DUAL-MICROPHONE ROBUST FRONT-END FOR ARM'S-LENGTH SPEECH RECOGNITION

Holly Francois, David Pearce, James Rex

Motorola Labs, Basingstoke, UK
Holly.Francois@motorola.com

ABSTRACT

This paper describes a novel method of improving the performance of a speech recognition front-end in non-stationary background noise. A two-microphone array has been designed that both enhances the speech and provides a continuous estimate of the background noise. This processing has been integrated with the standard ETSI DSR Advanced Front End so that the continuous noise estimate is an input to the first stage of Wiener filtering, increasing the noise suppression and hence the recognition performance. Tests with real-world noise have shown the recognition error rate is reduced by up to 50% when compared with single-microphone input to the same Advanced Front End.

1. INTRODUCTION

Robustness is an essential issue in practical deployment of automatic speech recognition (ASR) technology. In portable devices such as cell phones, various acoustic environments interfere with speech and reduce recognition performance. The dual-microphone algorithm described in this paper improves recognition performance in non-stationary background noise. This is particularly important when devices are held at arm's length, rather than close-talking next to the head.

The algorithm described in this paper uses microphone array processing that provides beamforming to improve the SNR of the signal and also generates a continuous estimate of the noise. This noise estimate is then used for further noise suppression in the short-term spectral domain.

Single-microphone noise estimation is based on identifying the gaps in the speech and then averaging the noise in the gaps. The noise estimate is then fixed during the speech and updated again in the next gap. This is effective for stationary noise, but does not respond sufficiently to noise which is rapidly time-varying. In addition the accuracy of this noise estimate is dependent on the performance of the voice activity detector (VAD). If the background noise is speech, for example, then the VAD cannot distinguish between desired and undesired speech, so the noise estimate is never updated.

Using a continuous noise estimate enables much better performance in non-stationary noise, particularly in background speech. For speech recognition, this algorithm has been integrated with the ETSI Distributed Speech Recognition (DSR) Advanced Front-End (AFE) [1]. Dual-microphone processing can improve the performance of local recognition, embedded in the device, as well as DSR for distributed speech and multimodal services.

2. DUAL MICROPHONE PROCESSING

The array processing algorithm uses the input from two microphones in two different ways; to form a beam on the desired talker which slightly improves the SNR of the noisy speech signal and, more importantly, to form a null on the desired talker and thus obtain a continuous noise estimate.

The polar plot in Figure 1 shows the directivity of a beam formed by adding the signals from two microphones. The microphones are omnidirectional and placed 4cm apart. The speech source is in the x-direction, and the microphones are broadside to it, i.e. along the y-axis. The plot shows the response at five different frequencies, and it can be seen that the spatial directivity of the beam is only apparent above 2kHz. At lower frequencies the beam is very broad, and therefore provides only slight noise reduction.

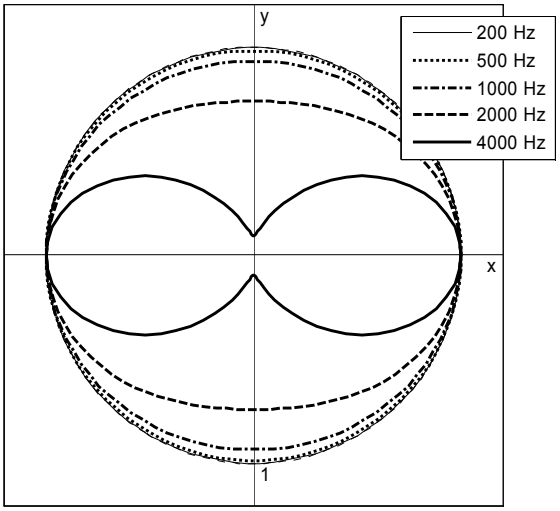

Figure 1. *Directivity plot of Beamforming array – talker is in the x direction*

The plot in Figure 2 shows the directivity of a null formed by subtracting the signals from the same two microphones [3]. The null is much narrower than width of the corresponding beam shown in Figure 1. The noise-estimating array is thus able to cancel the talker whilst receiving the noise coming from most other directions. In this way an estimate of the noise is formed. The advantage of cancelling the speech is that it allows the noise to be estimated during speech utterances, as well as during gaps.

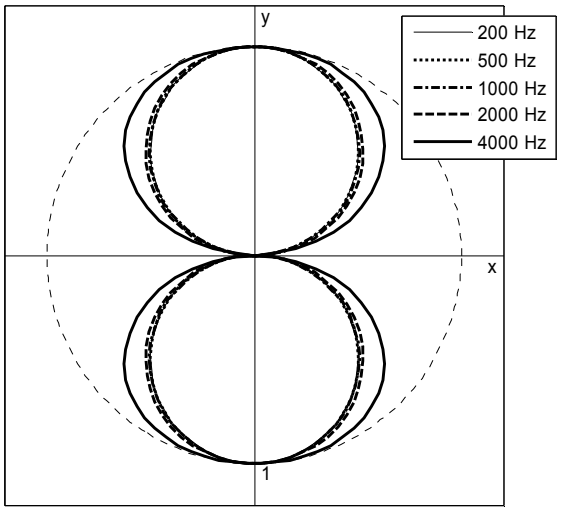

Figure 2. *Directivity plot of Noise Estimating array – talker is in the x direction*

Figure 3 shows a block diagram of the array-processing algorithm. The output from the two microphones is used to form two arrays that feed signals into the Dual-Channel Advanced Front-End (DCAFE) for speech recognition. The beamforming array provides a speech signal (A), which is slightly less noisy than the speech from its microphones, but still has noise present. An estimate of this noise component (B) is provided to the DCAFE by the noise-estimating array.

The first stage of processing is to compensate for any differences between the microphones. The frequency response of pairs of microphones of the same type does not vary significantly, however the sensitivities of microphones is more variable. A scalar gain-equalisation factor must therefore be applied to one of the input signals. The gain-equalised signals are then summed to form the beam, slightly increasing the SNR of the output speech signal (A). Simultaneously the signals are also subtracted, cancelling the speech and putting a null on the talker. Since the output of the speech-cancelling array (B) only covers a slice of the spatial noise field, the resulting noise estimate is not equal to the noise component of the noisy speech signal (A). Also, since the directivities of both the beam and the null are frequency-dependent, the difference between the noise estimate and the noise in the signal also varies with frequency. The noise estimate is therefore passed through a filter, EQN, which is designed to equalise the spectrum of the noise estimate to that of the noise component

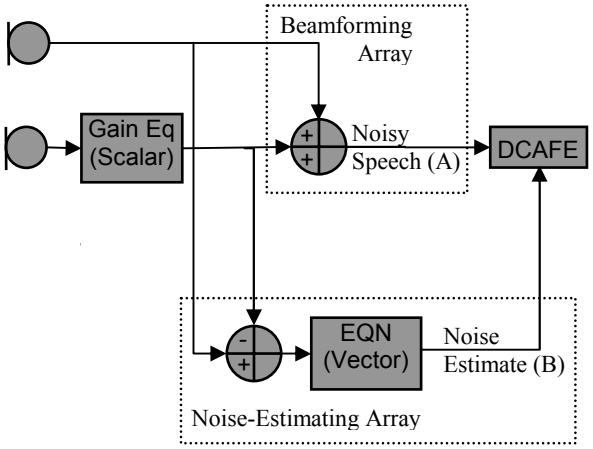

Figure 3. *Block diagram of array-processing algorithm*

of the noisy speech signal (A). The noise estimate (B) and the noisy speech signal (A) are then passed to the DCAFE. The gain-equalisation factor and the noise-equalisation filter are implemented by weighting the signals in the frequency domain.

3. ADVANCED FRONT-END

The "front-end" of an automatic speech recogniser extracts features from a speech waveform, which are then passed to a "back-end" for recognition. These features are usually related to the power spectra of short-time windows (10ms frames) of the waveform.

3.1. Original Single-Channel AFE

In the DSR Advanced Front-End [2], noise-reduced cepstral features are calculated from the incoming digital signal. The algorithm uses a two-stage Mel-warped Wiener filter noise reduction scheme (see Figure 4). After noise reduction, SNR-dependent waveform processing (SWP) is applied to the denoised signal. The output signal from SWP is used for cepstrum calculation. Finally, blind equalization is applied to the cepstral features. Only the first stage of Wiener filtering in the noise-reduction block will be discussed further, since this is where the modifications to form the DCAFE are made.

The input signal, at 8ksample/s, is split into frames of 200 samples. Each frame is Hann-windowed, and its spectrum is estimated using an FFT of length 256. The spectral resolution is reduced from 129 to 65 frequency bins by averaging

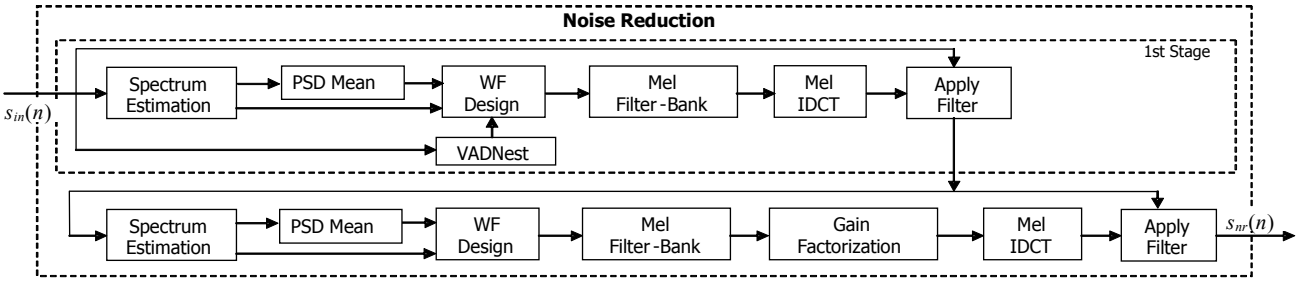

Figure 4. *Block scheme of noise reduction for single channel AFE*

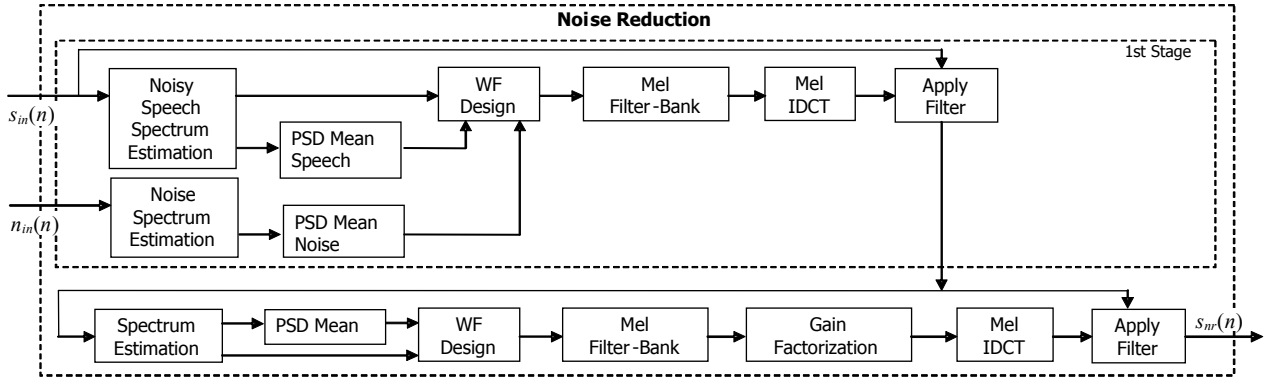

Figure 5. *Block scheme of noise reduction for dual channel AFE*

consecutive pairs of bins. Next, the *PSD Mean* (Power Spectral Density mean) block estimates the power spectrum, averaged over two successive frames to reduce its variance.

The current-frame spectrum and corresponding speech/non-speech decision from the *VADNest* block (Voice Activity Detector for Noise estimation) are used in the *WF Design* block to estimate the Wiener filter frequency characteristic. *VADNest* is an energy-based voice activity detector. The current frame is labeled as speech when the difference between the current frame log energy and the long-term estimate of non-speech log energy exceeds a threshold. The frames labeled as non-speech are used to update the noise estimate by adding a small portion of the current noise frame to the previous noise estimate. In other words, the single-channel noise estimate is a rolling average of the noise in the gaps.

3.2. Dual-Channel AFE

The Dual-Channel Advanced Front-End is an extension of the AFE that takes two inputs, one from the beamforming array and one from the noise-estimating array. Initially, the two array outputs are processed identically, to estimate power spectra of both the noisy speech and the noise, averaged in both frequency and time to reduce variance. The DCAFE then uses the continuous noise estimate from the array in the Wiener filter design process, see Figure 5, in place of the single-channel noise estimate derived from recent gaps. Using this noise estimate gives greater noise suppression than a single-microphone AFE, because the continuous noise estimate is more accurate during utterances than one derived from recent gaps, especially when the noise is varying rapidly.

4. SPEECH DATABASE

A new two-microphone speech database was required for evaluating the dual-channel processing. The database was recorded at 48 kHz sample rate, 24 bits, and then downsampled to 8 kHz, 16 bits, for all subsequent processing. The omnidirectional microphones were those used in a production mobile phone handset, and were placed 4 cm apart. The source speech was played out from a B&K mouth simulator, placed 40cm from the microphones.

4.1. Recording Environments

The database was recorded in typical phone-usage environments, so each microphone recorded a realistic acoustic mixture of ambient noise and speech. The noise came from both diffuse and point sources, and reverberation was also present. Recordings were made in the following environments:

- Meeting Room: Quiet, but with some stationary air-conditioning noise, short reverberation time, approximately 20 dB SNR.
- Quiet Canteen: Some background noise, occasional loud noises, long reverberation time, 10 dB SNR.
- Busy Canteen: Continuous loud background speech, long reverberation time, -3 dB SNR.
- Interfering talker: Meeting room with additional interfering speech played from a loudspeaker, 2 dB SNR.

The microphone pair was sited in front of the mouth simulator, and broadside to it. The interfering talker was simulated by an additional loudspeaker, placed at a similar distance, at an angle of approximately 45 degrees, see Figure 6.

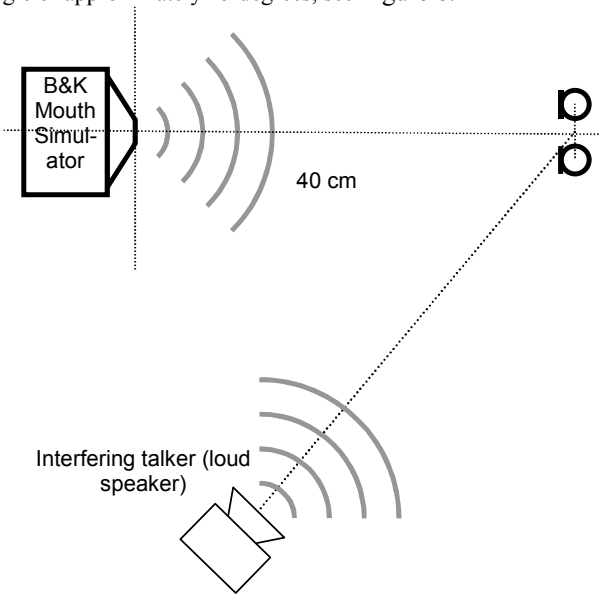

Figure 6. *Speech & noise recording set-up*

4.2. Source Sounds

One minute of speech-coloured noise was recorded in quiet conditions for equalising the microphone gains. The short-term power spectrum of speech-coloured noise is stationary and equal to a long-term average speech spectrum.

The section of the TI Digits database that was used in Aurora 2 [4] was used for testing speech recognition accuracy. This data consists of English digits, both isolated and strings, spoken by various North American male and female adults, originally recorded with a close-talk microphone under quiet conditions. The training set (used for training the speech recogniser) contains 8441 utterances and has a duration of 248 minutes, while the test set contains 1001 utterances with a total duration of 30 minutes. The training set was only recorded in quiet conditions, while the test set was recorded in the four environments listed above.

5. ARRAY TRAINING

No knowledge of the noise type or acoustic environment has been assumed since the handset will be used in a wide variety of scenarios. The offline array training simply involves compensating for variation between the microphones, then adjusting the power spectrum of the noise estimate so that it matches the noise present in the noisy speech signal, as described in Section 2.

5.1. Weights for Gain Equalisation

The beamforming weights are calculated by recording one minute of speech-coloured noise on both microphones in a quiet environment and then finding the difference in the gain of microphone 2 relative to microphone 1. This gain factor is then applied to each FFT bin of the complex spectrum of microphone 2, so that the gain of the microphones is equalised, this is represented by Gain Eq in Figure 3.

5.2. Weights for Noise Estimation

The weights for the noise estimation are calculated in two stages. First the equalised signals from the two microphones are subtracted, forming a null in the broadside direction. A noisy test file recorded in the busy canteen is then processed with these weights, with EQN set to 1. This test file is then used to find EQN. The power spectral density of the noise estimate in the gaps is equalised to the noise in the noisy speech signal from the beamforming array.

6. RESULTS

The recognition performance was tested using the Aurora 2 set up of HTK [4] as the back-end recognizer, with both the original AFE and the new DCAFE as front-ends.

It can be seen from the results in Table 1 that the DCAFE improves the speech recognition performance for all noise types, particularly in the presence of highly non-stationary noise. The reduction in word error rate is greatest for the single interfering talker, as this case is an extreme example of a non-stationary background noise. The multiple talkers and general noise in the canteen environment produce a babble effect which is more stationary, hence the smaller, but still significant improvement.

Test Condition	AFE	Dual Channel AFE	
	Word Accuracy	Word Accuracy	Error Reduction Relative to Single Channel
Meeting Room	98.6%	98.6%	2%
Normal Canteen	89.7%	93.7%	39%
Busy Canteen	26.4%	52.0%	35%
Interfering Talker	47.3%	75.3%	53%

Table 1. *Recognition Performance*

7. CONCLUSIONS

This paper has described a two microphone algorithm for improving the noise robustness of the DSR Advanced Front End. Microphone array processing is used both to cancel noise and to provide a continuous estimate of the noise, which replaces the VAD-based noise estimate in the AFE. Performance measurements in environments with non-stationary noise (e.g. in a canteen or near an interfering talker) have shown that the addition of second microphone reduces speech recognition error rate by between 34% and 50%.

8. REFERENCES

[1] ETSI standard ES 202 050 “Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithm”, Oct 2002

[2] D Macho, L Mauuary et al, “Evaluation of a Noise-Robust DSR Front-End on Aurora Databases”, ICSLP 2002, Sept 2002.

[3] A Alvarez et al, “Speech Enhancement and Source Separation based on Binaural Negative Beamforming” Eurospeech 2001, Sept 2001.

[4] H G Hirsch & D Pearce, “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions”, ISCA ITRW ASR2000 “Automatic Speech Recognition: Challenges for the Next Millennium”, Sept 2000.