PERFORMANCE EVALUATION OF AN AUTOMATIC SPEECH RECOGNISER INCORPORATING A FAST ADAPTIVE SPEECH SEPARATION ALGORITHM

Kutluyıl Doğançay

School of Electrical and Information Engineering
University of South Australia
Mawson Lakes, Australia

Jason Littlefield

Command & Control Division
Defence Science & Technology Organisation
Edinburgh, Australia

Ahmad Hashemi-Sakhtsari

Command & Control Division
Defence Science & Technology Organisation
Edinburgh, Australia

ABSTRACT

This paper addresses the performance evaluation of a speaker-dependent automatic speech recogniser (ASR) that employs a speech separation algorithm as a front-end processor. The ASR software used is Dragon NaturallySpeaking (NS) Professional Version 6.1. The word recognition accuracy of NS is known to be very sensitive to background noise due to competing speakers, as well as ambient and environmental disturbances. In this work, a reduced complexity fast-converging adaptive decorrelation filter (ADF) is used to successfully reduce the interference from competing speakers. The recognition accuracy of NS for speech utterances before and after front-end separation was measured. A significant improvement has been observed with the proposed front-end processing.

1. INTRODUCTION

Modern ASRs perform well in quiet environments, but very poorly in the presence of background noise and interference from competing speakers [1]. ASRs used in conferences, meetings, and command and control centers are particularly susceptible to cross-talk interference from competing speakers. In these situations a front-end speech processing unit that cancels interference from competing speakers is desirable to improve the percentage recognition accuracy of the ASR. Cancellation of interference from competing speakers can be done effectively by using speech separation algorithms.

In cases where a single microphone is employed to record multiple speakers, the effectiveness of speech separation techniques based on pitch and harmonic estimation [2] is rather limited. However, techniques that use multiple microphones for multiple speakers provide more effective separation by exploiting the diversity and independence of speech signals (see e.g. [3]).

This paper describes an experiment that evaluates the effectiveness of a reduced complexity fast-converging ADF speech separation algorithm for two competing speakers. The separated speech signals are applied to a commercially available speaker-dependent ASR, *Dragon NaturallySpeaking (NS) Professional Version 6.1*.

The major contributions of this paper can be summarized as follows:

- Performance evaluation of a speaker-dependent ASR subjected to interference from competing speakers.
- Development of a reduced complexity fast-converging ADF front-end processing for separation of competing speakers.

The paper is organized as follows. Section 2 provides background information on NS. Section 3 describes the model for speaker separation and the ADF algorithm with improved convergence. The experimental method and results are described in Section 4. The conclusions are drawn in Section 5.

2. SPEAKER-DEPENDENT ASR

NS is a speaker-dependent, large vocabulary, continuous speech ASR that is commercially available and widely used. Speaker-dependent ASRs require a training phase prior to use. The training phase builds a user profile by combining an acoustic model based on phonetic analysis of the speaker voice with a language model. The ASR transcribes the speaker's speech into text with the aid of the user profile generated in the training phase.

This work was partially supported by a DSTO research contract.

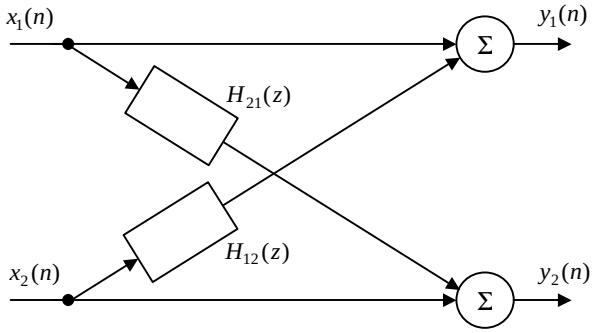

Fig. 1. Convolutive mixture model for two speakers.

Over eighty factors have been found to affect the performance of ASRs [4,5]. Some influential factors include speaker dialect or degree of enunciation, amount of training, speaking rate, vocabulary size and vocabulary confusability. ASRs are particularly susceptible to channel and environmental disturbances such as background noise and cross-talk from interfering speakers.

A previous study has shown that the best performance for a speaker-dependent ASR is achieved when the speech-to-interference ratio (SIR) during the testing of an ASR is maximized regardless of the SIR level during training [6]. However, at any given test SIR, the best performance is achieved when the training SIR matches the test value. Therefore, as expected, the best performance is achieved when the interference is minimal. This justifies the desirability of front-end processing to eliminate cross-talk interference from other speakers by means of speech separation.

3. ADAPTIVE SPEECH SEPARATION

3.1. Convolutive Mixture Model

In a multi-speaker, multi-microphone environment, a microphone will not only pick up the intended source signal, but also interferences from other competing speakers. In this paper, we consider the two-speaker, two-microphone case for the sake of simplicity. The interference signal is subjected to reverberation due to room acoustics, which implies that the interference is obtained by convolving the interfering source (speaker) signal with the reverberation channel impulse response. This gives rise to a convolutive mixture model for the microphone signals as shown in Fig. 1 [3]. The signals $x_1(n)$ and $x_2(n)$ are digitized source (speaker) signals, and $y_1(n)$ and $y_2(n)$ are digitized microphone signals that contain cross-talk from the interfering speaker. The reverberation channels have the transfer functions $H_{12}(z)$ and $H_{21}(z)$. The distances between the speakers and their microphones are assumed to be very small, so the direct

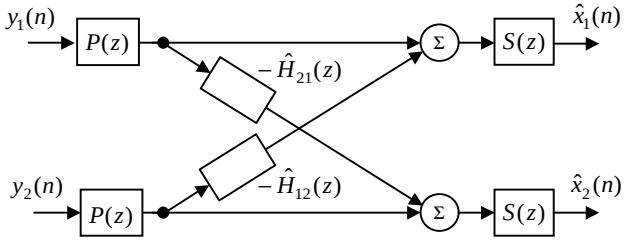

Fig. 2. Separation of speech signals. $P(z)$ is a preprocessor and $S(z)$ is a postprocessor.

reverberation channels can be safely ignored.

In the *z*-domain, the microphone signals are given by

$$\begin{bmatrix} Y_1(z) \\ Y_2(z) \end{bmatrix} = \begin{bmatrix} 1 & H_{12}(z) \\ H_{21}(z) & 1 \end{bmatrix} \begin{bmatrix} X_1(z) \\ X_2(z) \end{bmatrix}.$$

The solution to the separation problem involves estimation of the reverberation channels and matrix inversion:

$$\begin{bmatrix} X_1(z) \\ X_2(z) \end{bmatrix} = \frac{1}{1 - H_{12}(z)H_{21}(z)} \times \begin{bmatrix} 1 & -H_{12}(z) \\ -H_{21}(z) & 1 \end{bmatrix} \begin{bmatrix} Y_1(z) \\ Y_2(z) \end{bmatrix}.$$

Fig. 2 shows an implementation of the above equation with reverberation transfer functions replaced by their estimates. The postprocessing block $S(z)$ implements the inverse of $1 - H_{12}(z)H_{21}(z)$. The preprocessing block $P(z)$ is explained in Section 3.3.

3.2. Adaptive Decorrelation Filter (ADF)

ADF is based on the premise that the separated speech signals will be decorrelated if the source signals come from statistically independent sources [7]. This premise will be true for source signals from different speakers. A reduced complexity version of the ADF adaptation algorithm is given by [8]

$$\mathbf{a}(n+1) = \mathbf{a}(n) + \mu_a \mathbf{v}_2(n) v_1(n)$$
$$\mathbf{b}(n+1) = \mathbf{b}(n) + \mu_b \mathbf{v}_1(n) v_2(n)$$

where μ_a and μ_b are stepsizes,

$$\mathbf{a}(n) = [a_0(n), a_1(n), \ldots, a_{N_a-1}(n)]^T$$
$$\mathbf{b}(n) = [b_0(n), b_1(n), \ldots, b_{N_b-1}(n)]^T$$

$$\hat{H}_{12}(z) = \sum_{k=0}^{N_a-1} a_k(n) z^{-k}$$

$$\hat{H}_{21}(z) = \sum_{k=0}^{N_b-1} b_k(n) z^{-k}$$

$$\mathbf{v}_1(n) = [v_1(n), v_1(n-1), \ldots, v_1(n-N_b+1)]^T$$

$$\mathbf{v}_2(n) = [v_2(n), v_2(n-1), \ldots, v_2(n-N_a+1)]^T$$

$$v_1(n) = y_1(n) - \sum_{k=0}^{N_a-1} a_k(n) y_2(n-k)$$

$$v_2(n) = y_2(n) - \sum_{k=0}^{N_b-1} b_k(n) y_1(n-k)$$

The computational complexity of the above algorithm is $O(N_a + N_b)$. The conditions for convergence of ADF to true reverberation channels were derived in [9].

3.3. Convergence Improvement for ADF

The convergence rate of ADF is excruciatingly slow especially for speech signals because of nonstationarity and lowpass spectral characteristics of speech. The decorrelation citerion is insensitive to linear filtering of individual microphone output signals, and it would yield the same reverberation estimates if linear filters were inserted after the $y_i(n)$. The convergence rate of ADF can be greatly improved by preprocessing the microphone outputs $y_i(n)$ prior to ADF. The objective of the preprocessor is to flatten the spectrum of the $y_i(n)$. The simple preprocessor $1-0.95z^{-1}$, which is a fixed highpass filter (HPF) commonly used in LPC processors as a *preemphasizer*, provides a significantly faster convergence. The use of a HPF is in fact a simple, yet very effective, substitute for whitening. For white signals, ADF would attain its fastest convergence rate. The postprocessor in this case may include the inverse HPF. Fig. 3 compares the convergence rates of ADF with and without HPF preprocessing for speech signals of approximate duration 3 min. The squared coefficient error is given by $\|\mathbf{a}(n) - \mathbf{g}_1\|^2 + \|\mathbf{b}(n) - \mathbf{g}_2\|^2$ where the entries of $\mathbf{g}_1$ and $\mathbf{g}_2$ are defined through $H_{12}(z) = \sum_{k=0}^{N_a-1} g_{1,k} z^{-k}$ and $H_{21}(z) = \sum_{k=0}^{N_b-1} g_{2,k} z^{-k}$. The reverberation channel impulse responses are shown in Fig. 4. The stepsizes used were 0.03 and 0.1 for ADF with and without HPF preprocessing, respectively, and they were chosen to attain the fastest convergence.

4. SPEECH SEPARATION EXPERIMENT

The speech separation experiments were done in a room

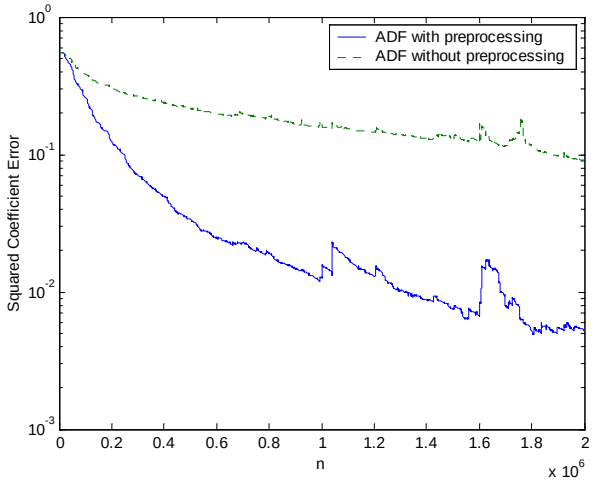

Fig. 3. Squared coefficient error curves for ADF with and without preprocessing.

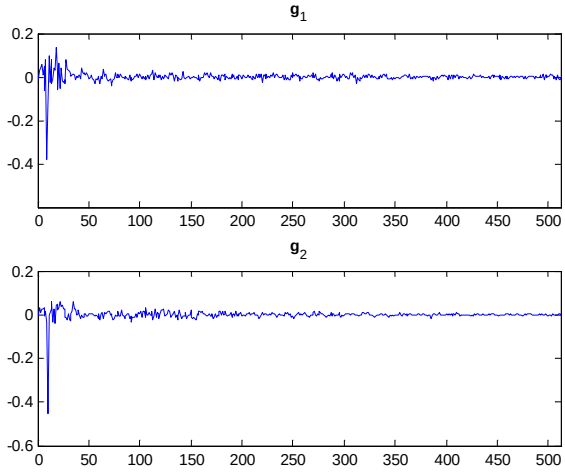

Fig. 4. Impulse responses for $H_{12}(z)$ and $H_{21}(z)$.

depicted in Fig. 5. To ensure repeatability of the experiments for different speakers, the reverberation channel impulse responses were first estimated for two speakers, using ADF with HPF preprocessing, and then the estimates were used to obtain the mixed microphone signals for other speakers as in Fig. 1. The microphone signals were recorded using a sampling frequency of 11025Hz and 16-bit resolution. The reverberation channels used for mixing speakers are depicted in Fig. 4.

Ten speakers performed the training process in NS to generate the NS user profiles where the "Australian English" language model and the "General" vocabulary were selected. In the second stage of the experiment, approximately three minutes (about 600 words) of spoken English from the ten speakers were recorded one at a time and digitized to produce the *source signals.* The speakers

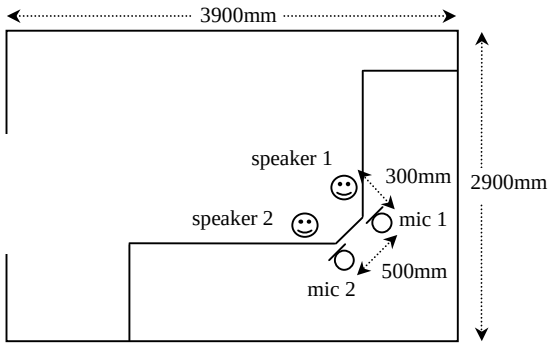

Fig. 5. Acoustic environment where cross-channel reverberations were measured.

read from two different selections of training text provided by NS. Five read one text selection and five read another. The *source signals* were grouped into pairs, $x_1(n)$ and $x_2(n)$, and mixed together using the convolutive mixture model. This produced $y_1(n)$ and $y_2(n)$, the microphone signals with cross-talk from the interfering speaker *before processing*. The variance estimates for the source signals are shown in Table 1. The ADF front-end processing with HPF preprocessor was applied to the microphone signals to obtain estimates of the source signals $\hat{x}_1(n)$ and $\hat{x}_2(n)$, i.e., the microphone signals *after processing*. The lengths of the ADF adaptive filters were set to $N_a = N_b = 512$.

The source signals, and the signals with cross-talk before and after processing were transcribed by an experimenter to produce reference patterns. These text patterns were compared to the hypothesis patterns produced by the recogniser. The texts were compared using a scoring program called *Sclite* from the US National Institute of Standards and Technology (NIST) to give the percentage word accuracy results (WA%) listed in Table 1. The average word accuracy for the ten speakers was 70.6% in the absence of cross-talk and dropped to 29.9% when the microphone signals were corrupted by interference. The fast converging ADF front-end processing improved the word accuracy from 29.9% to 62.9%.

5. CONCLUSION

We have evaluated the performance of NS incorporating a fast converging ADF front-end processor in an experiment that involved competing speakers. A significant performance improvement has been observed compared to the case of no front-end speech separation. The fast convergence of the proposed adaptive separation algorithm has been confirmed using real speech signals as spoken utterances.

Spk.	Mic.	Variance x 10^{-3}	Source Signal x_i	Before Proc. y_i	After Proc. $\hat{x}_i$
	i		WA %	WA %	WA %
1	1	0.254	80.7	24.6	72.5
2	2	1.865	85.4	49.6	80.9
3	1	0.156	59.9	25.8	58.4
4	2	0.106	69.2	27.6	66.2
5	1	0.759	71.1	50.4	66.4
6	2	0.061	67.1	9.2	52.8
7	1	2.550	68.7	47.6	61.7
8	2	0.202	67.6	10.3	48.5
9	1	0.337	76.7	37.7	69.7
10	2	0.410	59.1	16.1	51.7
Mean			70.6	29.9	62.9

Table 1. Performance of speech recogniser in terms of percentage word accuracy for source and processed speech from ten speakers.

6. REFERENCES

[1] R. Cole *et al.*, "The challenge of spoken language systems: Research directions for the nineties," *IEEE Trans. Speech Audio Proc.*, vol. 3, pp. 1-21, Jan. 1995.

[2] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 407-424, Sep. 1997.

[3] K. Yen and Y. Zhao, "Adaptive co-channel speech separation and recognition," *IEEE Trans. Signal Proc.*, vol. 7, pp. 138-151, March 1999.

[4] W. A. Lea, "What causes speech recognizers to make mistakes?" in *Proc. ICASSP*, pp. 2030-2033, Paris, May 1982.

[5] D. S. Pallett, "Performance assessment of automatic speech recognisers," *Journal of Research of the National Bureau of Standards (USA*), 90(5), pp. 1-17, 1985.

[6] J. Littlefield and A. Hashemi-Sakhtsari, "The effects of background noise on the performance of an automatic speech recogniser," *DSTO Internal Report* (in press).

[7] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Signal Proc.*, vol. 1, pp. 405-413, October 1993.

[8] K. Yen and Y. Zhao, "Improvements on co-channel speech separation using ADF," in *Proc. ICASSP*, pp. 1025-1028, 1998.

[9] S. Van Gerven and D. Van Compernolle, "Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness," *IEEE Trans. Signal Proc.*, vol. 43, pp. 1602-1612, July 1995.