Multi-Source Far-Distance Microphone Selection and Combination for Automatic Transcription of Lectures

Matthias Wölfel, Christian Fügen, Shajith Ikbal, and John W. McDonough

Institut für Theoretische Informatik, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
{wolfel|fuegen|shajith|jmcd}@ira.uka.de

Abstract

In this work, we present our progress in multi-source far field automatic speech-to-text transcription for lecture speech. In particular, we show how the best of several far field channels can be selected based on a signal-to-noise ratio criterion, and how the signals from multiple channels can be combined at either the waveform level using blind channel combination or at the hypothesis level using confusion network techniques to improve the accuracy of a far field lecture transcription system. Using the techniques described here, we ran a series of experiments on the test set used by the US National Institute of Standards and Technologies for the RT-05S evaluation. For the multiple distant microphones (MDM) task of RT-05S, our system achieved a word error rate of 38.5% which represents an improvement of over 13% absolute compared to the best reported results in the RT-05S evaluation.
Index Terms: far-distance, automatic speech recognition

1. Introduction

Ideally, automatic speech recognition systems working on lecture or meeting tasks operate on data recorded from distant microphones, freeing users from wearing body-mounted microphones. If applied wisely a combination of microphones can improve the performance over a single one. Therefore, an important effort in current speech research is focused on the processing of speech from multi-source far field microphones. This problem is surpassingly difficult, given that the speech signals collected by a varying number and types of microphones are severely degraded by both, background noise and reverberation and that their locations are unknown. As in many real-world applications, microphone arrays with known and fixed geometries are not be available, we focus in this work on speech recorded by several randomly placed table-top microphones. Furthermore, the chosen material is challenging on other aspects: lecture speech varies in speaking style from freely presented to read, comprising spontaneous events as well as hyper articulation [1]. The evaluated corpus contains mainly non-native speakers of English, some not even fluent.

Section 2 describes the development of a baseline system at the Universität Karlsruhe (TH). Section 3 covers in detail different multi-source selection and combination techniques. Section 3 presents and discusses a variety of speech recognition experiments and section 5 concludes the findings.

2. Task Description and Baseline System

The evaluated NISTs RT-05S lecture meeting data [2], selected under the European Commission integrated project CHIL [3], *Computers in the Human Interaction Loop*, which aims to make significant advances in the fields of speaker localization and tracking, speech activity detection and distant-talking ASR, presents significant challenges to both modeling components used in *automatic speech recognition* (ASR), namely the language and acoustic models. With respect to the former, the currently available lecture data primarily concentrates on technical topics with focus on speech and vision research. This is a very specialized task that contains many acronyms and therefore is quite mismatched to typical language models currently used in the ASR literature. Furthermore, on the acoustic modeling side, large portions of the data contain spontaneous, disfluent, and interrupted speech, due to the interactive nature of seminars and the varying degree of the speakers' comfort with their topics. In addition to the latter difficulty, the seminar speakers exhibit moderate to heavy German or other European accents in their English speech.

Three evaluation conditions using different type and number of microphones were defined for RT-05S lecture data:

- **MDM** Multiple Distant Microphones
- **SDM** Single Distant Microphone
- **IHM** Individual Head-set Microphone

The SDM condition can be derived from the MDM condition by disregarding all but one centrally-located channel for each meeting, which was specified by NIST in the task description. Using this channel, however, did not necessarily result in the lowest possible single-channel word error rate. A description of the system used for the IHM condition is not given here, but can be found in Fügen *et al* [4].

The above problems are compounded by the fact that not enough data is available for training new language and acoustic models matched to this lecture task, and thus one has to rely on adapting existing models that exhibit gross mismatch to the data. Clearly, these challenges present themselves in both close-talking microphone data, as well as far-field data captured using table-top microphones, where of course they are exacerbated by the much poorer quality of the acoustic signal.

A detailed description of the room layout, data collection, labeling as well as some preliminary experiments on the far field microphones can be found in [5].

2.1. Vocabulary Selection and Language Model Training

The dictionary contains 58,695 pronunciation variants over a vocabulary of 51,731. We used a 4-gram language model, which achieve a perplexity of 130 on the NIST RT-05S lecture meeting task. More details can be found in the system description [4].

2.2. Acoustic Pre-Processing

For the extraction of speech features we have used two different front-ends. One is identical to the one used in the RT-04S meeting evaluation [6] based on *Mel-frequency cepstral coefficients* (MFCC). The second uses a warped *minimum variance distortionless response* (MVDR) spectral envelope [7] of model order 30.

Both front-ends provided features every 10 ms (first pass) or 8 ms (following passes) obtained by the Fourier transformation followed by a Mel-filerbank or the warped MVDR. No filterbank was used in the warped MVDR case as the warped MVDR envelope already provides the properties of a Mel-filterbank, namely smoothing and Mel-frequency warping of the spectral estimate. Vocal track length normalization was applied either in the linear domain, MFCC, or in the warped frequency domain. The MFCC models used 13 cepstral coefficients while for the MVDR models the number of cepstral coefficients has been increased to 20. Thereafter, the mean and variance of the cepstral coefficients were normalized, and seven adjacent frames were appended, resulting in a feature of either 195 Mel-frequency or 300 warped MVDR cepstral coefficients. These features were then reduced to a final length of 42 by applying *linear discriminant analysis* (LDA) and a global semi-tied covariance (STC) transform.

2.3. Acoustic Model Training

The speech recognition experiments described below were conducted with the *Janus Recognition Toolkit* (JRTk), which was developed and is maintained jointly by the Interactive Systems Laboratories at the Universität Karlsruhe (TH), Germany and at the Carnegie Mellon University in Pittsburgh, USA.

Both, the MFCC and warped MVDR systems were trained in the same way, resulting in a size of 16,000 distributions over 4,000 models, with a maximum of 64 Gaussians per model. For faster turn around times we have optimized our system on close talking and later adapted the close talking models to far field by four, MFCC, or two, warped MVDR, additional Viterbi iterations on the close talking models using far field data. Details of the training procedure can be found in [4].

3. Multi-Source Selection and Combination

To develop the reliable and computationally efficient combination strategy presented here, we briefly review signal and text combination techniques, and discuss their advantages and disadvantages. We then explain how these strategies can be augmented by a *signal to noise ratio* (SNR) criterion applied on an utterance and speaker basis to select the utterances or channels per speaker to be combined. We show that a selection based on SNR can improve recognition speed as well as accuracy. Last but not least we give a signal combination in conjunction with text combination strategy which can increase the accuracy over either of the techniques with a limited amount of extra computation compared to the signal combination and a huge reduction of computation time compared to text combination techniques.

3.1. Signal Combination: Blind Channel Combination

To perform automatic transcription of seminar recordings made with far field microphones, it is necessary to separate the voice of a single speaker from the background noise to improve the signal quality, and therefore the recognition accuracy. There exists a large literature on various *blind source separation* techniques [8] that claim to be able to solve this problem. To the knowledge of the current authors, however, such techniques have never been successfully applied to the problem of automatic transcription of the type of speech encountered in the NIST RT evalaution. A simpler approach suggests itself, however: assuming the speech on all microphones is correlated while at least some of the noise is uncorrelated, we can simply sum up all channels pre-shifted by their relative estimated *time delays of arrival* (TDOA) and divide by the number of channels N to attenuate the noise. To estimate the TDOA, we can maximize

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} G_{x_1,x_2}(\omega)\, e^{j\omega\tau}\, d\omega \tag{1}$$

with the *generalized cross correlation* (GCC)

$$G_{x_1,x_2}(\omega) = \frac{X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})|} \tag{2}$$

To improve the estimate of the TDOA under realistic conditions where correlated noise is present we have subtracted the cross-correlation of the averaged noise where N_1 and N_2 is estimated at the time no speech is present [9]:

$$G(\omega) = G_{x_1,x_2}(\omega) - \frac{N_1(e^{j\omega\tau}) N_2^*(e^{j\omega\tau})}{|N_1(e^{j\omega\tau}) N_2^*(e^{j\omega\tau})|} \tag{3}$$

3.2. Signal Combination: SNR Based Channel Combination

As not all channels have similar quality for different utterances or speakers it might not be sufficient to weight the channels equally. In the literature it was suggested to weight the channels by their SNR values. Even though this approach led to a small improvement in SNR per speaker and recognition accuracy we found that a combination of only the best channels led to better results in SNR as well as accuracy. To decide which channels should be used for combination, we compared SNR values over all channels on an utterance basis and divided by the best SNR value. This ratio of SNRs was then compared to a threshold and the corresponding channel was chosen if the value was above the threshold. On our data we found that a threshold of 0.95 led to the best result in SNR, an improvement of 2 dB. Hence, our selection rule was, choose channel X if

$$\frac{\mathrm{SNR_X}}{\max_{\mathrm{channel}} (\mathrm{SNR_{channel}})} > 0.95$$

Thereafter, this combined channel was compared to all single channels on a speaker basis and the channel with the best SNR value was chosen as the final channel. These steps are depicted in the gray boxes numbered one, two and three of Figure 1.

3.3. Text Combination: Confusion Network Combination

Confusion networks reduce the complexity of lattice representations to a simpler form that maintains all possible paths from the lattice, but transforms the space to a series of slots which each have word hypotheses (and null arcs) and associated posterior probabilities. Therefore, by combining the hypotheses or lattices of the same time segment of recognition runs on different microphones into a single word confusion network the networks can be used to optimize the *word error rate* (WER) over different microphones by selecting the word with the highest probability in each particular slot [10].

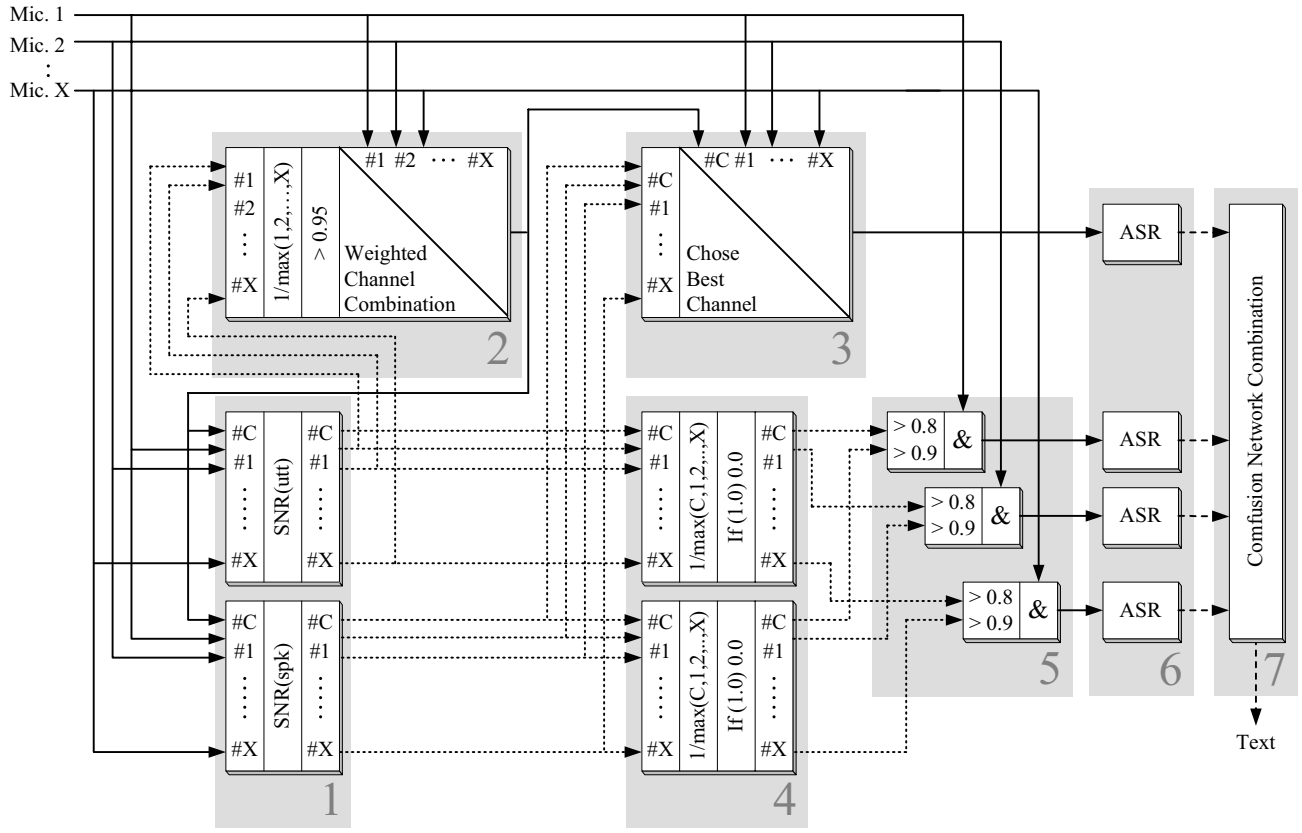

Figure 1: Flowchart of channel combination. Solid lines show the flow of the audio files while dotted lines are float values (SNR or values between zero and one). The dashed lines stand for word hypos.

3.4. Text Combination: SNR Based Confusion Network Combination

The mayor drawback of confusion network combination is a very high computation load as an individual adaptation and recognition is required for every available channel. Hence, the overall computation grows linearly with the number of available channels. In this section, we aim on reducing the amount of computation time by selecting channels and utterances which might lead to an additional improvement in word accuracy over the best single channel (note that the best single channel could be a combination of more than one physical channel). To decide which channels and utterances should be used we once more compare the normalized SNR values. Only those utterances and channels are decoded which have a value above 0.9 for the speaker and 0.8 for the utterance. These steps are depicted in the gray boxes numbered one, four and five of Figure 1. This strategy yielded an overall reduction in additional decoding effort of 70% as compared to decoding over all channels with no increase in word error rate.

3.5. Dual Combination

Both, the *blind channel combination* (BCC) and the *confusion network combination* (CNC), approaches have their advantages and disadvantages and on a variety of experiments no clear decision can be taken which of the approaches is leading to the best possible performance. To profit from both techniques the combined channel has to be added as an additional channel to the confusion network combination approach.

4. Speech Recognition Experiments

For the preliminary experiments scored in WER without overlap, Table 1, we have used an unadapted one pass system with the MVDR front-end and slightly worse language and acoustic models as in later experiments. For the BCC system we see nice gains using a threshold; e.g., we gain 3.8% by a threshold of 0.9. For the CNC system we can observe a small improvement in word accuracy, 0.4% for a threshold of 0.9, by decoding only on an limited numbers of additional utterances and channels if compared to the decoding over all utterances and channels. Combining the two approaches are leading to the best numbers, again we profit from deleting bad utterances and channels chosen by a threshold, in speed as well as in accuracy. Note that in this case the lower threshold of 0.8 is leading to a better result.

For the following experiments three passes of decoding are processed on automatic segmentation. Automatic segmentation and clustering steps for MDM and SDM are the same, except for the fact that in MDM a best single distant microphone (BSDM) is chosen by the SNR measure. The resulting segments from segmentation are further tagged with speaker labels using a hierarchical agglomerative speaker clustering technique as explained in [11].

The first decoding used speaker based incremental VTLN estimation and incremental FSA. The following decodings were

Combination	WER		
Cutoff	0.0	0.8	0.9
BCC	61.0%	57.7%	57.2%
CNC	58.5%	58.3%	58.1%
BCC & CNC	57.5%	56.6%	56.9%

Table 1: *Word error rates* (WER)s for different channel combination techniques and cutoffs.

Spectral Estimation	WER			
Pass	1*	1	2	3
Fourier	53.2%	49.3%	44.0%	41.8%
warped MVDR	52.3%	49.3%	43.0%	40.2%
CNC	50.9%	46.9%	42.0%	39.0%
+ all channels		-	-	38.7%
+ selected channels		-	-	38.5%

Table 2: *Word error rates* (WER)s for channel combination. Pass 1* has used an SNR based weighting, all other passes have used the proposed selection of channels.

adapted on either confidence-weighted hypothesis of the MVDR system, for the MFCC front-end or on confusion network combined confidence-weighted hypothesis for the MVDR front-end with *maximum likelihood linear regression* (MLLR), VTLN and FSA.

The first and second pass in Table 2 uses only the SNR based BCC approach. On the final pass the combined SNR based BCC and SNR based CNC approach is leading to an improvement of 0.5% over the SNR based BCC approach. By comparing the single channel WER of Table 3 with the mutiple channel WER of Table 2 we see a significant gain by using multiple far field microphones wisely over a single far field microphone. If we apply simple blind channel combination we don't see gains over the single channel. This is consistent to the numbers published by ICSI [12] last year.

On preliminary results of the RT-06S lecture evaluation data the proposed system shows a similar improvement between the MDM and SDM condition, which is significantly higher than the improvements reported by other sides.

5. Conclusions

The paper has presented our progress in multi-source far field automatic speech transcription. Based on the SNR selection we were able to improve in speed as well as in accuracy over our previous system. In addition to this we have successfully combined the two different combination approaches with further improvements over the single approaches with limited additional decoding time.

6. Acknowledgment

The work presented here was partly funded by the *European Union* under the project CHIL, *Computers in the Human Interaction Loop*, contract number IST-506909.

The authors would like to thank the additional members of the RT-06S evaluation team Kenichi Kumatani, Florian Kraft, Kornel Laskowski and Sebastian Stüker.

Spectral Estimation	WER		
Pass	1	2	3
Fourier	52.9%	48.2%	46.4%
warped MVDR	52.6%	47.4%	44.8%
CNC	50.9%	45.9%	43.4%

Table 3: *Word error rates* (WER)s for single channel

7. References

[1] M.C. Wölfel and S. Burger, "The ISL baseline lecture transcription system for the TED corpus," *Technical Report TR0001, http://isl.ira.uka.de/~wolfel*, 2005.

[2] NIST, "Rich transcription 2005 spring meeting recognition evaluation," *www.nist.gov/speech/tests/rt/rt2005/spring*.

[3] "Computers in the human interaction loop," *http://chil.server.de*.

[4] C. Fügen, Wölfel. M., J. McDonough, S. Ikbal, Kraft; F., K. Laskowski, M. Ostendorf, S. Stüker, and K. Kumatani, "Advances in lecture recognition : The ISL RT-06S evaluation system," *Proc. of Interspeech*, 2006.

[5] M.C. Wölfel and J.W. McDonough, "Combining multi-source far distance speech recognition strategies: Beamforming, blind channel and confusion network combination," *Proc. of Interspeech*, 2005.

[6] F. Metze, C. Fügen, Y. Pan, T Schultz, and H. Yu, "The ISL rt-04s meeting transcription system," *Proc. of ICASSP Meeting Recognition Workshop*, 2004.

[7] M.C. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[8] X.R. Cao and R.W. Liu, "General approach to blind source separation," *IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 44, no. 4, pp. 562 - 571*, 1996.

[9] Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," *Proc. of ICASSP*, 2004.

[10] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer, Speech and Lanuage, vol. 14, no. 4*, vol. 14, no. 4, pp. 373–400, 2000.

[11] Q. Jin and T. Schultz, "Speaker segmentation and clustering in meetings," *Proc. of ICSLP*, 2004.

[12] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grezl, A. Janin, A. Mandal, C. Peskin, B. Wooters, and J. Zheng, "Further progress in meting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," *Proc. of the Rich Transcription 2005 Spring Meeting Recognition Evaluation*, 2005.