METHODS FOR MICROPHONE EQUALIZATION IN SPEECH RECOGNITION

L. Fissore, G. Micca and C. Vair

CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, Italy
E-Mail fissore/micca/vair@cselt.stet.it

ABSTRACT

This paper presents a review of current research carried on at various laboratories aiming to increase the robustness of speech recognition systems to channel and microphone variations. A comparative analysis of several techniques, used in recent studies on microphone-independence, are discussed and compared: these include Cepstral High-Pass Filtering, Cepstral-Mean Normalization, Ratz algorithm and Bayesian learning. Also, some results obtained at CSELT labs using the methods above mentioned are reported, specifically addressing the issue of robustness of ASR systems to microphone variations.

1. INTRODUCTION

Interactive voice systems with automatic speech recognition (ASR) capabilities are often deployed in difficult and dynamically varying acoustical environments; therefore robust recognition paradigms are required in real applications (e.g. over telephones, in cars or outdoors) to maintain an acceptable level of recognition accuracy even in adverse conditions [1, 15, 14]. A limitation of today's ASR technology is the severe performance degradation when the acoustical environment of the training phase is substantially different from that of the operating phase. This paper addresses the problem of heavy performance degradation observed when the used microphone differs from the one on which the recognizer was trained. The adoption of various different microphones (often low-cost and low-quality ones) is requested by some applications (e.g. running on PC/workstations) but it is detrimental to recognition performance. In the last few years, a great deal of interest has been devoted to equalization issues concerning the change of microphone and a large number of methods has been proposed; however further improvements are still required, mainly for the case of very low data available from the operating environment. Methods have been grouped in three classes:

- **spectral equalization** techniques applied in a preprocessing step (e.g. cepstral normalization and filtering);
- **feature compensation** techniques based on some type of mapping, possibly conditioned to a specific variable (SNR, codeword, etc...); these in turn are subdivided in "stereo" techniques which require bi-channel speech input, and "blind" techniques which compute an approximation of the spectral tilt by means of optimization algorithms (CDCN [2], RATZ [3], VTS [4], VPS [5]);
- **model compensation** based on some iterative or incremental procedure to modify and adapt the stochastic structure of the statistical models (Bayesian adapting learning [6] and MLLR [7]).

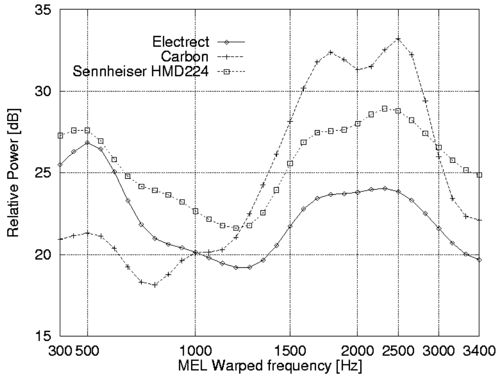

Figure 1: Spectral behaviour of 3 different microphones

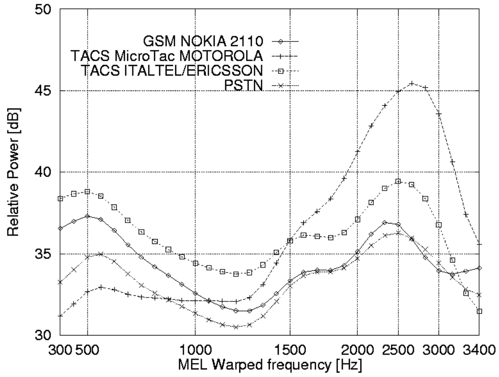

Figure 2: Spectral behaviour of 4 different microphones

2. GENERAL CHARACTERISTICS

The choice of the appropriate microphone is conditioned by type and constraints of the application, by the background noise conditions and by the cost of the overall system. The sensitivity to the noise level and to the reverberation depends on the way the speech signal is acquired. Microphones have different transductional and directional characteristics which alter the speech signal in different ways. The positioning of the microphone also causes distortion. Generally, if the distance between the microphone and the speaker's mouth increases, more environmental noise is acquired.

For some applications (e.g. in office environments) the speech recognizers are usually trained and tested by using a high-quality, head-mounted, close-talking, noise canceling microphone. On the other side, telephone applications

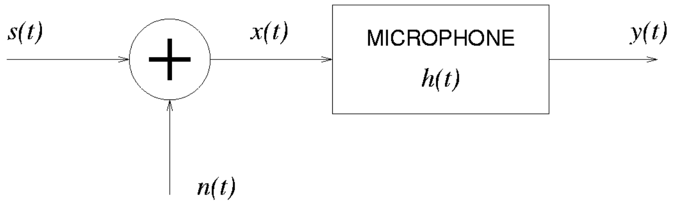

Figure 3: Model of the acquisition chain

rely upon speaker phones, cellular phones and ordinary telephones owned by the users and characterized by different microphones. The spectral characteristics of speech provided to the recognizer and collected through different microphones and communication channels are shown in Fig. 1 and Fig. 2.

Each plot represents the average spectrum and is computed as the DFT of the average cepstral vector. This representation includes the spectral modeling of the microphone transfer function, the preemphasis filtering of the front-end and the average vocal tract characteristics of the pool of speakers. The average cepstrum has been obtained from a large corpus of utterances pronounced by a few hundreds of subjects. The logarithmic frequency scale reflects the MEL-based band grouping. The spectral behaviours of two telephone microphones, one of carbon type and the other one of electret type, and of a high quality, head-mounted microphone (Sennheiser HMD224) are compared in Fig.1. The plots of the electret and HMD224 mics are quite similar, while the carbon mic plot shows an enhanced gain above 1.5 Khz. Fig. 2 displays the comparison among speech data collected on different networks, fixed (PSTN) and mobile (TACS and GSM); besides, two different TACS classes are represented, corresponding to two different handsets. The plots of the NOXIA 2110 GSM, PSTN and ITALTEL/ERICSSON TACS terminals are similar and are at less than 5 dBs from one another. The MOTOROLA MicroTac TACS terminal shows a high peak at around 2.5 KHz. The spectral differences justify the degradation of recognition performance in unmatched conditions.

3. CHANNEL MODEL

Fig. 3 shows a simplified model of the acquisition chain. The input speech signal $s(t)$ is merged with an additive ambient noise $n(t)$. The resulting signal $x(t)$ is then filtered by the microphone transfer function $h(t)$ to produce $y(t)$ at the recognition system input. In the time domain:

$$y(t) = [s(t) + n(t)] * h(t) \tag{1}$$

where $*$ is the convolution operator. If the $n(t)$ term is neglected, the equation (1) can be rewritten:

$$y(t) = s(t) * h(t) \tag{2}$$

The convolutional filtering performed by the transfer function of the microphone in the time domain in (2) becomes a sum in the cepstral trajectories domain:

$$C_y[q,i] = C_x[q,i] + C_H[q,i] \tag{3}$$

where q is the quefrency index and i is the frame. The equation (3) shows how the microphone transfer function acts as an additive term in the cepstral domain. Moreover the microphone characteristics can be regarded as time invariant: its contribution is then steady in time:

$$C_y[q,i] = C_x[q,i] + C_H[q]. \tag{4}$$

4. SPECTRAL EQUALIZATION

Among the techniques developed to increase microphone robustness, the most common and simplest ones are the preprocessing techniques (Cepstral Mean Normalization and High Pass Filtering of cepstral coefficients). These techniques are applied at the front-end level before the pattern matching stage in an unsupervised mode; they are useful to partially remove slowly varying components in the feature space when there are distortions due to a change of microphone [8]. These methods are effective in compensating for the effects of unknown linear filtering, but they are not able to handle distortions due to the presence of additive noise.

4.1. High Pass Filter (HPF)

This technique acts on cepstral trajectories, wiping out the low frequency spectral components; these represent the long term characteristics of speaker and microphone which are not informative for speech recognition purpose. The filtering is carried out by a first order IIR filter [9], acting on cepstral trajectories $C_y[q,i]$:

$$H(z) = \frac{1 - z^{-1}}{1 - (1-\lambda)z^{-1}} \tag{5}$$

$$C_z[q,i] = C_y[q,i] + C_y[q,i-1] + (1-\lambda)C_z[q,i-1] \tag{6}$$

4.2. Cepstral Mean Normalization (CMN)

The aim of Cepstral Mean Normalization, like HPF technique, is to wipe out the bias due to the time invariant microphone characteristics. If it is supposed the non-coherency of the input signal, the channel cepstrum $C_H[q]$ can be directly derived from (4) :

$$< C_y[q] >=< C_x[q] > + C_H[q] \simeq C_H[q] \tag{7}$$

where $< C[q] >= \frac{1}{T}\sum_{i=1}^{T} C[q,i]$ is the average cepstrum at the quefrency q. A large variety of techniques has been proposed to estimate the spectral bias from the input data; the bias vector can be estimated as the average of the feature frames over the whole utterance or can be conditioned to the vector quantization codeword, phone label or HMM state. In order to make the CMN suitable for real-time applications, the computation of the spectral bias is performed through short-term averages over a moving window superimposed to the input signal.

5. FEATURE COMPENSATION

Researchers at CMU have developed a large number of techniques able to cope with the joint effects of additive noise and channel distortions. All these methods estimate an additive correction to be applied in the cepstral domain. These techniques can be grouped according to their requirements of knowledge about the testing environment (stereo or no-stereo data), the dependency between compensation vectors and measurements (SNR, codevector, phoneme identity) and their adaptation capability. Among the initial approaches proposed by CMU, the most known is the Codebook-Dependent Cepstral-Normalization (CDCN) algorithm, which applies the Maximum-Likelihood estimation to determine parameters representing additive noise and spectral-tilt and uses a Minimum Mean Square Error estimator to minimize VQ codeword distances.

More recently, CMU proposed a multi-variate Gaussian-based Cepstral Normalization technique (RATZ), which uses an optimal estimation procedure.

The RATZ algorithm [3] (named BRATZ in blind conditions not requiring stereo data) performs a Gaussian mixture modeling of the universal [M] acoustic space by means of the Maximum Likelihood criterion; given an input frame vector $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_T]^t$ we can write:

$$p_M(\mathbf{X}) = \sum_{k=1}^{K} P_M[k] \mathcal{N}(\mathbf{X} \mid \mu_{k,M}, \mathbf{\Sigma}_{k,M}) \qquad (8)$$

where $P_M[k]$ is the *a priori* probability and $\mu_{k,M}$,$\mathbf{\Sigma}_{k,M}$ are the mean vector and covariance matrix of each multivariate Gaussian mixture k. The algorithm models the change of environment as an additive offset for each mean and covariance matrix:

$$\mu_{k,T} = \mu_{k,M} + \mathbf{r}_k \qquad \mathbf{\Sigma}_{k,T} = \mathbf{\Sigma}_{k,M} + \mathbf{R}_k \qquad (9)$$

In blind condition the estimation of $\mathbf{r}_k$ and $\mathbf{R}_k$ is done by the EM technique [3]. The current frame $\mathbf{x}_{i,T}$ is then mapped from the actual acoustic space back to the "universal" acoustic space $\hat{\mathbf{x}}_{i,M}$ according to a Maximum Mean Square Error (MMSE) distance:

$$\hat{\mathbf{x}}_{i,M} \simeq \mathbf{x}_{i,T} - \sum_{k=1}^{K} P(k|\mathbf{x}_{i,T})\, \mathbf{r}_k \qquad (10)$$

Recently, the CMU speech group proposed new procedures, named Vector Taylor Series (VTS) [4] and Vector Polynomial Expansion (VPS) [5] based on series approximations of the non-linear environment function. CMU showed that the new model-based algorithms, tested on the ARPA 5000-word WSJ and on the Census corpora outperform CDCN and RATZ at all SNRs.

6. MODEL COMPENSATION

6.1. Bayesian learning

Bayesian learning - also known as MAP (Maximum A Posteriori) learning - allows to adapt HMM parameters to a new acoustic environment. The original formulation [10] provided a mathematical framework to optimally combine training data with a priori knowledge about the stochastic distributions of the models. MAP has been extensively experimented as a technique for adaptation [10, 6, 11, 12]; its fair property is that of being asymptotically equivalent to the ML algorithm as the amount of adaptation data increases. MAP is usually implemented as an off-line, supervised procedure; recently an on-line, incremental and unsupervised version has been proposed [11], with joint reestimation of the models parameters and a-priori parameters [16]. The accurate estimation of the parameters of the acoustic-phonetic units provided by the MAP algorithm is traded-off by a slowing down of the convergence to the target acoustic environment. MAP can be profitably coupled to other techniques to improve the compensation effect. Specifically:

- the **exponential forgetting** mechanism [16] to speed up the convergence to the working environment;
- the combination of **feature compensation** or of **signal equalization** techniques to MAP learning, to improve the efficiency of the adaptation process [13, 12];
- the use of explicit **correlation constraints** among model parameters: this class includes the tying technique, the adoption of hierarchic or tree-based structures [17], the modelisation of the correlation itself in the MAP framework [18];
- **Vector Field Smoothing** techniques [19] for interpolation of unobserved parameters.

Bayesian learning is particularly effective in cases where both the acoustic impairment must be compensated and speaker adaptation is to be performed. At CSELT we focussed on an on-line, unsupervised version of the algorithm, combined with some simple compensation of the acoustic space variations [12]. Tab. 1 presents the results obtained by combining Cepstral Mean Normalization and MAP adaptation in a given microphone equalization task in speaker-dependent mode.

7. EXPERIMENTAL RESULTS

A fair and thorough evaluation of the most important microphone equalization methods has not yet been reported in the literature, due to the variety of conditions in which these techniques have been experimented. A valuable comparison of some techniques for ARPA 5000-word WSJ microphone compensation task appears in [14].

The following mismatch condition was experimented at CSELT to assess different compensation techniques: the recognizer was trained by means of a wide-band, head-mounted microphone (Senheiser HMD 241) [M], while the test was performed with a telephone microphone [T] connected to a PABX. We used a down-sampled 8KHz version of source 16Khz sampled signal for both telephone and microphone.

The speech database BI_MICRO consisted of 1600 continuous utterances recorded by each one of 4 speakers in a quiet room, in stereophonic mode from the two ([M] and [T]) microphones. A set of 1400 utterances was used to train HMMs, while the test set consisted of the remaining 200 sentences. Out of the 1400 training utterances, 1120 belonged to a phonetically balanced corpus, the other 280 plus the test utterances were related to a railway timetable enquiry domain. The test vocabulary was made up of 247 words and no grammar was used during recognition. The average utterance duration was 2.84 sec.

Fig. 4 shows the recognition performance comparison as a function of a different number of adaptation utterances for the recognition task [M:T] (training with microphone [M] and test with telephone [T]) in term of R_{WA}[1].

The BRATZ technique was the most effective when few seconds of speech were used, while the MAP algorithm combined with CMN was the best one with a larger amount of adaptation data.

HPF required a minimum amount of computational load and still gave 35% of error rate reduction. Tab. 1 summarizes the recognition results obtained using 100 adaptation utterances, corresponding to about 5 minutes of speech. It

[1] R_{WA} represents the relative error rate reduction obtained by the algorithm.

Algorithm	CDHMM	
	W.A.	$R_{WA}{}^1$
Baseline[M:M]	86.0%	100%
Mismatch[M:T]	75.7%	0%
HPF	79.5%	36.9%
CMN	82.3%	64.1%
BRATZ	83.0%	70.9%
MAP	80.4%	45.6%
CMN-MAP	83.1%	71.8%

Table 1: Performance comparison of each adaptation algorithm with 100 adaptation utterances

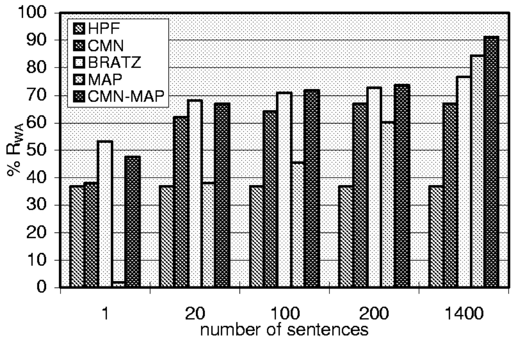

Figure 4: Relative error rate reduction for tested compensation techniques versus the number of adaptation sentences

should be noted that the adaptation for BRATZ, MAP and CMN-MAP was off-line computed due to the large amount of speech data needed for its estimation. On the other hand the computational load for on-line adaptation is minimum for CMN, MAP and CMN-MAP and equivalent to the multivariate Gaussian functions computation (10) for the BRATZ algorithm.

8. CONCLUSIONS

A variety of techniques for microphone equalization in ASR systems has been presented and evaluated. Relationships among amount of available data in the working environment and recognition performance have been shown. Current research focuses on reducing the adaptation time by procedures not involving stereo data. It appears that successful implementations of compensation methods will require a combination of different techniques, and will exploit the constraints of the specific application. For instance, IVR applications with long dialogue sessions should incrementally and jointly adapt to microphone, environment noise and speaker's vocal characteristics, in a mode totally transparent to the user; this methodology is best suitable for multimedia devices, open to a pool of users, where even longer sessions can be allowed. Further research is still required for effective quasi-instantaneous compensations in cases where only few words are available from the speaker.

9. REFERENCES

[1] S. Furui "Recent advances in robust speech recognition", in Proc. of ESCA Workshop on "Robust Speech Recognition for Unknown Communication Channels", Pont-a-Mousson, France, April 1997

[2] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition", Kluwer Academic Publishers, 1992

[3] R.M. Stern, P.J. Moreno, B.Raj, "A unified approach for robust speech recognition", in Proc. of Eurospeech, Madrid, 1995

[4] P.J. Moreno, B.Raj and R.M. Stern, "A Vector Taylor Series Approach for Environment Independent Speech Recognition", in ICASSP-96, pp. 733-736, Atlanta, May 1996

[5] B.Raj, E.Gouvea, P.J. Moreno, and R.M. Stern, "Cepstral Compensation by polinomial approximation for Environment-Independent Speech Recognition", in Proc. of ICSLP, pp. 2340-2343, Philadelphia, October 1996

[6] J.-L. Gauvain, C.-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains" in IEEE Trans. on Speech and Audio Processing, Vol.2, N.2, pp. 291-298, April 1994

[7] C.J. Leggetter, P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol. 9, pp. 171-185, 1995

[8] J. Chang, V. Zue, "A study of speech recognition system robustness to microphone variations. Experiments in phonetic classification", in ICSLP '94, pp. 995-998

[9] H.Hermansky, N. Morgan, A. Bayya and P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)", in Proc. of Eurospeech, pp. 1367-1370, 1991

[10] C.-H. Lee, C.-H. Lin, B.-H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models", in IEEE Trans. on ASSP, Vol.ASSP-39, N.4, pp. 806-814, April 1991

[11] Y. Zhao, "Self-learning Speaker and Channel Adaptation Based on Spectral Variation Source Decomposition", in Speech Communication, N.18, pp. 65-77, 1996

[12] C. Vair, N. Chiminelli, L. Fissore, G. Micca, "Comparison of Algorithms for Microphone Equalization in Continuous Speech Recognition", in Proc. of ESCA Workshop on "Robust Speech Recognition for Unknown Communication Channels", Pont-a-Mousson, France, April 1997

[13] C.-H. Lee, "On feature and model compensation approach to robust speech recognition" in Proc. of ESCA Workshop on "Robust Speech Recognition for Unknown Communication Channels", Pont-a-Mousson, France, April 1997

[14] R.M. Stern, B. Raj, P.Moreno, "Compensation for environmental degradation in automatic speech recognition", in Proc. of ESCA Workshop on "Robust Speech Recognition for Unknown Communication Channels", Pont-a-Mousson, France, April 1997

[15] J.C. Junqua, J.P. Haton, "Robustness in Automatic Speech Recognition", Kluwer Academic Publishers, 1996

[16] Q. Huo, C.-H. Lee, "A Study of On-Line Quasi-Bayes Adaptation for CDHMM-Based Speech Recognition", in Proc. of ICASSP-96, pp. 705-708, Atlanta, May 1996

[17] G. Zavaliagkos, R. Schwartz, J. Makhoul, "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition", in Proc. of ICASSP-95, pp. 676-679, Detroit, May 1995

[18] Q. Huo, C.-H. Lee, "On-Line Adaptive Learning of the Correlated Continuous Density Hidden Markov Models for Speech Recognition", in Proc. of ICSLP, pp. 985-988, Philadelphia, October 1996

[19] J. Takahashi, S. Sagayama, "Vector-Field-Smoothed Bayesian Learning for Incremental Speaker Adaptation", in Proc. of ICASSP-95, pp. 696-699, Detroit, May 1995