A STUDY ON RECOGNIZING DISTORTED SPEECH OVER LOCAL DISTRIBUTED TRANSDUCER NETWORKS

Yong Zhao, Sunghwan Shin, Enrique Robledo-Arnuncio & Biing-Hwang (Fred) Juang

Center for Signal and Image Processing
Georgia Institute of Technology
{yongzhao,sshin31,era,juang}@ece.gatech.edu

ABSTRACT

In a collaborative scenario, a multiplicity of portable devices may constitute a network of distributed microphones, without a clearly defined geometric configuration or synchronization that can be taken advantage of for traditional microphone array processing to enhance the acquired signal. This application scenario represents a severe, but interesting challenge for automatic speech recognition systems. In this paper, we investigate a variety of robust speech recognition techniques with a focus on the distributed transducer scenario. We also report some important study results that lead to new thinking in the design of robust speech recognition for broadened applications. Two issues that are inherent to distributed transducer networks are specially investigated. First, we study the effect of the sampling rate skew of microphones to the system performance; second, we explore the possibility of combining recognition hypotheses from multiple transducer channels for improved recognition accuracy.

Index Terms— Robust speech recognition, distributed transducer network, sampling rate skew, system combination.

1. INTRODUCTION

The increasing prevalence of mobile devices such as cellphones, PDAs, and laptop PCs in place of traditional telephones, brings new possibilities for speech recognition applications. In one example, a user may hold his/her cellphone or PDA in the front (rather than at the ear) for short message (SMS) dictation. In another collaboration application, the use of speech recognition tools to transcribe audio streams captured by portable devices has been proposed [1], [2]. In a collaborative (meeting) scenario, a multiplicity of these mobile devices may be used, constituting a network of distributed transducers or microphones, without a clearly defined geometric configuration or synchronization that can be taken advantage of for traditional microphone array processing to enhance the acquired signal. These application scenarios represent a severe, but interesting challenge for speech recognition systems.

The challenge here may be considered within the realm of robust automatic speech recognition, the goal of which is to maintain satisfactory recognition accuracy under arbitrary operating conditions. It is well known that the current technology suffers substantial performance degradation if it is operated in a mismatch condition, i.e. the condition in which the recognizer was not designed for.

The mismatch between training (design) and testing (deployment) conditions can be the result of a number of factors. The mismatch from inter-speaker variability include accent, dialect, and speaking rate differences; the speaking environment mismatches involve interfering noise, noise-induced change in speaking styles (the Lombard effect), (linear) channel distortion due to variations in the microphone response, and voice network trans-coding. These factors have been extensively studied in the past two decades in a certain degree [3], [4].

Earlier approaches dealing with the mismatch problem can be broadly classified into three categories. The first category is often referred to as speech enhancement methods that attempt to clean up the signal so as to produce acoustic features that are very close to those obtained under a matched condition. The second category aims at robust feature selection and comparison, which chooses a feature representation or a dissimilarity measure that is not overly sensitive to environmental changes, thereby achieving robust performances. The third category is in the area of adaptation methods, which are designed to transform either the observation (feature) vector or the model in response to an estimated change of the operating conditions (including a change in speaker). As an example, central mean normalization (CMN) [5] and its extension, cepstrum bias removal methods, are simple and robust methods that are suitable for dealing with convolutional distortions. These traditional methods tend to address one single dimension of the general robustness problem. The aforementioned new application scenarios, nevertheless, contain additional dimensions of the potential mismatch condition that need to be accounted for.

It is thus the purpose of this paper to elaborate on the new dimensions of the mismatch condition, with a focus on the distributed transducer scenario, and to report some important study results that lead to new thinking in the design of robust speech recognition techniques for broadened applications.

In the aforementioned application scenario, distributed devices are connected via a local area network, likely to be of an ad hoc type. Speech data acquired at these distributed devices need to be coded and transmitted to a server where further processing is performed. Thus the server has access to

multiple channels of the acoustic signal, although each of which is independently acquired. This configuration gives rise to many new issues that significantly affect the performance of an intelligent system that converts speech data into text via the use of an automatic speech recognition system. Key issues involved in this scenario include non-linear distortion introduced by the speech codecs, the trans-coding distortions that are usually difficult to model, room acoustic responses that are hard to estimate or track (a talker speaking at an unspecified distance and location relative to the transducers), background noise and interference (say, stray conversations on the side irrelevant to the main talker's speech), and the heterogeneity of the devices such as sampling rate skews and a lack of synchronization and calibration of signals. These issues constitute new challenges as we expand the scope of applications of automatic speech recognition systems.

In the following section, we elaborate these somewhat new adverse conditions and discuss our methodology of study in hopes of developing some insights as well as effective methods to mitigate the potential problems that come with these factors. An additional issue, also worthy of study, is the possibility of integrating multiple channels of speech input with multiple speech recognition systems for enhanced recognition performance. The new scenario provides an opportunity for the application of known algorithms, such as the ROVER [6], which fuses individual channel results to produce improved recognition performance.

The remaining of the paper is organized as follows: Section 2 discusses the issue of sampling rate mismatch among distributed microphone array. Section 3 describes the method of combining output from multiple microphones for better recognition performance. The procedure of collecting the speech database is reviewed in Section 4, and experiments and results are described in Section 5.

2. ISSUES OF SAMPLING RATE SKEW

The lack of a common clock reference is a fundamental problem when dealing with audio streams originating from or heading to different distributed sound capture or playback devices. Traditional multichannel digital signal processing tools, which take audio from multiple transducers as their input, such as blind source separation (BSS) and acoustic echo cancellation (AEC), will not work as expected if the incoming audio streams are not synchronized. Thus, the benefit of multiple audio inputs to speech recognition may be substantially discounted as a result of the lack of sampling synchronization.

In discrete commercial components, the tolerance of sampling rate skew, usually specified in parts per million (ppm), can range from just a few ppm's to many hundreds of ppm's. Furthermore, the specific operating frequency of these devices is temperature-dependent. For example, the data sheet of a popular audio chipset for portable devices specifies that its sampling frequency closest to 8 kHz can actually be 8.0182 kHz for certain master clock frequencies. This sampling frequency corresponds to a mismatch of more than 2000 ppm.

Regarding the estimation of the rate mismatch, one approach described in [7] is to transmit some known calibration sound signal to all of the acquisition devices. To avoid the interference with the ongoing application, the calibration needs to be carried through a quiet channel, such as an FM radio receiver, available to all the devices. Unfortunately, such a channel may not be available in many applications. Another possibility for rate estimation is to use network packet time-stamps, as proposed in [8]. The time-stamp information is readily available, but the estimation is challenging due to the jitter in the network delays. Obtaining an accurate and fast estimate of the rate mismatch remains a challenge.

All of these challenges make it necessary to first understand what the exact synchronization requirements for a given application are. In [9], we have presented a set of experimental results illustrating the impact of sampling rate mismatches in batch BSS and adaptive AEC applications. The results reveal that adequate synchronization is a key for the performance of the speech enhancement algorithms. In this paper, we present an experimental study to analyze the sensitivity of speech recognition to variant rate mismatch factors.

3. COMBINING OUTPUTS FROM MULTIPLE MICROPHONES

A straightforward approach using multiple microphones to improve the speech recognition performance is through beamforming. With prior knowledge of the microphone locations, a microphone array can enhance the signal coming from the direction of a desired speaker and suppress undesired interference and noise from other directions. However, traditional microphone array techniques are not directly applicable to the current situation of distributed microphones. The main challenges of the distributed microphones are as follows: the spatial topology of the transducer network is uncertain and may change over time, and there is no common clock to synchronize all microphones. In addition, these microphones are heterogeneous and have different gains, system responses and signal-to-noise ratios.

In this paper, we investigate the possibility of combining multiple signals at the word hypothesis level. Specifically, we apply ROVER [6] to combine the recognition outputs from multiple microphones. The ROVER algorithm was originally proposed to improve the performance of speech recognition by combining multiple speech recognizers. It first constructs a confusion network by aligning all the system outputs through an iterative procedure. Then a rescoring process follows to select a word sequence with the best score among all sequences that traverse the network. Since the nature of ROVER is to extract a consensus hypothesis from multiple decision alternates of the same objective, this technique could be applied to combine recognition hypotheses from multiple microphone channels.

4. SPEECH DATABASE

An experimental database based on the original TIMIT dataset was constructed for research in practical distributed transducer

network applications. The recording was performed in a conference room equipped with sound attenuating wall panels and acoustic ceiling tiles. Utterances of the original TIMIT database were played back through a loudspeaker, and captured by a variety of commercial portable devices for a realistic signal quality. The resulting speech corpus is referred to as TIMIT DM. Table 1 briefly describes the types of these recording devices. Built-in automatic gain control (AGC) and noise suppression provisions were disabled in those portable devices that support such features. In addition to the microphone channels of these portable devices, a wired microphone was used as a reference channel to record synchronously with the playback device.

The portable devices and the reference microphone were located on a round conference table, about 3 feet away from the PDAs. The reverberation time of the room during the recordings was approximately 200 milliseconds. The ambient acoustic noise was primarily due to AC air flow and pipe-transmitted vibrations. The acoustic noise level measured between 33 and 36 dBA at the location of the portable devices.

Table 1. Recording devices used for speech data acquisition.

Device ID	Model
HP1	HP iPAQ rx3100 PDA
DL2	Dell Axim x51v PDA
HT3	HTC S710 cellphone
NB4	Dell Inspiron 8500 notebook
AT5	Audio Technica AT899 omnidirectional condenser microphone

5. EXPERIMENTS AND RESULTS

In this section, we evaluate the capabilities of different robust speech recognition techniques in the context of the distributed transducer network using the TIMIT DM corpus. We built a baseline phoneme recognition system based on the hidden Markov model (HMM) paradigm. The experimental conditions are similar to those established by Lee and Hon in their benchmark experiments [10]. All phones are modeled as 3-state strict left-to-right context-independent HMMs. Each state observation density is modeled by a 64-component Gaussian mixture density with diagonal covariance matrices. The input feature is a 39-dimension vector of 12 MFCC's and log energy, and their first and second order time derivatives. These models are trained with the maximum likelihood (ML) method implemented by HTK [1]. Recognition is carried out by a Viterbi search that uses a phone bigram language model.

5.1. Baseline system

Table 2 shows phoneme accuracy with respect to different microphone channels. The recognizers are trained and tested within the data set of the same channel. The results under this matching condition can be seen as an upper bound for robust recognition in mismatched conditions. The first column, labeled as *Clean*, is the result using the original TIMIT corpus for both training and testing.

The phoneme accuracy of the portable devices ranges from 48.39%-61.16%. It is observed that the three PDA devices (HP1, DL2, and HT3) perform substantially better than Channel NB4 of the laptop PC, which is in large part due to the amount of noise generated by moving parts of the laptop, such as CPU fans and hard disks in operation. Moreover, the recognition rates of the PDA devices are comparable with that obtained using the reference microphone, AT5, which partially indicates that in terms of sound quality, portable devices are approximately equivalent to wired microphones as unit components in deploying distributed recognition networks.

Table 2. Phone recognition results with respect to different distributed devices evaluated under matching conditions.

	Clean	HP1	DL2	HT3	NB4	AT5
Phone acc. (%)	70.41	61.16	58.88	63.05	48.39	59.50

A second experiment is performed to evaluate the performance of distributed recognition systems under mismatching conditions. Two kinds of recognition systems are examined. The first system is trained using TIMIT clean data and the second is trained using data from Channel HT3 of TIMIT DM, which is chosen due to its best recognition rate among all portable devices in the matching condition. The two recognizers are evaluated using the testing data from portable devices of TIMIT DM. The experiment with the second recognizer can be interpreted as the recognition under the mismatch that is mainly from the switch of microphone channels, while the experiment with the first recognizer confronts with various kinds of mismatches, including background noise, room reverberation, and characteristic of microphones.

Table 3. Phone recognition results with respect to different distributed devices evaluated under mismatching conditions.

Clean	HP1	DL2	HT3	NB4
Baseline	34.70	33.56	35.59	23.00
CMN	37.60	37.54	51.89	24.34

HT3	HP1	DL2	HT3	NB4
Baseline	35.31	42.62	-	30.30
CMN	48.54	49.80	-	33.88

Table 3 shows the recognition performance of the two recognizers in a baseline configuration, and one configuration using CMN to compensate for feature mismatch. First, the recognizer trained with Channel HT3 usually outperforms the recognizer trained with clean data, since the data of Channel HT3, though noisy and distorted, matches better to the testing condition than does the clean data. CMN improves the accuracy of speech recognition as expected. Another observation concerning the CMN method is that it does not improve the

accuracy of Channel NB4 as much as that of the other channels. Noticing that Channel NB4 is intervened by a high additive noise from surrounding components of the laptop, we could say that CMN falls short of mitigating the additive noise, and thus the gain of CMN decreases in the condition of noise level mismatch.

5.2. Effects of sampling rate skew

We evaluated the effect of recognizing speech as a result of sampling rate mismatch between the training data and the testing data where the recognizers are trained and tested within data set from the same channel (except that the testing data is re-sampled into a different sampling rate).

Table 4 shows the range of the recognition results of different channels when we exponentially increase values of sampling rate mismatch from 0 to 8,192 ppm's. It is hard to observe any substantial effect of the sampling rate mismatch to the recognition system in the context of recognition using individual channels. We attribute the robustness of the recognition systems against sampling rate skew to the characteristics of MFCC acoustic features, which represent speech waveforms in the form of filter banks and to a good degree tolerate the deviation of frequencies.

Table 4. Phone recognition results with respect to different sampling rate mismatches. The first row records the actual average sample rate mismatches of each channel relative to the AD/DA converters used for playback and reference recording.

	HP1	DL2	HT3	NB4
Rate mismatch (ppm)	141	5860	942	-6140
Min phone acc. (%)	61.07	58.60	62.94	48.23
Max phone acc. (%)	61.23	58.92	63.20	48.41

5.3. Combining hypotheses from microphone channels

We tested ROVER for combining the recognition outputs from the distributed microphone network. All of the four portable devices are configured for recognizing under matching conditions, i.e. the recognizer of each channel is trained and tested using data acquired from the same channel, and no sampling rate skew is introduced. The recognition accuracies of component devices are listed as in Table 2. The consensus output yields a phone accuracy of 63.47%, which is a slight improvement over the best single system of a phone accuracy of 63.05%.

6. CONCLUSION

In this paper, we described our ongoing research in the use of distributed transducer network for improved robust speech recognition. We studied different robust speech recognition techniques in the context of the distributed transducer network using the TIMIT DM corpus. We showed that CMN provides complementary benefits to speech recognition in channel distortion. We did not observe the effect of sampling rate skew of individual channels to the recognition performance. We also found that the system combination algorithm that combines the recognition outputs from multiple channels yields a slight improvement over the output of the best single channel.

Nevertheless, in a typical telecollaboration setting where automatic transcription of speech needs to deal with time-varying processes of high reverberation and non-linear distortions, the current robust speech recognition techniques are not powerful enough to handle them, and new robust modeling approaches need to be developed.

REFERENCES

[1] R.C. Rose, S. Parthasarathy, B. Gajic, A.E. Rosenberg, and S. Narayanan, "On the implementation of ASR algorithms for hand-held wireless mobile devices," in *Proc. ICASSP*, 2001.

[2] B. Zhou, Y. Gao, J. Sorensen, D. Dchelotte, and M. Picheny, "A hand-held speech-to-speech translation system," in *Proc. IEEE ASRU workshop*, 2003.

[3] B.H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, vol. 5, pp. 275-294, 1991.

[4] Y. Gong, "Speech recognition in noisy environments: A survey," *Computer Speech and Language*, vol. 16, pp. 261-291, 1995.

[5] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp.1304-1312, 1974.

[6] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. IEEE ASRU Workshop,* 1997.

[7] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "on the importance of exact synchronization for distributed audio signal processing," in *Proc. ICASSP*, 2003.

[8] D. Miljkovic, T. Trump, and G. Petrovic, "Clock Skew Compensation by Speech Interpolation," in *Proc. Int. Conference on Digital Telecommunications*, 2006.

[9] E. Robledo-Arnuncio, T.S. Wada, B.H. Juang, "On Dealing with Sampling Rate Mismatches in Blind Source Separation and Acoustic Echo Cancellation," in *Proc. IEEE ASPAA workshop*, 2007.

[10] K.-F. Lee and H.-W. Hon, "Speaker-Independent Phone Recognition using Hidden Markov Models," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 11, pp. 1641-1648, 1989.

[11] S. Young, G. Evermann, M. Gales, et al, "The HTK Book for HTK Version 3.4," *Cambridge University Press*, Cambridge, UK, 2006.