set there were 2245 phrases in total, spoken by 30 speakers, at 74 position/orientations. In both sets the positions of the speakers were located within a single room, i.e., the living-room. The room was characterized by an average T60 of 0.7 seconds. All the phrases were of the type "read commands". A total of 15 synchronized microphones were available in the room under study; 6 on the ceiling (LAx) and 9 on the walls (L1x, L2x, L3x, L4x) (see Figure 2).

4.2. Speech Recognition

The speech recognition system used in this work was built on the HTK toolkit [12]. A standard front-end processing was employed, with a pre-emphasis step and a feature vector composed of 12 Mel Frequency Cepstral Coefficients plus the energy, and their first and second derivatives. Mean and energy normalization were applied. Various pruning conditions were explored at experimental level. Here we contrast

the results of applying beam-search with a beam of 100 and no beam, a more computational demanding setting.

A set of 27 phonemes was selected. Acoustic context-independent phone units, modeled with three states and 32 Gaussian mixtures per state, were trained. APASCI database [13] was contaminated [14], and this version was used to train the acoustic models. This database includes 20 phonetically rich sentences spoken by 164 speakers.

The language model was a bigram, trained on a mixture of read and spontaneous commands, collected under the DIRHA project. The size of the dictionary was of 380 words. Language model scale(s) and word insertion penalty(p) were optimized in the devset, using only one microphone per each microphone group (e.g., mic LA6 in the group LAx). A single combination of parameters (s=11, p=16) was then used, for all the microphones in the test set.

4.3. Combination Techniques

MMCN is compared to signal and hypothesis combination techniques. Beamformed (BF) signals were extracted using the BeamformIt tool [15]. The results of word level hypothesis combination (ROVER) were derived with SCTK Toolkit [16]. In order to apply CNC in the given experimental conditions, the SRILM toolkit [17] was used. With the lattice-tool we obtained CNs from the lattices derived by decoding each channel, and later CNC was performed using the nbest-lattice tool. All the CNs were assigned a uniform weight.

5. RESULTS

Table 1 shows the performance of MMCN and other combination techniques, on the dev and test sets. The results of applying a beam of 100 (b100) and no pruning (None) are presented. Narrowing this beam would lead to significantly worse performance. The results of decoding every single distant microphone (SDM) are also presented for reference purposes. ORACLE displays the results of selecting the channel with the lowest WER per utterance. For simplicity, three mic-configurations are presented in the tables, which are denoted as C_5, C_{10} and C_{15}. The mics composing the configurations are: C_5) L1L-L2R-L3L-L4R-LA6, C_{10}) L1L-L1C-L2L-L2R-L3L-L3R-L4L-L4R-LA3-LA6, C_{15}) L1L-L1C-L1R-L2L-L2R-L3L-L3R-L4L-L4R-LA1-LA2-LA3-LA4-LA5-LA6. Concerning ROVER and CNC, the order of the elements used in the combination affects the final hypothesis. According to the number of mics to combine, there would be a large number of possible permutations to explore (e.g., given 5 mics, there are 120 possible permutations). Due to this reason, in this work, given N mics, only N permutations were addressed with ROVER and CNC. Each permutation is created in a cyclic fashion, starting at element k, with $k = 1..N$. The average performance of these permutations is reported in the Table.

Experimental results show that hypothesis combination approaches achieve better performance than SDM and BF, as expected. Furthermore, with the optimal set of parameters, MMCN achieves a performance comparable to CNC.

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

		Dev		Test	
	Mic	b100	None	b100	None
SDM	L1L	11.41	11.41	14.86	14.33
	L1C	17.11	16.11	15.07	14.55
	L1R	16.11	16.11	14.62	14.19
	L2L	9.40	9.4	14.66	13.84
	L2R	11.07	11.07	14.32	13.57
	L3L	6.38	6.38	14.68	14.05
	L3R	10.74	10.07	14.58	14.12
	L4L	16.78	16.44	16.52	15.88
	L4R	12.08	12.08	17.42	16.63
	LA1	17.45	17.45	16.22	15.62
	LA2	15.44	15.44	16.34	15.46
	LA3	19.46	18.46	15.14	14.49
	LA4	17.45	17.45	16.24	15.73
	LA5	14.77	13.76	15.70	15.05
	LA6	12.42	12.42	15.55	14.96
ORACLE		2.35	2.35	4.73	4.51
BF	C_5	10.74	10.74	16.44	15.38
	C_{10}	10.07	10.07	15.21	14.23
	C_{15}	10.40	10.40	15.47	14.07
ROVER	C_5	7.38	7.38	12.69	12.20
	C_{10}	8.66	8.69	12.21	11.76
	C_{15}	9.08	8.99	12.38	11.93
CNC	C_5	8.05	8.05	12.18	11.82
	C_{10}	8.92	9.26	11.87	11.50
	C_{15}	9.37	9.44	11.99	11.57
MMCN	C_5	7.72	7.72	12.76	12.22
	C_{10}	9.73	9.40	12.42	12.20
	C_{15}	9.40	9.40	12.60	12.26

In order to evidence the effect of the order in which the mics are incorporated into a combination, we explored other configurations, such as one based on ranking mic-group WERs, i.e.: X_6) the mics on the ceiling (6 mics.), then X_{10}) those in X_6 plus one mic of each wall group (10 mics.), and finally X_{15}) adding the remaining wall sensors (15 mics.). Note that in this case only one permutation is analyzed for each configuration. With CNC, in the case of X_{15}, its performance is different from the one reported in C_{15} Table 1. Starting with lattices that achieved the highest SDM WER, the balanced inclusion of mics leads to a reduction of the WER. In Figure 3, it can be observed that, for most of the explored mic-configurations, now MMCN shows a lower WER than CNC. This illustrates the impact of the arrangement of mics on CNC, which is not an issue for MMCN.

From the different setups evaluated, it can be observed that the number of mics is not the unique factor affecting hypothesis combination approaches; the quality of the source lattices, an information not available a priori, is also relevant.

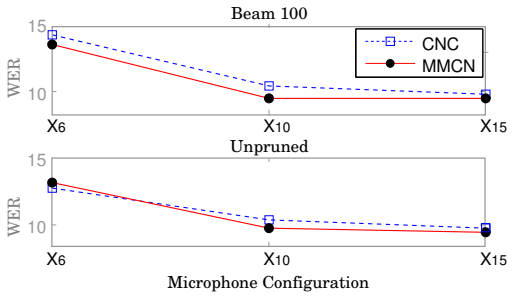

Fig. 3. WER variation with different mic-configurations.

Although there is no WER reduction with MMCN, its performance is invariable independently of the arrangement of the mics. This is an advantage over CNC in a multi-microphone scenario where an analysis of a proper ordering is hard to achieve. Note that the algorithms in MMCN are just at starting level, which leaves a window for improvement concerning the identification of optimal parameters.

6. CONCLUSIONS AND FUTURE WORK

A new method of multi-lattice agreement for hypothesis combination was presented and evaluated. Given the large number of microphones available in the research scenario, it became evident the advantage of applying a method which obtains consistent results independently of the arrangement of the sensors. With the optimal operation parameters, MMCN achieves comparable results to those of the state of the art. With the support of more efficient algorithms for the extraction of key components in the MMCN approach, experiments suggest that an improvement could be achieved. The technique has already been validated on real data, confirming the performance outlined in this work.

REFERENCES

[1] K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Hands-free Speech Communication and Microphone Arrays, Joint Workshop*. IEEE, 2011, pp. 1–6.

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

[3] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*. Baltimore, 2000, vol. 27.

[4] L. Mangu, E. Brill, and A. 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.

[5] B. Lecouteux, M. Vacher, F. Portet, et al., "Distant speech recognition in a smart home: Comparison of several multisource ASRs in realistic conditions," *Proc. of International Conference on Spoken Language Processing*, pp. 2273–2276, 2011.

[6] C. Guerrero and M. Omologo, "Word boundary agreement to combine multi-microphone hypothesis in distant speech recognition," in *Hands-free Speech Communication and Microphone Arrays, Joint Workshop*. IEEE, 2014.

[7] V. Goel and W. J. Byrne, "Minimum bayes-risk automatic speech recognition," *Computer Speech & Language*, vol. 14, no. 2, pp. 115–135, 2000.

[8] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Workshop on ASR and Understanding*. IEEE, dec 1997, pp. 347–354.

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

[10] 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 International Conference on Spoken Language Processing*, 2006, pp. 361–364.

[11] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, may 2014.

[12] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, vol. 2, Entropic Cambridge Research Laboratory Cambridge, 1997.

[13] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus," *Proc. of International Conference on Spoken Language*, pp. 1391–1394, 1994.

[14] M. Matassoni, M. Omologo, D. Giuliani, and P. 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.

[15] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2011–2022, 2007.

[16] "NIST-SCTK Speech Recognition Scoring Toolkit," http://www.nist.gov/speech/tools/, 2009.

[17] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. of International Conference on Spoken Language Processing*, 2002, pp. 901–904.