Robust speech recognition using microphone arrays and speaker adaptation

Ofentse Noah, Nicholas Zulu, Daniel Mashao
Department of Electrical Engineering, University of Cape Town
Rondebosch, Cape Town, South Africa
ofentse@star.za.net pzulu@crg.ee.uct.ac.za daniel@eng.uct.ac.za

Abstract—This paper proposes the use of a low complexity microphone array combined with speaker adaptation for a speech recognition system. Microphone arrays are known to reduce different sources of acoustic degradation by producing directionally sensitive gain patterns in a desired speaker's direction and reducing the gain in the direction of undesired sound sources. Speaker adaptation aims at achieving speaker dependent (SD) performance for speaker independent (SI) recognition systems. In this contribution we propose the combination of microphone array and adaptation techniques as processing stages for a Hidden Markov Model (HMM) based text-independent speech recognition system. Our aim is to show that the combination of speech enhancement and speaker adaptation would greatly improve speech recognition when a system is tested in adverse conditions.

1. INTRODUCTION

There are many situations where the use of a close talking microphone is undesirable, inconvenient or impossible such as in a car, a conference or a dictation system where the position of a speaker is neither known nor fixed. A four-microphone array acquisition system is proposed for an HMM-based speaker recognition system. Work done previously [1] on speech recognition in a noisy office environment showed some performance improvements due to the use of a microphone array. The use of microphone arrays for speech recognition is dependent on the possibility that microphone arrays can obtain signals of improved quality compared to signals recorded using a single microphone [2]. Microphone arrays reduce noise and reverberation components in the input speech. Using a linear array of a four microphones we aim to evaluate the performance of various beamforming algorithms combined with speaker adaptation on a speech recognition task.

Speech recognition systems can be designed in two ways; *speaker dependent* (SD) and *speaker independent* (SI) [3]. An SD system is a system which is trained using one speaker. This system has very good performance for the speaker it was trained with due to the fact that it models the speaker very well. A well trained SD system requires a few hours of training data from a speaker. It is not feasible to design a speech recognition system using this design if it will be used by a very large population. For example, it is not feasible to build such a system for the South African population because we would require over 40 million hours of speech data. However an SI system is a system that is trained using many different speakers. This system does not have SD type performance for a particular speaker due to the fact that it doesn't model a particular speaker well. Furthermore if the system is trained with native speakers, it may not work well for non-native speakers [5]. Such a system requires hundreds of hours of speech data for it to be well trained [6]. This data is collected by sampling parts of the population where 30 seconds of speech or more from a speaker in the population sample is collected. It is therefore feasible to implement this system for a large population; albeit it does not work as well as the SD system.

Speaker adaptation is the process of improving the performance of an HMM-based speech recognition system [3]. In particular, speaker adaptation is used to obtain SD-like performance from the initial SI speech recognition system [5]. Two basic speaker adaptation approaches exist [4]. The first is a spectral mapping speaker adaptation approach and the second is a model mapping speaker adaptation approach. The first approach includes techniques which convert the speech from the new speaker into feature vectors which match the training speech feature vectors. The second approach modifies the SI speech recognition system such that its modeling better matches the adaptation data. In this paper, we focus on using model mapping speaker adaptation with microphone arrays for robust speech recognition.

2. SPEAKER ADAPTATION AND MICROPHONE ARRAYS

Speaker adaptation is the process of improving the performance of an HMM speech recognition system, most importantly using considerably less data than in training. As described before, speaker adaptation can be done using two approaches, which are spectral mapping and model mapping. In this paper we look into model mapping speaker adaptation where we adapt the initial speech recognition engine using speech acquired using a microphone array. As we see below in Figure 1 the adaptation data is used in the transformation process of the SI speech recognition system. The transformation transforms the SI speech recognition system model into the new SD speech recognition system model indicated in the figure. Adaptation speech is usually obtained from a source which may be contaminated by extraneous noise like background noise, environment acoustics, reverberation effects and channel distortions [10].

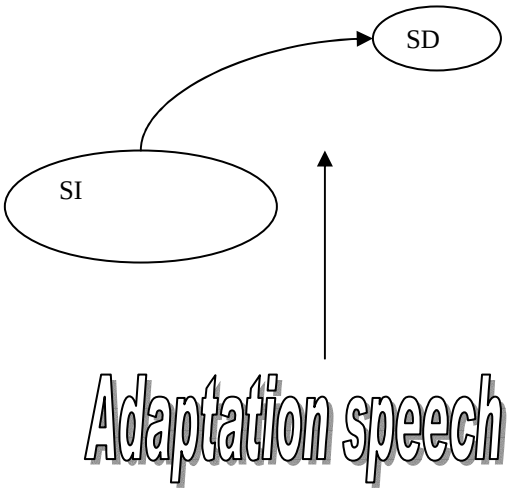

Figure 1: Speaker adaptation process.

Microphone arrays are known to reduce the effects of background noise and reverberation effects, thereby making them an alternative speech input to speech recognition systems for hands free applications. Figure 2 below shows a block diagram that includes the basic components of the proposed system. The microphone array acquires the speech signals and array processing is done using beamforming algorithms. The beamformed signal then acts as input to a speech recognition module that has an active speaker adaptation module.

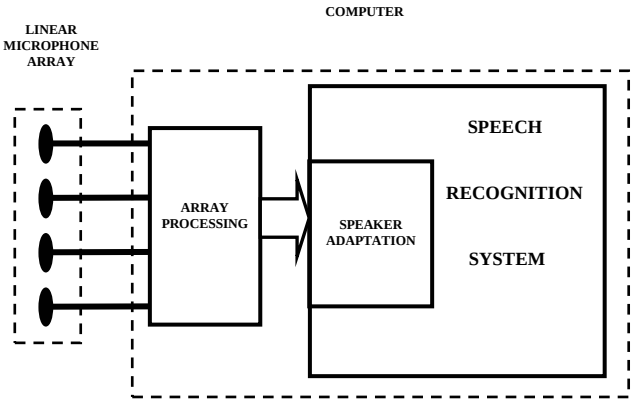

Figure 2: Block diagram of the proposed speech recognition system.

[1] D. Giuliani, M. Omologo, and P. Svaizer, "Robust Continuous Speech Recognition using a Microphone Array," presented at Eurospeech, Madrid, 1995.

[2] C. Che, Q. Lin, J. Pearson, B. d. Vries, and J. Flanagan, "Microphone Arrays and Neural Networks for Robust Speech Recognition," presented at ARPA Workshop on Human Language Technology, NJ, 1994.

[3] C. J. Leggetter and P. C. Woodland, “Maximum likeli-hood linear regression for **speaker adaptation** of contin-uous density hidden Markov models,” *Comp. Spch. &Lang.*, vol. 9, pp. 171–185, Apr. 1995.

[4] Jonathan E. Hamaker, MLLR: A **Speaker Adaptation Technique** for **LVCSR**, *Lecture for a course at ISIP -Institute for Signal and Information Processing*, Department of Electrical and Computer Engineering, Mississippi State University, 1999.

[5] M.J.F. Gales and P. C. Woodland, “Mean and variance adaptation within the MLLR framework,” *Comp. Spch. &Lang.*, vol. 10, pp. 249–264, Apr. 1996.

[6] Language Data Consortium, http://www.ldc.upenn.edu/, Last Accessed: 31 May 2005.

[7] Adcock.J, Mashao.D et. al., "Microphone-array speech recognition via incremental MAP training", *In proc: ICASSP*, Vol 2. pp.879-900, 1996.

[8] C. Che, Q. Lin, J. Pearson, B. de Vries, J.L. Flanagan, Microphone arrays and neural networks for robust speech recognition, in: Proceedings of ARPA Human Language Technology (HLT), 1994, pp. 342–348

[9] D. Giuliani, M. Omologo, P. Svaizer, Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation, in: Proceedings of ICSLP, 1996, pp. 1329–1332

[10] S. Nakamura, T. Takiguchi, K. Shikano, Noise and room acoustics distorted speech recognition by HMM composition, in: Proceedings of IEEE ICASSP, 1996, pp. 69–72

O. Noah and N. Zulu are currently pursuing MSc degrees in Electrical Engineering at the University of Cape Town and are both in their final years of study.

Dr. D. Mashao is a senior lecturer at the University of Cape Town and head of the Speech research and Technology Group. He is also the supervisor of the above-mentioned authors.