Ad-Hoc Meeting Transcription on Clusters of Mobile Devices

Michele Cossalter, Priya Sundararajan, Ian Lane

Carnegie Mellon University

{michele.cossalter, priya.sundararajan, ian.lane}@sv.cmu.edu

Abstract

For all the time invested in meetings, very little of the wealth of information that is exchanged is explicitly preserved. In this paper, we propose a novel platform for meeting transcription using cellular phones for recognition. As most meeting participants carry cellular phones with them, this platform will allow meetings to be transcribed wherever they take place, without requiring any additional infrastructure. In this paper, we introduce our proposed platform, and compare three approaches for combining audio from multiple devices: microphone selection, either at signal or feature level, and combination of decoder outputs via confusion network combination. We evaluated the effectiveness of our cellular phone based platform on speech collected in a meeting environment, and found that the early microphone selection at signal level obtained a 16% improvement in speech recognition accuracy compared to using a single recording device. Moreover, this approach offered a comparable performance to multi-system confusion network combination, while requiring significantly lower computational cost.

Index Terms: far-field speech recognition, automatic meeting transcription, mobile devices

1. Introduction

Although many people spend a substantial portion of their work week in meetings, technological support for the meeting process is scant. Meeting records usually take the form of brief minutes and personal notes which are labor intensive to produce, and usually fail to capture much of the content of a meeting. A key engineering challenge for preserving the information that is exchanged is capturing it in a reliable, unobtrusive and flexible way. In this paper, we propose to use cellular phones for this task, as this enables transcription to be performed without any additional infrastructure anywhere a meeting may take place. Also, as most participants will carry a cell phone, by leveraging resources (microphone, CPU, and network bandwidth) across clusters of devices transcription accuracy may be improved compared to using a single device.

The challenges present in this task are three-fold. First, recognition of distance speech (even at a distance less than $1m$) is difficult, as reverberation, background noise, and overlapping speech from multiple speakers severely degrade the quality of the speech signal. Additionally, beamforming, which is a common approach for performing recognition of distance speech using multiple microphones, is impractical due to the lack of synchronization for acoustic sampling and the inherent clock drift on mobile devices. Second, recognition of conversation speech remains a significant challenge, due to poor enunciation, dis-fluencies, and lack of clear sentence boundaries. Additionally, with multiple participants barge-in effects and collisions due to overlapping speech frequently occur. Third, performing complex signal processing on networks of mobile devices and back-end servers while minimizing bandwidth and computational cost requires novel approaches to task management.

In this paper we focus on the first of these challenges, specifically addressing automated recognition of distance speech using multiple mobile devices. The main contributions of this work are:

1. We develop a novel platform for meeting transcription using clusters of mobile devices.
2. We compare three approaches for combining acoustic input from multiple devices: microphone selection, either at signal or feature level, and combination of decoder outputs via confusion network combination.
3. We evaluate the effectiveness of the above approaches on speech collected in a meeting environment on multiple mobile devices.

The remainder of the paper is structured as follows. Section 2 provides an overview on related work, while Section 3 presents the proposed transcription platform. The investigated approaches are described in Section 4. Experimental setup and results are discussed in Section 5. Finally, Section 6 concludes the paper with our future research directions.

2. Related Work

Transcription of meetings using multi-microphone automatic speech recognition has been investigated in a number of recent works [1, 2, 3]. In these works three main approaches were evaluated: acoustic space combination, best microphone selection, and output combination. If signals are sampled synchronously, multiple channels can be combined in the acoustic space, via beamforming techniques [4, 5]. As not all channels have similar quality for different utterances or speakers, channels may be weighted based on signal-to-noise ratio (SNR) [6].

Second, the best channel can be selected and used for decoding. Channel selection is usually performed based on SNR, using either a fixed or adaptive [1] threshold. A different approach consists of combining channels after feature computation. A channel selection method based on class separability was proposed by Wölfel [7]. Class separability, based on within-class and between-class scatter matrices, is computed for a number of classes derived by merge and split training, and the channel with the best class separability is chosen.

Finally, channels can be combined at the output level after decoding. The simplest approach for leveraging results from multiple decoders consists of choosing the channel with the highest likelihood [8]. A different selection criterion was proposed by Obuchi [9], based on a feature compensation metric. The most popular strategy is combination of decoder outputs through confusion networks [10]. Since this solution demands very high computational load, some approaches only use for combination a subset of channels selected based on SNR [6].

28 – 31 August 2011, Florence, Italy

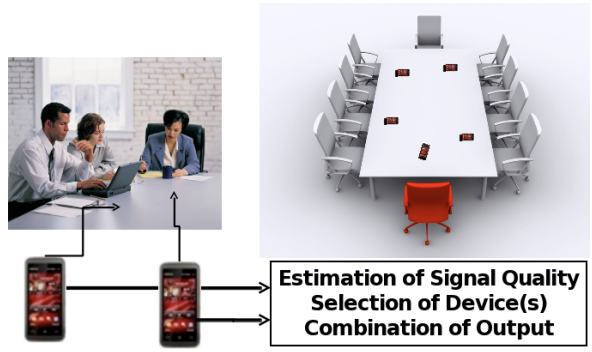

Figure 1: Using cellphones for meeting transcription allows for flexibility, since no infrastructure is required, and improved recognition by three steps: estimation of signal quality, selection of device(s), and combination of output.

3. A Platform for Meeting Transcription on Mobile Devices

A pictorial representation of the proposed transcription platform is shown in Figure 1. Meetings are recorded on the participants' cellular phones which are placed on the table as shown. Transcription involves three steps: estimation of signal quality, selection of devices (negotiation) and combination (in this work we focus on combination of speech recognition hypotheses across multiple devices).

The selected audio streams can be sent to a back-end server where the transcription process will take place. Hypotheses combination requires a bandwidth of approximately 250 Kbps per microphone of uncompressed audio, since each individual audio stream must be sent to the server. Microphone selection based on SNR or class separability is more appealing, as it is less computationally intensive and bandwidth demanding, and can be done dynamically over a cluster of phones.

4. Microphone Selection and System Combination Schemes

To effectively leverage multiple cellular phones for meeting transcription, we investigated three approaches to select or combine inputs sampled across a cluster of devices. In the first approach signal-to-noise ratio (SNR) is estimated for each device, and the single device with maximum SNR for a speech segment is selected as the active microphone at that time. Automatic speech recognition (ASR) is then performed on the audio collected from the best single device to generate the meeting transcript for that segment. The second approach applies the same method as above but uses a class separability measure (CSM) of the input MFCC features [7] for device selection. The third approach involves generating speech recognition hypotheses from the audio streamed from all devices and combining the hypotheses via confusion network combination (CNC) [10]. In this approach, the SNR estimated for each device is used to select the component weightings applied during combination. These three approaches are described in detail in the following subsections.

4.1. Microphone Selection via SNR Estimation

One simple approach for leveraging a cluster of devices to perform meeting transcription is to select the single best device based on a signal quality metric. Intuitively, the channel with the highest SNR will provide the decoder with a "cleaner" signal, thus increasing the likelihood that the recognition hypothesis will be correct. We compute SNR on each device as the ratio between estimates of the observed signal level and noise level. In this work, the noise level is estimated as the average root mean square (RMS) of the signal over all non-speech frames, while the signal level is estimated as the 85th percentile of the RMS distribution over speech frames. This metric, which is similar to the one used in the NIST Speech Quality Assurance package[1], was found to be reasonably robust, and it generally selected the same device for utterances from the same speaker.

4.2. Microphone Selection via Class Separability Measures

As an alternative to SNR, we evaluated device selection based on the class separability measure (CSM) proposed in [7]. The objective of CSM is to find the channel providing the sequence of feature vectors that have minimum variance within an acoustic class and maximum discrimination between classes. The channel maximizing class separability is selected for decoding. The separability measure is given by:

$$d_I = tr(S_w^{-1} S_b) \tag{1}$$

where the within-class (S_w) and between-class (S_b) scatter matrices are defined as:

$$S_w = \sum_i^c \left[\sum_j^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T \right] \tag{2}$$

$$S_b = \sum_i^c n_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{3}$$

n_i being the number of samples in the i-th class, μ_i the mean vector for the i-th class, and μ the mean vector over all classes. A more stable measure which does not rely on the non-singularity of S_w can be used instead of (1):

$$d_R = tr(S_b)/tr(S_w) \tag{4}$$

Based on the approach described in [7], we derived Gaussian Mixture Model (GMM)-based class models by applying the merge and split procedure on the training data. Two different approaches to model classes were investigated: the first, a single *combined model* trained and shared among all the channels; the second, a set of *individual models* trained for each channel independently. We also investigated two possible approaches to deal with silence, an *implicit model* in which all frames are considered without eliminating silence or the pauses, and an *explicit model* where silence frames are neglected and only speech frames are considered for computing the scatter matrices. As in [7], we found that using explicit silence and independent device models outperformed implicit silence and combined model. These results are described in the experimental evaluation in Section 5.2.

[1] http://www.itl.nist.gov/iad/mig//tools/

	imp-com	exp-com	imp-ind	exp-ind
d_I	24.1	23.9	22.1	27.1
d_R	23.5	23.4	22.8	23.7

Table 1: Accuracy (%) for CSM-based microphone selection using different approaches, evaluated on the AMI corpus.

4.3. System Combination via Dynamic Confusion Network Combination

One popular method for combining outputs from multiple decoders is confusion network combination (CNC), where word lattices are computed from each channel and then combined into a single confusion network [10]. We extend this approach by applying dynamic weights for combination, based on channel SNR. We found that weighting the channel with maximum SNR at 1.0 and other channels at 0.75 obtained the best performance for combination over multiple evaluation scenarios. We evaluate this approach in Section 5.3.

5. Experimental Evaluation

We evaluated the effectiveness of the three approaches described in Section 4 on two separate meeting corpora. In the first evaluation we simulated the placement of cellular phones in a meeting by selecting four far-field channels within each session of the AMI Meeting Corpus[2]. For the second experiment we evaluated performance using the Mobile Device Meeting Corpus, a small corpus collected locally consisting of speech recorded in a quiet meeting room using high-end cellphones.

5.1. Baseline System

A multi-condition acoustic model was trained using HTK [11] with speech collected from headset, lapel and far-field microphones from 168 sessions within the AMI Corpus. In total 202 hours of speech data were used during training. The resulting acoustic model consisted of 8000 codebooks with a maximum of 64 Gaussian components per codebook. For the AMI task, a 12k vocabulary, trigram language model was built using the SRILM toolkit [12]. The language model was trained using modified Kneser-Ney smoothing on the transcriptions of the training portion of the AMI corpus, 900k words in total. This language model obtained a perplexity of 94.20 on the evaluation set, consisting of the remaining 4 sessions in the AMI corpus (in total 1 hour of non-overlapping speech). Speech was manually segmented into utterances before performing recognition.

In the first evaluation, speech recognition was performed using a single pass speaker independent recognition system. The average speech recognition accuracy over all 16 speakers was 61.2% when a headset microphone was used, 56.5% for lapel microphones and 25.7% for the average far-field case. This performance is similar to that obtained from other groups for single pass, single microphone recognition on this data [7].

5.2. Single Microphone Selection

First, we evaluated the effectiveness of single microphone selection for transcription using SNR as the selection metric. For each utterance, SNR was computed independently for each microphone using the method described in Section 4.1, and the single microphone with maximum SNR was selected for decoding. Using this approach speech recognition accuracy improved from 25.7% to 29.0%, a relative improvement of 12% compared to the average far-field case.

[2] http://corpus.amiproject.org

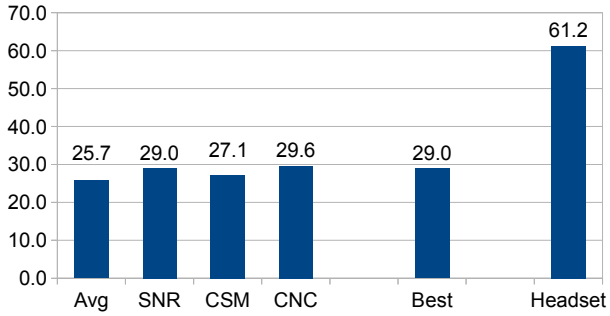

Figure 2: Accuracy (%) for different multi-channel decoding approaches, evaluated on the AMI corpus. *SNR* outperforms the average single channel (*Avg*) and achieves comparable performance to *CNC*.

Next, we evaluated the effectiveness of single microphone selection for transcription using the class separability measure (CSM) described in Section 4.2. For each utterance, CSM was computed independently for each microphone, and we compared the effectiveness of the three techniques introduced in Section 4.2: *implicit* (**imp**) and *explicit* (**exp**) modeling of silence; the use of *combined* (**com**) or *individual* (**ind**) models for each microphone channel; and computing the separability measure using d_I (1) or d_R (4).

Table 1 shows the results for CSM selection at the utterance level. We found that using *individual* models, *explicit* modeling silence and computing CSM using d_I (1) obtained the best performance. This outcome agrees with what was reported in [7], where the same result is stated with only partial supporting data. Using the best CSM method (*dI - exp - ind*) an accuracy of 27.1% was obtained, a 4% relative improvement in ASR accuracy compared to the average far-field case. In our experimental evaluation the SNR-based microphone selection was found to outperform CSM. This outcome disagrees with what published in [7], where a 7.7% relative improvement of CSM with respect to SNR is reported. This can be partially explained by the limited amount of data that we used to train the models. Also, the silence class was in some cases hard to select, as several classes got a similar number of assigned feature vectors.

5.3. System Combination using Dynamic CNC

Next, we evaluated the effectiveness of the Dynamic Confusion-Network Combination (CNC) approach described in Section 4.3. In this approach, speech recognition was performed on audio from each microphone channel and confusion networks were generated for each speech segment. The outputs were then combined using confusion network combination. The microphone with maximum SNR was assigned a weight of 1.0 and all other channels were set to 0.75. Using this approach an accuracy of 29.6% was obtained, which is significantly higher (14% relative) than the average far-field case, and slightly better (2% relative) than the SNR-based single microphone case. Compared to traditional CNC, where equal weights were applied for all utterances, Dynamic CNC was much more effective (29.6% compared to 27.0%).

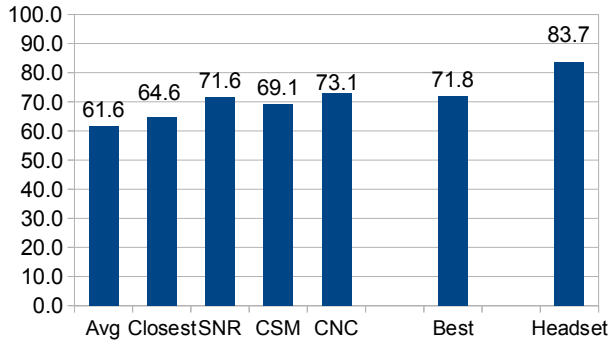

Figure 3: Accuracy (%) for different multi-channel decoding approaches, evaluated on the Mobile Device corpus. *SNR* outperforms the average single channel (*Avg*) and achieves comparable performance to *CNC*.

Figure 2 shows the speech recognition accuracy achieved for the three proposed approaches, *SNR*, *CSM*, and *CNC*, compared to the average single channel (*Avg*) and best single channel (*Best*, chosen a-priori) cases. All three approaches obtain higher speech recognition accuracy than the average far-field case. The best results are obtained by CNC, followed by SNR and finally by CSM. Although slightly higher accuracy is obtained with CNC compared to the single microphone SNR case, the computational cost and bandwidth requirements are significantly higher, in this case four times as large.

5.4. Evaluation on Mobile Devices

In the second experiment we evaluated the performance of the same three approaches described in Section 4 on speech collected on high-end cellular phones. In this evaluation six speakers sat in a $4.7m$ x $3.8m$ conference room, which contained a $2.6m$ x $1.1m$ table, and phones were placed naturally on the conference table in front of each speaker, similar to the meeting scenario shown in Figure 1. Each participant was also equipped with a headset microphone. Speakers read five news stories selected from recent newspaper articles for approximately 30 minutes. In total 790 utterances (consisting of 20k words) were collected with an average of 130 utterances per speaker. Speech recognition was performed applying the acoustic model described in Section 5.1 and a 64k language model trained on the English Gigaword corpus [13]. Recognition was performed using a two-pass decoding scheme, applying a speaker adapted acoustic model on the second pass.

Figure 3 shows the speech recognition accuracy achieved for the different approaches. Similar to the previous experiment, the single *SNR* obtained a 16% relative improvement in ASR accuracy compared to the average far-field microphone, and a 3.6% relative improvement with respect to *CSM*. Also, *CNC* obtained a 2% relative compared to the *SNR* case. Compared to the simulated experiments on the AMI corpus, the mobile devices and microphones used here are of higher quality and more realistically distributed. This might be the reason why the microphone selection approach is more effective in this case. Interestingly, *SNR* based selection obtained significantly higher accuracy than manually selecting the *closest* microphone for each speaker. This is likely due to the relative orientation of the speaker and the cellular phone.

6. Conclusion and Future Work

In this paper we addressed far-field recognition of meetings using multiple cellphones. We developed a research and data collection platform, and compared different approaches to improve speech recognition accuracy by combining audio from multiple randomly placed devices. Experimental results show that higher recognition accuracy can be achieved by taking advantage of multiple far-field microphones for decoding and a significant saving of computational effort is obtained by early selection of the best channel at the signal level with negligible loss of accuracy compared to output combination. In future work, we aim to develop a real-time implementation of the proposed approach and deploy it for transcription of meetings. The three most significant challenges in this development is accurate syncronization across devices, robust speech/non-speech detection and effectively handling multiple overlapping speakers.

7. References

[1] A. Janin, A. Stolcke, X. Anguera, K. Boakye, O. Çetin, J. Frankel, and J. Zheng, "The ICSI-SRI Spring 2006 Meeting Recognition System," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2006, vol. 4299, pp. 444–456.

[2] F. Metze, Q. Jin, C. Fügen, K. Laskowski, Y. Pan, and T. Schultz, "Issues in Meeting Transcription - The ISL Meeting Transcription System," in *Proc. of INTERSPEECH'04*, Jeju Island, South Korea, Oct 2004, pp. 1709–1712.

[3] M. Wölfel, S. Stüker, and F. Kraft, "The ISL RT-07 Speech-to-Text System," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Springer Berlin / Heidelberg, 2008, pp. 464–474.

[4] X. Anguera, C. Wooters, and J. Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[5] A. Stolcke, "Making the Most from Multiple Microphones in Meeting Recognition," in *to appear in Proc. of ICASSP'11*, Prague, Czech Republic, May 2011.

[6] 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'06*, Pittsburgh, PA, USA, Sep 2006, pp. 361–364.

[7] M. Wölfel, "Channel Selection by Class Separability Measures for Automatic Transcriptions on Distant Microphones," in *Proc. of INTERSPEECH'07*, Antwerp, Belgium, Aug 2007, pp. 582–585.

[8] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech Recognition Based on Space Diversity Using Distributed Multi-Microphones," in *Proc. of ICASSP'00*, Istanbul, Turkey, Jun 2000, pp. 1747–1750.

[9] Y. Obuchi, "Multiple-Microphone Robust Speech Recognition Using Decoder-Based Channel Selection," in *Proc. of SAPA'04*, Jeju, Korea, Oct 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 Language*, vol. 14, no. 4, pp. 373–400, 2000.

[11] P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev, and S. Young, "The 1994 HTK large vocabulary speech recognition system," in *Proc. of ICASSP'95*, Detroit , MI, USA, May 1995, pp. 73–76.

[12] A. Stolcke, "SRLIM - An Extensible Language Model Toolkit," in *Proc. of ICSLP'02*, Denver, CO, USA, Sep 2002, pp. 901–904.

[13] D. Graff and C. Cieri, "English Gigaword," Linguistic Data Consortium, ISBN: 1-58563-260-0, Philadelphia, Jan 2003.