Far-field ASR on Inexpensive Microphones

Laura Docio-Fernandez[1,2], *David Gelbart*[2], *Nelson Morgan*[2]

[1]E.T.S.I. Telecomunicacion
Dpto. Teoria de la Señal y Comunicaciones
Universidad de Vigo – 36200 Vigo-Spain
ldocio@gts.tsc.uvigo.es

[2]International Computer Science Institute*
1947 Center Street, Suite 600
Berkeley, CA 94704–1198
{ldocio,morgan,gelbart}@icsi.berkeley.edu

Abstract

For a connected digits speech recognition task, we have compared the performance of two inexpensive electret microphones with that of a single high quality PZM microphone. Recognition error rates were measured both with and without compensation techniques, where both single-channel and two-channel approaches were used. In all cases the task was recognition at a significant distance (2–6 feet) from the talker's mouth. The results suggest that the wide variability in characteristics among inexpensive electret microphones can be compensated for without explicit quality control, and that this is particularly effective when both single-channel and two-channel techniques are used. In particular, the resulting performance for the inexpensive microphones used together is essentially equivalent to the expensive microphone, and better than for either inexpensive microphone used alone.

1. Introduction

As part of the research conducted within the Meeting Recorder Project at ICSI [1], we are considering the issue of robust speech recognition of meetings using far-field microphones, particularly *inexpensive* far-field microphones. We envision a future personal digital assistant (PDA) which would have the ability to record and transcribe multi-party meetings. The need to recognize speech coming from all the talkers participating in the meeting means that the distance between talker and microphone is higher than for most PDA speech recognition applications.

The relatively uncontrolled acoustic environment, variation in talker position with respect to the microphone, and the widely variable characteristics of mass-produced inexpensive microphones all contribute to the difficulty of such a task. Background noise such as fans, door slams, and music can all contribute to the acoustic background. Reverberation can also be a significant problem using far-field microphones.

Two types of distant microphones are being used in our work: high-quality Crown Pressure Zone ("PZM") (roughly $300), and inexpensive conventional electret microphones ("ELC") (roughly 50 cents). Such electret microphones display significant variability of frequency response and sensitivity (output level) across individual mics. Recognition experiments have shown major differences between the two electret microphones we used.

*This work has been partially supported by the Spanish Ministry of Science and Technology, the Natural Sciences and Engineering Research Council of Canada, the German Ministry for Education and Research, and Qualcomm. Thanks to Jim Beck and Dan Ellis for assembling the PDA mockup used in experiments.

Omnidirectional microphones are, by definition, sensitive to sounds from all directions: many sounds from several meters away including keyboard clicks and general desktop noises can induce recognition errors. On the other hand, highly directional single microphones tend to be overly sensitive to changes in talker position. When possible, it is attractive to adopt microphone arrays, although their application is significantly limited by their additional cost and size.

The companies that develop speech recognition portable devices now use inexpensive electret microphones. These microphones work surprisingly well, but the lack of effective quality control leads to a large variability in the properties of the output signal (e.g., frequency response).

This paper compares the ASR performance of an omnidirectional PZM table-mounted microphone with that of two *cheap and low-quality* omnidirectional electret (ELC) elements mounted about 8 cm apart on a mock-up of a future PDA, placed on the same conference room table (see [1] for a picture of the mock-up).

In addition, we present new sets of robust speech features obtained from the combination of the two ELC microphone signals. We show that (1) relatively simple signal processing steps can make the performance of inexpensive microphones much closer to that of high-quality microphones, and (2) that adding the second microphone, even though it is spatially very close to the first one, is useful in a moderately noisy and reverberant environment.

2. Two-channel signal processing techniques

Although humans can hear with one ear only – "monaural hearing" – hearing with two functioning ears is clearly superior. This binaural hearing is a key part of the robustness of the human auditory system for recognizing speech in a wide range of environmental conditions. In this section, we describe two methods for extracting robust speech features using the signals from two mics.

2.1. Delay-and-sum processing

In delay-and-sum beamforming [2], delays are inserted after each microphone to compensate for the arrival time differences of the speech signal to each microphone. These time-aligned signals are added together. This has the effect of reinforcing the desired speech signal while the unwanted off-axis noise signals are combined in a more unpredictable fashion. The signal-to-noise ratio of the total signal is greater than (or at worst, equal to) that of any individual microphone's signal.

In order to compute the delay between the speech signals

we have used the "information theoretic delay criterion" delay estimation algorithm proposed in [3].

Delay-and-sum processing is very attractive because of its simplicity. It is easy to implement and can be done in a very cost-effective manner.

2.2. The coherence function as weighting factor

The coherence function, which is a complex function of frequency, is defined as follows,

$$C_{xy}(f,m) = \frac{S_{xy}(f,m)}{\sqrt{S_{xx}(f,m)S_{yy}(f,m)}}$$

where $S_{xx}(f,m)$ and $S_{yy}(f,m)$ are the power spectral densities of signals x and y for the speech frame m respectively, and $S_{xy}(f,m)$ is the cross-spectral density between x and y for the speech frame m. We estimated these spectral densities using a time averaging by a simple first order low-pass filter, e.g.,

$$S_{xy}(f,m) = \lambda S_{xy}(f,m-1) + (1-\lambda)X(f,m)Y^*(f,m)$$

where λ is a forgetting factor, and $X(f,m)$ and $Y(f,m)$ represent the FFT spectra of the two channels for speech frame m.

In [4] a classical approach for estimating the incoherent or reverberant part of the signal energy using the coherence function is proposed. This approach was also used in a noise reduction algorithm for binaural hearing aids by Peissig [5]. Our related algorithm works with the magnitude squared coherence (MSC), Γ_{xy}:

$$\Gamma_{xy}(f,m) = \frac{|S_{xy}(f,m)|^2}{S_{xx}(f,m)S_{yy}(f,m)}$$

A 3-point median filter is applied to this MSC to remove spurious values and smooth the MSC sequence.

$\Gamma_{xy}(f,m)$ is used to weight the sum of the power spectra $|X(f,m)|^2$ and $|Y(f,m)|^2$ resulting in a single spectral representation which is then mel-filtered and used to compute cepstral coefficients. The coherence weighting is meant to turn off uncorrelated signals and pass correlated signals.

3. Single-channel signal processing techniques

The ideal case for two-channel processing is likely to be when the noises at the two microphones are totally uncorrelated. This is only true in the ideal case of an incoherent noise field. In practice, for a real noise field, the cross-power spectrum between the noises is never exactly zero. The noises (and reverberant speech echoes) at two microphones are likely to become less correlated if the distance between the microphones is increased; but in many practical applications this distance is constrained.

We investigated the combination of the two-channel methods with two single-channel signal processing methods that do not exploit the uncorrelatedness of noise and seemed likely to be complementary.

3.1. Noise reduction by spectral subtraction

Basic noise reduction algorithms are an easy and effective way to reduce mismatch between noisy test conditions and clean training conditions. We used the generalized form of spectral subtraction defined in [6] as,

$$D(f,m) = |Y(f,m)|^{2\gamma} - \alpha_{SS}|\hat{N}(f,m)|^{2\gamma}$$

$$|\hat{X}(f,m)|^2 = \max([D(f,m)]^{\frac{1}{\gamma}}, \beta|\hat{N}(f,m)|^2)$$

where $|\hat{X}(f,m)|^2$ is an estimate of the clean speech power spectrum at frame m, $|\hat{N}(f,m)|^2$ is an estimate of the noise power spectrum at frame m, $|Y(f,m)|^2$ is the observed noisy speech power spectrum at frame m, α_{SS} is an over-subtraction factor, and β represents a noise flooring factor.

From [6][7] we chose the parameter combination $\beta = 0.1$, $\alpha_{SS} = 4.1$ and $\gamma = 1$, for our recognition experiments.

The choice of an estimator for background noise is a key problem for noise reduction algorithms. Generally, noise is assumed to be additive and stationary with respect to speech. We use a a simple method of adaptive noise estimation to get a continuous noise estimate in order to avoid the use of a VAD (Voice Activity Detector) [8]. The noise estimate is continually updated but is allowed to increase much more slowly than it is allowed to decrease. Thus the noise estimate will increase only slowly during speech intervals and collapse quickly back during speech gaps. The noise estimate update (omitting the slow increase / fast decrease control) is as follows:

$$\hat{N}(f,m) = \rho\hat{N}(f,m-1) + (1-\rho)Y(f,m)$$

where $\hat{N}(f,m)$ is the short-time noise spectrum estimate at frame m, and $Y(f,m)$ is the short-time spectrum of the noisy signal at frame m.

3.2. Cepstral mean normalization

To compensate for the effects of a communication channel (room acoustics, microphone response, or transmission channel) many proposals have been made. Cepstral Mean Normalization (CMN) is a very simple but efficient method to compensate for the so-called convolutive noise that arises from channel distortions. It has found widespread use in many systems. In earlier work with our PZM speech data [9][10], we used the "long-term log spectral subtraction" (LTLSS) method to compensate for the communications channel.

Here we use an online implementation (without delay) of CMN [11] because we decided to process each utterance independently of the others, and the LTLSS method is inappropriate for the independent processing of single utterances. Thus, the cepstral means are estimated using a weighted sum of the current feature vector components and the previous estimate,

$$m_c(t) = (1-\alpha)m_c(t-1) + \alpha x_c(t)$$

where $m_c(t)$ is the cepstral mean vector at t-th frame, $x_c(t)$ are the original cepstral coefficients at frame t, and $m_c(t-1)$ is the cepstral mean vector at frame (t-1).

Having updated the means, the normalized cepstral coefficients are obtained by subtraction of the means.

4. Automatic speech recognition experiments

4.1. Test corpus

In order to provide a simple task, and to isolate acoustic issues from other effects, the Meeting Recorder Project at ICSI has recorded connected digits. Each digit string contains from one to ten digits ("zero" and "oh" are both possible as digits). There are a total of 7604 digits in 2323 digit strings. These recordings were made simultaneously with close-talking mics, with table-mounted PZM mics and with the two cheap ELC mics on the

mock-up PDA. We have used these test data in all the experiments.

The digit recordings were made in recording sessions held before or after group meetings, with the talkers seated in the same positions around the table as in the meetings.

We measured the 60 dB reverberation time of the room as about 0.25 s (averaged reverberation time over the octaves from 125 to 4000 Hz). The C50 clarity (the ratio of the sound energy arriving in the first 50 ms to the sound energy arriving later) was measured at 26 and 16 dB for speaker locations respectively about 2 and 3 feet from the recording microphone. The measurements were made by estimating room impulse responses with maximum-length sequences, and then calculating reverberation time and C50 from the impulse responses, using Schroeder's integrated-impulse method for reverberation time [12].

4.2. Training corpus

The training set consisted of 4220 male and 4220 female digit string utterances from the TIDigits training set, downsampled to 8000 Hz. This file set was the same as that used to obtain the Aurora [13] clean training set, except that we omitted the G.712 telephone bandwidth filtering.

Because we did not have two-channel training data, we treated the single-channel training data as two identical channels when using the two-channel front ends.

4.3. ASR system

We used the Aurora reference recognizer system described in [13]. This system uses the HTK recognition toolkit and is based on Gaussian mixture HMMs. The digits are modeled as whole word HMMs, and two pause models are defined: one for modeling pauses before and after the utterance and the other for modeling pauses between words. This system has a 1% word error rate when training and testing on TIDIGITS.

The front-end of the reference system calculates 39 features: 12 mel-frequency cepstral coefficients, log frame energy, and first- and second-order delta features. The two-channel front ends replace the reference front end and calculate a similar set of 39 features, with cepstral coefficients calculated using the two-channel processing described in Section 2, and the log frame energy set to the average of the log frame energies of the two channels.

There was considerable low-frequency noise present in the PZM and, especially, ELC signals, so a high-pass filter with a 50 Hz cutoff frequency was applied to all waveform data as the very first stage of processing. After the filtering, the NIST "stnr" tool [14] reported a 11 dB average SNR for the PZM recordings and a 9 dB average SNR for the ELC recordings.

5. Experimental results

In this section we present a range of experiments that evaluate the recognition performance of the cheap ELC microphones. As a baseline experiment we compare their individual recognition performance with that of the high quality PZM mic. Table 1 presents the word error rates obtained in the recognition tests on the three mics using the front-end MFCC reference system. The row "Ener=yes" means that the log frame energy is included in the feature vector, and the row "Ener=no" means that only the delta and delta-delta log frame energy are included. The "Ener=no" case was tried because of a volume level mismatch between the training data and the testing data. We will only present the "Ener=no" case from here on. For the experiments shown in Tables 4 and 5, we tried the "Ener=yes" case with mean normalization of the cepstral and frame energy features and found that this greatly reduced the gap between "Ener=no" and "Ener=yes" cases, but "Ener=yes" still did not perform better than "Ener=no".

Table 2 presents the word error rates obtained in the recognition tests on the PZM mic using the front-end MFCC reference system and table 3 illustrates the recognition results on both ELC mics. The column "baseline" corresponds to no robust speech processing; the column "CMN" corresponds to the CMN processing; the column "NR" corresponds to the additive noise reduction processing; and the column "NR+CMN" corresponds to the noise reduction processing followed by the online cepstral mean normalization processing.

Several observations can be drawn from these initial experiments. The first one is that both noise reduction and cepstral mean normalization reduced the error rates. The improvement achieved by the additive noise reduction technique is greater than the one achieved by the cepstral mean normalization processing. Perhaps the CMN is doing a worse job of taking out convolutive noise than the noise reduction is doing of taking out additive noise. Another observation is that these processing techniques seem not to have a cumulative effect when used together.

If we compare the recognition results for the two ELC mics, we can see that one of the mics (MIC-1) outperforms the other one (MIC-2) in all cases. Now comparing the results obtained by both the PZM and ELC mics, we can see that the error rate of ELC MIC-1 (12.6%), after noise reduction and cepstral mean normalization, compares well with that of the PZM mic (11.3%), while the ELC MIC-2 error rate is still relatively poor (16.6%).

Ener	PZM	ELC-1	ELC-2
yes	25.3	47.5	56.3
no	24.1	30.1	36.4

Table 1: WER using different quality microphones.

	PZM-MIC			
Ener	baseline	CMN	NR	NR+CMN
no	24.1	22.9	11.2	11.3

Table 2: WER using good quality PZM microphone.

ELC-MIC-1				
Ener	baseline	CMN	NR	NR+CMN
no	30.1	29.5	12.9	12.6
ELC-MIC-2				
Ener	baseline	CMN	NR	NR+CMN
no	36.4	36.2	16.5	16.6

Table 3: WER using the cheap ELC microphones independently.

In a second set of recognition experiments we have used both ELC mics signals jointly in order to test the performance of the feature extraction techniques described in section 2. Tables 4 and 5 detail the recognition performance in terms of word error rate (WER) for the delay-and-sum and coherence weighting approaches, respectively. As observed for the single-mic results, the additive noise reduction technique (which was applied to each channel separately prior to the two-channel processing) exhibits a very good performance. Similarly, the improvement

given by the CMN technique is not as large as the improvement given by the noise reduction.

The two-channel feature-extraction methods show themselves to be valuable, since they slightly outperform the better-performing mic (ELC-1) and greatly outperform the worse-performing mic (ELC-2). The two-channel methods result in quite similar error rates; the best error rate (11.8%) is given by the coherence technique.

	MIC-1 + MIC-2			
Ener	baseline	CMN	NR	NR+CMN
no	27.5	25.5	12.3	12.3

Table 4: WER using delay-and-sum processing.

	MIC-1 + MIC-2			
Ener	baseline	CMN	NR	NR+CMN
no	26.4	23.3	11.8	11.8

Table 5: WER using coherence weighting.

In order to perform significance testing on our results, we produced Table 6, which compares the hypotheses produced for the 2323 test sentences between systems: the PZM mic, the ELC mics ELC-1 and ELC-2 used independently, the coherence weighting processing using both ELC mics (COHW), and the delay-and-sum processing using both ELC mics (DS).

Noise reduction and CMN. No Log-Energy					
Compare	same	diff	better	worse	signif
(ELC-1,ELC-2)	1876	447	330	117	7×10^{-25}
(COHW,ELC-1)	2034	289	179	110	3×10^{-5}
(COHW,DS)	2142	181	106	75	0.013
(PZM,ELC-1)	1805	518	285	233	0.012
(PZM,COHW)	1806	517	260	257	0.46

Table 6: This table compares the sentence hypotheses produced by systems (a,b). The column "same" counts the times that the two systems produced the same sentence hypothesis, the column "diff" count the times the hypotheses were different, the column "better" counts the times that system "a" had a better hypothesis and the column "worse" counts the times that the system "a" has a worse hypothesis. Column "signif" gives the approximated probability that the higher performance of system "a" is random chance, assuming a binomial distribution with each differing sentence seen as an independent trial.

From the table, we see that the inferiority of the ELC-2 mic to the ELC-1 mic is strongly significant, as is the superiority of the COHW combination to the ELC-1 mic used alone. The difference between the COHW and DS combinations and the difference between the PZM mic and the ELC-1 mic are less strongly significant. The difference between the PZM mic and the two ELC mics in the COHW combination is not significant.

6. Conclusions and future work

For the task explored here, it appears that we can achieve similar results with both high and low-quality microphones so long as we can use more than one microphone and at least one microphone performs well. This could have practical significance for the case of mass-produced portable devices for which extensive quality control is impractical.

This experiment is suggestive of such a conclusion. We will focus our research efforts to improve the effectiveness of the proposed techniques and to validate them. For instance, it would be desirable to repeat the experiment for a larger number of inexpensive microphones so that we could derive a distribution of performance. Also, even the best performance we observed is much poorer than the results for the near-mic case (which for these tests was in the vicinity of 3%–4%). Clearly much more work is required to provide low error rates for far-field microphones in cases where large arrays are not practical. This is a major area of further research for us.

7. References

[1] "The ICSI Meeting Recorder Project," http://www.icsi.berkeley.edu/ Speech/mr/.

[2] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal of the Acoustical Society of America*, vol. 78, pp. 1508–1518, November 1985.

[3] R. Moddemeijer, "An information theoretical delay estimator," in *Ninth Symposium on Information Theory in the Benelux*, Enschede (NL), 1988, pp. 121–128, Werkgemeenschap Informatie- en Communicatietheorie.

[4] J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone signal processing technique to remove room reverberation from speech signals," *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[5] B. Kollmeier, J. Peissig, and V. Hohmann, "Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain," *Scand. Audiol. Suppl.*, vol. 38, pp. 28–38, 1993.

[6] M. Berouti, B. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, 1979, pp. 208–211.

[7] P. Lockwood, C. Baillargeat, J.M. Gillot, J. Boudy, and G. Faucon, "Noise reduction for speech enhancement in cars: Non-linear spectral subtraction / kalman filtering," in *Proc. EUROSPEECH*, 1991, vol. 1, pp. 83–86.

[8] L. Arslan, A. McCree, and V. Viswanathan, "New methods for adaptive noise suppression," in *Proc. ICASSP*, Detroit, USA, 1995, pp. 812–815.

[9] D. Gelbart and N. Morgan, "Double the Trouble: Handling Noise and Reverberation in Far-Field Automatic Speech Recognition," in *Proc. ICSLP*, Denver, USA, 2002.

[10] D. Gelbart, "Mean subtraction for automatic speech recognition in reverberation," M.S. thesis, Univ. of California Berkeley, 2003.

[11] L. Docio-Fernandez, *Aportaciones a la mejora de los sistemas de reconocimiento*, Ph.D. thesis, University of Vigo (Spain), 2001.

[12] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *Journal of the Acoustical Society of America*, vol. 66, no. 2, August 1979.

[13] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," in *ISCA ITRW ASR2000*, Paris, France, 2000.

[14] "SPQA Version 2.3," http://www.nist.gov/speech/tools/.