Combining Multi-Source Far Distance Speech Recognition Strategies: Beamforming, Blind Channel and Confusion Network Combination

Matthias Wölfel and John McDonough

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

10.21437/Interspeech.2005-270

Abstract

Interest within the automatic speech recognition (ASR) research community has recently focused on the recognition of speech captured with a microphone located in the medium field, rather than being mounted on a headset and positioned next to the speaker's mouth. The capacity to recognize such speech is a primary requirement in making ASR a viable modality for so-called *ubiquitous computing*. This is a natural application for multiple microphones whose signals can be combined in different ways: On the signal side, combination can be accomplished by beamforming techniques using a microphone array or by blind source separation. On the word hypothesis side, combination can be achieved through confusion network combination. In this work, we compare the effectiveness of the several combination techniques, and compare their performance to that achieved with a close talking microphone.

1. Introduction

Interest within the *automatic speech recognition* (ASR) research community has recently focused on the recognition of speech captured with a microphone located in the medium field, rather than being mounted on a headset and positioned next to the speaker's mouth. Using a combination of microphones can improve the performance with respect to a single microphone. To combine the multiple sources we can identify two main approaches: on the signal side, through beamforming techniques using a *microphone array* (MA) or *blind channel combination* (BCC); or on the word hypothesis side through *confusion network combination* (CNC).

In this work, we present a variety of ASR results using different types of microphones and their combinations with the aforementioned techniques. The speech corpus used for the experiments reported here was collected as part of the European Commission integrated project CHIL [1], *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. The corpus is comprised of lectures and oral presentations collected by both near and far-field microphones. In addition to the audio sensors, the seminars were also recorded by calibrated video cameras. This simultaneous audio-visual data capture enables the realistic evaluation of component technologies as was never possible with earlier data bases. One of the long-term goals of the project is to develop the ability to recognize speech in a real reverberant environment, without any constraint on the number or the distribution of microphones in the space nor on the number of sound sources active simultaneously. This problem is surpassingly difficult, given that the speech signals collected by a given set of microphones are severely degraded by both background noise and reverberation. Moreover, the speech material is inherently challenging for several reasons: Lecture speech varies widely in speaking style as compared to read speech and contains spontaneous events as well as hyper-articulation effects [2]. Moreoever, the corpus contains mainly non-native speakers of English, some of whom are not even fluent in English.

The remainder of this works is organized as follows. Section 2 describes the development of a baseline system at the Universität Karlsruhe (TH). Section 3 gives a short description of the data collection and labeling. Finally, Section 4 reports the results, conclusions and plans for future work of the speech recognition experiments.

2. Baseline System

The CHIL seminar data present significant challenges to both modeling components used in ASR, namely the language and acoustic models. With respect to the former, the currently available CHIL data primarily concentrates on technical topics with a focus on ASR research. The speech material is very specialized with many technical terms and acronyms; hence, the language modeling corpora typically used in the ASR literature are ill-suited to this particular ASR task. Due to the interactive nature of the seminars and the varying degree of the speakers' comfort with their topics, large portions of the data are characterized by spontaneous, disfluent, and interrupted speech. Moreover, the seminar speakers exhibit moderate to heavy German or other European accents in their English speech. These problems are compounded by the fact that, at this early stage of the CHIL project, not enough data is available for training new language and acoustic models for the seminar task. Thus one has to rely on adapting existing models that exhibit gross mismatch to the CHIL data. Clearly, these challenges present themselves in both close-talking microphone data, as well as the far-field data captured using the MAs and table-top microphones, where of course they are exacerbated by the poorer quality of the acoustic signal.

For the experiments reported here, a test set containing 16,395 words was chosen from five seminars, providing a total of approximately 130 minutes speech material.

2.1. Language Model Training

To train *language models* (LM) for interpolation we used corpora consisting of broadcast news (160M words), *proceedings* (17M words) of conferences such as ICSLP, Eurospeech, ICASSP or ASRU and *talks* (60k words) by the Translanguage English Database. Our final LM was generated by interpolating a 3-gram LM based on broadcast news and proceedings, a class based 5-gram LM based on broadcast news and proceedings and a 3-gram LM based on the talks. The perplexity is 144 and the vocabulary contains 25,000 words plus multi-words and pronunciation variants.

2.2. 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.

As relatively little transcribed data is available for acoustic model training, the acoustic model used in the experiments reported here was trained on the *Broadcast News* [3] corpora and merged with the close talking channel of several meeting corpora [4, 5]. A total of 300 hours of speech material were used for system training.

The speech data was sampled at 16kHz. Speech frames were calculated using a 10 ms Hamming window. For each frame, 13 *Mel-Minimum Variance Distortionless Response* (Mel-MVDR) cepstral coefficients were obtained through a discrete cosine transform from the Mel-MVDR spectral envelope [6]. Thereafter, linear discriminant analysis was used to reduce the utterance based cepstral mean normalized features plus 7 adjacent to a final feature number of 42. Our baseline model consisted of 300k Gaussians with diagonal covariances organized in 24k distributions over 6k codebooks.

2.3. Acoustic Adaptation: Close Talking Speech

The adaptation of the close talking acoustic model was done in three consecutive steps:

1. A supervised Viterbi training of the CHIL adaptation speakers followed by a *maximum a posteriori* (MAP) combination of this model with the acoustic model of the original system: To find the best mixing weight, a grid search over different mixing weights was performed. The weight, which reached the best likelihood on the hypotheses of the first pass of the unadapted speech recognition system, was chosen as the final mixing weight.
2. A supervised *maximum likelihood linear regression* (MLLR) in combination with *feature space adaptation* (FSA) and *vocal tract length normalization* (VTLN) on the close talking CHIL development set: This step adapts to the speaking style of the lecturer and the channel. In the case of non-native speakers the adaptation should also help to cover some 'non nativeness'.
3. A second, now unsupervised MLLR, FSA and VTLN adaptation based on the hypothesis of the first recognition run: this procedure aims at adapting to the particular speaking style of a speaker and to changes within the channel.

2.4. Acoustic Adaptation: Far Field Speech

The adaptation of the far distance acoustic model was done in three consecutive steps:

1. Four iterations of Viterbi training on far distance data from NIST [7] and ICSI [8] over all channels on top of the acoustic trained models to better adjust the acoustic models to far distance.
2. A supervised MLLR in combination with FSA and VTLN on the far distance (single distance or MA processed) CHIL development set: This step adapts to the speaking style of the lecturer and the channel (in particular to the room reverberation). In the case of non-native speakers the adaptation should also help to cover some non-native speech.
3. A second, now unsupervised MLLR, FSA and VTLN adaptation based on the hypothesis of the first recognition run: this procedure aims at adapting to the particular speaking style of a speaker and to changes within the channel.

2.5. Signal Combination: Beamforming

A basic ingredient of classic beamforming techniques is the speaker location. Hence, to apply such techniques, a source localization algorithm is required.. The source localizer used for the experiments reported in Section 4 is based on the estimation of *time delays of arrival* (TDOA) with the *phase transform* (PHAT) [9],

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

where

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

The estimated TDOA $\hat{\tau}_i(t)$ is that which maximizes $R_{12}(\tau)$. This estimate is then compared to the predicted TDOA, given by

$$T_i(\mathbf{x}) = T(\mathbf{m}_{i1}, \mathbf{m}_{i2}, \mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{m}_{i1}\| - \|\mathbf{x} - \mathbf{m}_{i2}\|}{s} \tag{3}$$

where $\mathbf{m}_{i1}$ and $\mathbf{m}_{i2}$ are the positions of the microphones in the i-th microphone pair, $\mathbf{x}$ is the speaker location, and s is the speed of sound. The estimated speaker location is then that $\mathbf{x}$ which minimizes the squared error criterion

$$\epsilon(\mathbf{x}) = \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} \left[\hat{\tau}_i - T_i(\mathbf{x})\right]^2 \tag{4}$$

Klee *et al* [10] propose an algorithm whereby (4) is recursively minimized with a variation of a Kalman filter to effectively track a moving speaker.

In this work, we used a simple *delay and sum* (D&S) beamformer implemented in the subband domain. Subband analysis and resynthesis was performed with a *cosine modulated filter bank* (CMFB) [11, §8]. In the complex subband domain, beamforming is equivalent to a simple inner product

$$y(\omega_k) = \mathbf{v}^H(\omega_k)\mathbf{X}(\omega_k)$$

where ω_k is the center frequency of the k-th subband, $\mathbf{X}(\omega_k)$ is the vector of subband inputs from all channels of the array, and $y(\omega_k)$ is the beamformed subband output. The speaker position comes into play through the *array manifold vector* [12, §2]

$$\mathbf{v}^H(\omega_k) = \begin{bmatrix} e^{j\omega_k\Delta_0(\mathbf{x})} & e^{j\omega_k\Delta_1(\mathbf{x})} & \ldots & e^{j\omega_k\Delta_{N-1}(\mathbf{x})} \end{bmatrix}$$

where $\Delta_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{m}_i\|/s$ is the propagation delay for the i-th microphone located at $\mathbf{m}_i$.

2.6. Signal Combination: Blind Channel Combination

The channel combination techniques seek to separate the voice of a single speaker from the background noise and room reverberation, and thereby to improve the signal quality and concomitant recognition accuracy. One way to address this problem is *blind source separation*, for which several approaches have been proposed in the literature [13]. Assuming that the speech on all microphones is correlated while at least some of the noise is uncorrelated, we can simply time shift each channel by its delay with respect to a reference channel, sum together all shifted channels, then divide the sum by the number of channels N to improve the signal-to-noise ratio. This leads to a very simple yet effective means for combining the several channels; in future, we hope to investigate more elaborate blind techniques that have appeared in the literature. To estimate the TDOA, we can maximise (1). 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 [14]:

$$G(\omega) = G_{x_1,x_2}(\omega) - \frac{N_1(e^{j\omega\tau})N_2^*(e^{j\omega\tau})}{|N_1(e^{j\omega\tau})N_2^*(e^{j\omega\tau})|} \quad (5)$$

2.7. Text Combination: Confusion Network Combination

Confusion networks reduce the complexity of lattice representations to a simpler form that maintains all possible paths through the lattice, but transforms the space to a series of slots, each of which contains either a word hypothesis or a null arc, and an associated posterior probability. 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 WER over different microphones by selecting the word with the highest probability in each particular slot [15].

3. Data Collection and Labeling

The data used for the experiments described in this work was collected during a series of seminars held by students and visitors at the Universität Karlsruhe (TH), in Karlsruhe, Germany in November, 2004. The students and visitors spoke English, but mainly with German or other European accents, and with varying degrees of fluency. This data collection was done in a very natural setting, as the students were far more concerned with the content of their seminars, their presentation in a foreign language and the questions from the audience than with the recordings themselves. Moreover, the seminar room is a common work space used by other students who are not seminar participants. Hence, there are many "real world" events heard in the recordings, such as door slams, printers, ventilation fans, typing, background chatter, and the like.

The seminar speakers were recorded with a Sennheiser *close-talking microphone* (CTM), a 64-channel Mark III MA developed at the NIST (National Institute of Standards and Technologies) mounted on the wall, four T-shaped MAs with four elements mounted on the four walls of the seminar room and three Shure Microflex table-top microphones located on the work table where the position was not fixed. A diagram of the seminar room is shown in Figure 1. All audio files have been recorded at 44.1 kHz with 24 bits per sample. The high sample rate is desireable to permit more accurate speaker position estimation, while the higher bit depth is necessary to accommodate the large dynamic range of the far field speech data. For the recognition process, the speech data was down-sampled to 16kHz with 16 bits per sample. In addition to the audio data capture, the seminars were simultaneously recorded with four calibrated video cameras with a rate of 15 frames per second.

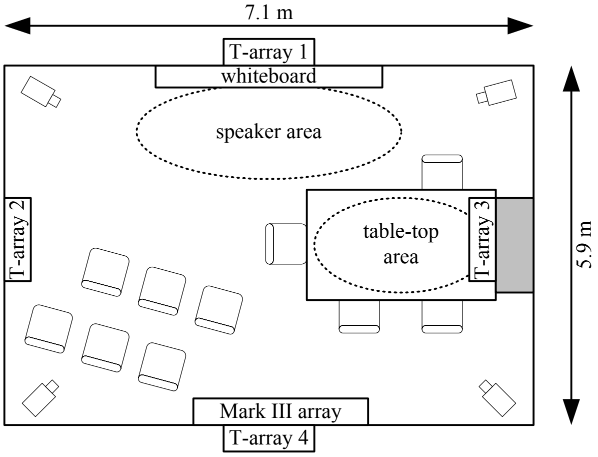

Figure 1: The CHIL seminar room layout at the Universität Karlsruhe (TH).

The data from the CTM was manually segmented and transcribed. The data from the far distance microphones was labeled with speech and non-speech regions. The location of the centroid of the speaker's head in the images from the four calibrated video cameras was manually marked every 0.7 second. Based on this marks the true position of the speaker's head in three dimensions could be calculated within an accuracy of approximately 10 cm [16].

4. Speech Recognition Experiments

All tests used the language and acoustic models described above for decoding. Even though the techniques to combine multiple microphones in our approach are very simple, by comparing the WER of Table 1 we see a significant gain by using multiple far distance microphones over a single distance microphone for three different types of combination; namely the use of MA processing, BCC and CNC. The disadvantage of the MA is that the speaker position has to be known. An estimate of this knowledge compared to the true speaker position results in a decrease in WER of 2.2%. Using a MA with an estimated speaker position over a single channel we gain back 26.9% of the accuracy compared to the CTM. In BCC and CNC no knowledge of the microphones geometry and speaker position has to be known. The BCC is a very simple technique while the latter has to use recognition runs on every single microphone which is very time consuming. Nevertheless, these two approaches are good methods to combine the table-top microphones as their signal quality is nearly equal. For the T-shaped microphones these two approaches fail to improve the performance over the best channel as the variance over the different channels is high, in particular as channel one is much better than the three other channels. Therefore, selecting only good channels could improve the overall performance.

The best performance was reached by beamforming the MA data, blindly combining all table top microphones, blindly combining a single microphone from every T-array and combining the three recognition runs by confusion networks. The blind channel combination of all microphones from the T-arrays is

Microphone Type	WER
close talk	34.0%
table-top	
single microphone (mic. 1)	62.4%
single microphone (mic. 2)	61.7%
single microphone (mic. 3)	62.2%
blind channel combination	59.3%
confusion network combination	61.0%
T-arrays (single microphone)	
single microphone (array 1)	60.9%
single microphone (array 2)	64.8%
single microphone (array 3)	66.5%
single microphone (array 4)	66.1%
blind channel combination	62.4%
confusion network combination	61.8%
Mark III array	
single microphone	66.5%
estimated position of the speaker	58.0%
true position of the speaker	55.8%
confusion network combination	
table-tops (CNC) & T-arrays (CNC)	59.8%
table-tops (BCC) & T-arrays (BCC)	59.7%
table-tops (CNC) & array (estimated)	59.3%
table-tops (CNC) & array (true)	59.2%
table-tops (BCC) & array (estimated)	56.9%
table-tops (BCC) & array (true)	55.7%
T-arrays (CNC) & array (estimated)	60.3%
T-arrays (CNC) & array (true)	60.2%
T-arrays (BCC) & array (estimated)	58.0%
T-arrays (BCC) & array (true)	57.0%
table-tops, T-arrays (CNC) & array (estimated)	58.4%
table-tops, T-arrays (CNC) & array (true)	58.3%
table-tops, T-arrays (BCC) & array (estimated)	55.6%
table-tops, T-arrays (BCC) & array (true)	55.0%

Table 1: *Word error rates* (WER)s for different single microphones and multiple microphones.

expected to lead to further gain.

In the future we want to use advanced techniques such as cepstral domain maximum likelihood beamformer [17] for the MA and replace BCC by blind source separation techniques. On the text level, incorporating a larger number of hypotheses on different microphones has improved results in all experiments where similar types of microphones has been used and similar WER has been reached. Therefore, we would expect this trend to continue for additional accuracy using more microphones, but time constraints limited our ability to run these larger experiments which will be done in the future. On different types of microphones and WER the use of CNC could not always lead to an improved accuracy. Furthermore, cross adaptation from hypothesis generated by a different microphone is expected to slightly improve the accuracy and could be explored in the future.

5. Acknowledgment

The work presented here was partly funded by the *European Union* (EU) under the project CHIL (Grant number IST-506909).

The authors wish to thank Kai Nickel, Keni Bernardin, Ulrich Klee and Tobias Gehrig for their assistance in collecting, labeling and converting the data used in this work.

6. References

[1] "Computers in the Human Interaction Loop", *http://chil.server.de*.

[2] M.C. Wölfel and S. Burger, "The ISL Baseline Lecture Transcription System for the TED Corpus", *Submitted to Eurospeech*, 2005.

[3] Linguistic Data Consortium (LDC), "English Broadcast News Speech (Hub-4)"," www.ldc.upenn.edu/Catalog/LDC97S44.html.

[4] F. Metze, C. Fügen, Y. Pan, T Schultz, and H. Yu, "The ISL RT-04S meeting transcription system", *in Proc. ICASSP-2004 Meeting RecognitionWorkshop. Montreal; Canada: NIST*, 2004.

[5] S. Burger, V. Maclaren, and H. Yu, "The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style", *ICSLP*, 2002.

[6] M.C. Wölfel, J.W. McDonough, and A. Waibel, "Warping and Scaling of the Minimum Variance Distortionless Response", *ASRU*, 2003.

[7] V. Stanford, C. Rochet, M. Michel, and J. Garofolo, "Beyond Close-talk - Issues in Distant speech Acquisition, Conditioning Classification, and Recognition", *ICASSP 2004 Meeting Recognition Workshop*.

[8] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI Meeting Project: Resources and Research", *ICASSP 2004 Meeting Recognition Workshop*.

[9] M. Omologo and P. Svaizer, "Acoustic Event Localization Using a Crosspower-spectrum Phase Based Technique"," in *Proc. ICASSP*, 1994, vol. II, pp. 273–6.

[10] U. Klee, G. Gehrig, and J.W. McDonough, "Kalman Filters for Time Delay of Arrival-Based Source Localization", *submitted to Eurospeech*, 2005.

[11] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, Englewood Cliffs, 1993.

[12] H. L. Van Trees, *Optimum Array Processing*, Wiley-Interscience, New York, 2002.

[13] 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.

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

[15] 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*, 2000.

[16] D. Focken and R. Stiefelhagen, "Towards vision-based 3-d people tracking in a smart room", *IEEE Int. Conf. Multimodal Interfaces*, 2002.

[17] D. Raub, J.W. McDonough, and M.C. Wölfel, "A Cepstral Domain Maximum Likelihood Beamformer for Speech Recognition", *ICSLP*, 2004.