FEATURE COMPENSATION WITH SECONDARY SENSOR MEASUREMENTS FOR ROBUST SPEECH RECOGNITION

Bhiksha Raj[1], *Rita Singh*[2]
1. Mitsubishi Electric Research Labs, Cambridge, MA, USA
2. Haikya Corp., Boston, MA, USA

ABSTRACT

This paper investigates the use of secondary sensor measurements to augment feature compensation methods for robust speech recognition. Secondary sensors measure secondary phenomena associated with human speech production. While such measurements do not provide sufficient information for speech recognition per-se, they do not degrade with the noise that corrupts the acoustic signal and can be used to guide algorithms that attempt to estimate noise compensation algorithms by restricting the region of the acoustic space within which the recorded speech must lie. In this paper we specifically, we investigate the use of measurements obtained from a Glottal ElectroMagnetic Sensor (GEMS) to improve the noise estimation performance of the Vector Taylor Series algorithm. We and show that this can result in significant improvement in performance of the VTS algorithm, and, consequently, recognition performance.

1. INTRODUCTION

It is well known that the recognition performance of automatic speech recognition (ASR) systems degrades in the presence of interfering noises. The reason for this is well known: speech recognition systems attempt to perform Bayesian classification. For this to work, the distribution of the feature vectors of incoming speech must match the distributions of the various sound classes known to the recognizer. Noise transforms the distribution of the incoming speech so that they no longer match the distributions within the recognizer, resulting in degraded recognition.

Noise compensation algorithms attempt to compensate for the effect of the noise by transforming either the incoming data (either the speech signal, or the features computed from them) or the parameters of the recognizer itself [1, 2], such that after the transformation the distribution of the data match those stored in the recognizer. In this paper we focus on data-compensation algorithms - algorithms that modify the incoming data - since they do not require access to the internal data structures of a recognizer; however the underlying principle carries over to other techniques as well.

The problem, roughly stated is as follows: Noisy recordings comprise speech that has been drawn from the distribution of clean speech and transformed in some manner. The distribution of the noisy speech is hence a transformed version of the region of the clean speech distribution that the speech samples have been drawn from, as illustrated by Figure 1a. An ideal denoising transformation transforms the noisy signal, such that the distribution of the transformed noisy speech exactly matches the regions of the clean speech distribution that the speech was drawn from.

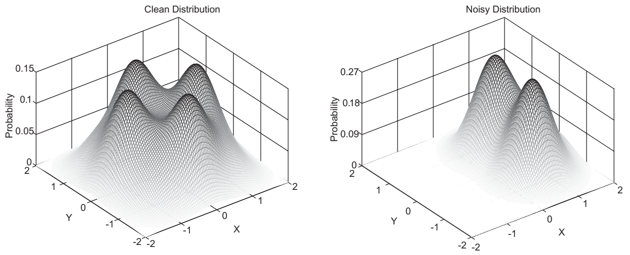

Figure 1a: Illustration of the relation of the distribution of noisy speech to that of clean speech through a hypothetical example. Left panel: Distribution of a hypothetical 2-dimensional feature representation of clean speech. Right panel: Distribution of features derived from a short recording of noisy speech. It represents data that have been drawn from only a part of the overall speech distribution and

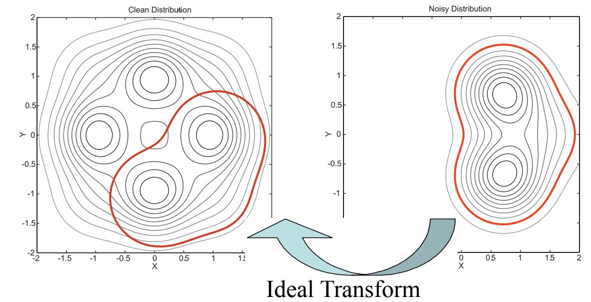

Figure 1b: Left panel: Contour plot of the clean speech distribution in Figure 1a. Right panel: Contour plot of the distribution of noisy speech vectors. The noisy vectors have been drawn from the region of the clean speech distribution outlined by the thick contour and transformed by noise. The ideal noise-compensating transformation would modify the contours of the noisy distribution to match the contours of the corresponding region of the clean speech distribution.

This is illustrated by Figure 1b. The goal of data-compensation algorithms is to determine this ideal denoising transformation and apply it to the incoming data.

Unfortunately the estimation of the ideal transformation is extremely difficult. The actual form of the transformation is not known, and must be assumed. Even if the true form of the transformation were known, the values of its parameters can be ambiguous: there are often multiple ways of transforming the noisy distribution to match the clean one, most of which are incorrect. This is illustrated by Figure 2a. Some of the incorrect solutions often result in a better statistical fit of the transformed noisy distribution to the clean distribution than the correct solution itself. In the absence of external supervision, the estimation procedure for any data-compensation algorithm might return any of the these possible solutions, resulting in suboptimal denoising. This problem is compounded by the fact that the form of the transformation is itself unknown and must be assumed.

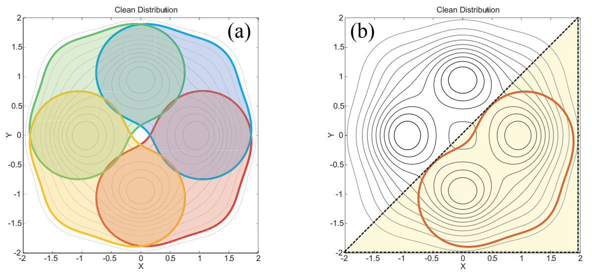

Figure 2: (a) The thick curves represent four possible contours, all of which could be matched against the noisy contour in Figure 1b by an unsupervised noise-compensation algorithm. (b) Even coarse information from a secondary sensor could localize the regions that the noisy speech vectors are drawn from, *e.g.* to the shaded triangle. This constraint is sufficient to identify the correct denoising transform.

The ambiguousness of the correct transformation can be greatly reduced through supervisory information that localizes the speech vectors in the acoustic space to guide the estimation algorithm. Such supervisory information can be obtained through secondary sensors that measure secondary phenomena related to the speech generation process, e.g. cameras that capture lip movements, bone sensors that capture bone vibrations associated with speech generation and Glottal Electromagnetic Sensors (GEMS) [3]. An important feature of such sensors is that their measurements are not corrupted by the noisy environment that corrupts the acoustic recordings of the speech signal. Thus, although they may not be sufficient for speech recognition per-se, they reliably localize the region of the acoustic space that the speech data have been drawn from. This information can be used to guide the estimation of denoising transformations, as illustrated by Figure 2b.

The use of secondary sensors for speech denoising has previously been explored by Hershey *et. al.* [4], who use bone sensor measurements to guide the denoising process. In addition to secondary sensor measurements, their procedure requires *a priori* knowledge of the distribution of the feature vectors noise, although this distribution may be updated during the estimation itself. Demiroglu *et. al.* [5] use GEMS sensor data to identify and delete noise corrupted frames of speech prior to recognition.This procedure is not strictly one of denoising; rather it is based on the identification of relatively uncorrupted components of the speech signal, to be used for recognition.

In this paper we present a Maximum-Likelihood algorithm for data compensation through the use of secondary sensors. Specifically, it is a *feature-compensation* algorithm, that attempts to compensate the log-spectral features computed from the speech signal for the effect of the noise, rather than the speech signal itself. For this work we use a GEMS sensors as secondary sensors; however the algorithm itself is generically applicable, with minor modifications, to other types of secondary sensors. We model the noisy recording environment as the combination of a linear filter and additive noise. The environment does not affect the measurements from the GEMS sensor. The parameters of the linear filter and the noise are assumed to be completely unknown. The compensation algorithm learns the parameters from the noisy recording itself and compensates the noisy log-spectral vectors for them. Experimental results show that the use of the secondary sensor can provide significant improvements in recognition performance, as compared to equivalent compensation performed without the use of secondary sensor measurements.

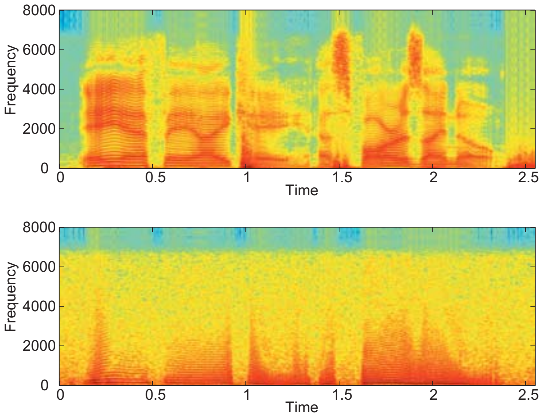

Figure 3: The upper panel shows the narrow-band spectrogram of 2.5 seconds of speech data. The lower panel shows the spectrogram of the corresponding GEMS recording obtained for the same speech.

The rest of the paper is arranged as follows: in Section 2 we briefly describe the GEMS sensor. In Section 3 we describe our model of the recording environment. In Section 4 we describe the noise compensation algorithm itself. In Section 5 we describe our experimental results and in Section 6 we present our conclusions.

2. THE GEMS SENSOR

The Glottal Electromagnetic Sensor (GEMS) is essentially a very low power radar. It is positioned near the glottis of the subject and measures the movement of the rear wall of the trachea. Measurements of tracheal wall motion are deconvolved from the wall-tissue response function to derive the pressure wave that forms the excitation function that drives the vocal tract. Details of the GEMS sensor can be found in [3]. Figure 3 shows the spectrogram of the output of a GEMS sensor, along with the spectrogram of the corresponding speech signal.

3. THE ENVIRONMENT MODEL

The recording environment is modelled as the combination of an unknown linear filter and unknown uncorrelated pseudo-stationary noise that is uncorrelated with the speech signal. Under these assumptions, the power spectrum for the noisy speech, $\boldsymbol{Y}(f)$ is given by

$$\boldsymbol{Y}(f) = |\boldsymbol{H}(f)|^2\boldsymbol{S}(f) + \boldsymbol{N}(f) \tag{1}$$

where $\boldsymbol{S}(f)$ and $\boldsymbol{N}(f)$ are the power spectra of clean speech and the noise, respectively, and $\boldsymbol{H}(f)$ is the frequency response of the linear filter. The relationship between the log spectrum (*i.e.* the logarithm of the power spectrum) of the noisy speech and that of the clean speech is given by

$$Y = S + H + \log(1 + e^{N-H-S}) = F(S, H, N) \tag{2}$$

where $Y = \log(\boldsymbol{Y}(f))$, $S = \log(\boldsymbol{S}(f))$, $H = \log(|\boldsymbol{H}(f)|^2)$ and $N = \log(\boldsymbol{N}(f))$. Here, we have dropped the frequency indicator f for brevity. We use the shortened notation $F(S, H, N)$ to represent $S + H + \log(1 + e^{N-H-S})$ in the rest of this paper.

The log spectrum of the GEMS signal, $G = \log(\boldsymbol{G}(f))$, is unaffected by the environment. Figure [4] shows the model pictorially.

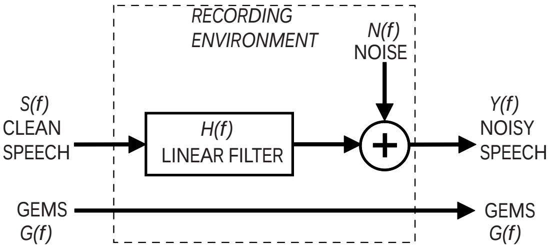

Figure 4: Model of the recording environment for primary and secondary sensor data. The speech signal (from the primary sensor) is affected by a linear filter followed by additive noise. The GEMS recordings (from the secondary sensor) passes through the environment undistorted.

4. THE COMPENSATION ALGORITHM

The compensation algorithm has two distinct stages: the estimation of noise and channel parameters, and the compensation of the noisy log spectra for the estimated parameters

4.1 ESTIMATING NOISE AND CHANNEL

We assume that the joint distribution of the log-spectral vectors of clean speech, X, and G, the log-spectra of the corresponding GEMS measurements is a multivariate mixture Gaussian:

$$\begin{aligned} P(X, G) &= \sum_k P(k)P(X|k)P(G|k) \\ &= \sum_k c_k \aleph(X;\mu_k^x, \sigma_k^x)\aleph(G;\mu_k^g, \sigma_k^g) \end{aligned} \tag{3}$$

Equation (3) explicitly represents the fact that the GEMS and speech data are assumed to be conditionally independent given Gaussian index *k*. $\aleph(X;\mu_k^x, \sigma_k^x)$ represents a Gaussian with mean μ_k^x and variance σ_k^x and c_k represents the *a priori* probability of the k^{th} Gaussian. The parameters of the joint distribution of clean speech and GEMS measurements are obtained from training recordings of the two signals via the EM algorithm.

We assume that over the course of the utterance, the corrupting noise N has a Gaussian distribution with mean μ_N and variance σ_N. The joint distribution of the log-spectral vectors of noisy speech, Y and their corresponding GEMS measurements G, given the channel parameter H and the parameters of the noise distribution is given by

$$P(Y, G|H, \mu_N, \sigma_N) = \sum_k c_k P(Y|k, H, \mu_N, \sigma_N)P(G|k) \tag{4}$$

In Equation (4) the distribution of GEMS measurements is not affected by noise. $P(Y|k, H, \mu_N, \sigma_N)$ cannot be derived in closed form, because of the nonlinearity of $F(S, H, N)$ in Equation (2). In order to make it tractable, we linearize Equation (2) around (μ_k^x, μ_N) using a truncated Taylor series expansion to obtain

$$P(Y|k, H, \mu_N, \sigma_N) = \aleph(Y;F(\mu_k^x, H, \mu_N), \Phi_k(H, \mu_N, \sigma_N)) \tag{5}$$

where $\Phi_k(H, \mu_N, \sigma_N)$ is given by

$$\begin{aligned} \Phi_k(H, \mu_N, \sigma_N) = {} & \nabla_S F(\mu_k^x, H, \mu_N)^T \sigma_k^x \nabla_S F(\mu_k^x, H, \mu_N) + \\ & \nabla_N F(\mu_k^x, H, \mu_N)^T \sigma_N \nabla_N F(\mu_k^x, H, \mu_N) \end{aligned} \tag{6}$$

The parameters of the environment that must be estimated are the linear filter, H and μ_N and σ_N, the mean and variance of the distribution of the noise log spectra. Based on the definition of the joint density of noisy speech and their corresponding GEMS data as given by Equations (4) and (5), a Maximum-Likelihood estimate of H, μ_N and σ_N is derived through an iterative EM algorithm in the following manner:

Let H^n, μ_N^n and σ_N^n denote the estimated values of H, μ_N and σ_N in the n^{th} iteration of the algorithm. $P^n(k|Y)$, the *a posteriori* probability of the k^{th} Gaussian in the mixture Gaussian density of Y computed in the n^{th} iteration is given by

$$P^n(k|Y, G) = \frac{c_k P(Y|k, H^n, \mu_N^n, \sigma_N^n)P(G|k)}{\sum_j c_j P(Y|j, H^n, \mu_N^n, \sigma_N^n)P(G|j)} \tag{7}$$

where $P(Y|k, H^n, \mu_N^n, \sigma_N^n)$ is given by Equation (5). The updated estimates of H, μ_N and σ_N are obtained as

$$\begin{aligned} \{H^{n+1}, \mu_N^{n+1}, \sigma_N^{n+1}\} &= \operatorname{argmax}_{H, \mu, \sigma}\{Q_n(\theta)\} \\ Q_n(\theta) &= \sum_Y \sum_k P^n(k|Y, G)\log(\aleph(Y;F(\mu_k^x, H, \mu), \Phi_k(H, \mu, \sigma))) \end{aligned} \tag{8}$$

where $Q_n(\theta)$ is obtained by summation over all log spectral vectors in the noisy utterance. The detailed solutions are not presented here for reasons of space; however they are easily obtained from Equation (8). Equations (7) and (8) are iterated until convergence to obtain the final estimates $\bar{H}$, $\bar{\mu}_N$ and $\bar{\sigma}_N$.

4.2 COMPENSATING NOISY LOG SPECTRA

Once the noise and channel parameters are estimated, an approximated minimum mean-squared error (MMSE) estimator is used to obtain the clean log-spectral vector underlying every noisy log-spectral vector Y:

$$\begin{aligned} \hat{X} &= E[X|Y, G] = \int_{-\infty}^{\infty} XP(X|Y, G)dX \\ &= Y - \sum_k P(k|Y, G)\int_{-\infty}^{\infty} R(X, N, H)P(X|k, Y, G)dX \end{aligned} \tag{9}$$

where $R(X, N, H) = H + \log(1 + e^{N-H-S})$. Approximating $R(X, N, H)$ with a zero order Taylor series expansion around $(\mu_k^x, \bar{\mu}_N)$ for the k^{th} Gaussian, we get the MMSE estimate:

$$\hat{X} = Y - \sum_k P(k|Y, G)R(\mu_k^x, \bar{\mu}_N, \bar{H}) \tag{10}$$

Cepstra derived from the MMSE estimates of the clean speech log spectra are used for recognition.

5. EXPERIMENTAL RESULTS

The proposed noise-compensation algorithm was evaluated on an in-house corpus of speech and GEMS recordings provided by Carnegie Mellon University and Intelligent Automation Inc. The data consisted of two and a half hours of speech+GEMS recordings obtained from each of three speakers. The utterances consisted of acronyms related to a naval task and their expansions

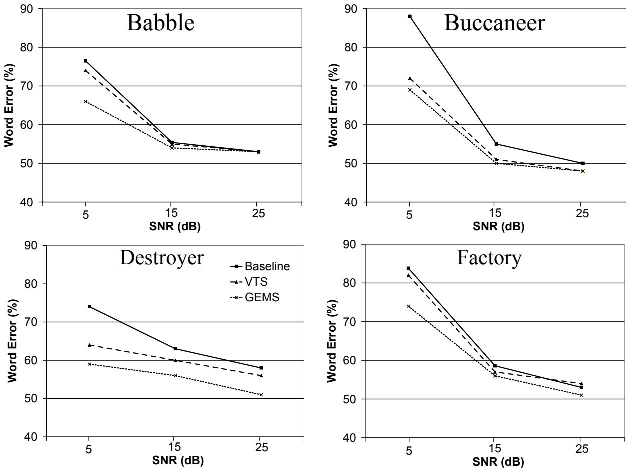

Figure 5: Recognition performance as a function of SNR for speech corrupted by babble, buccaneer noise, factory noise and destroyer operating room noise. In all cases, the solid line represents baseline performance with uncompensated data, the dashed line shows performance with VTS compensation and the dotted line shows performance with GEMS-based compensation.

(e.g. "FYI For Your Information"). Half an hour of the recordings from each speaker was kept aside as training data, to be used to learn the joint distribution of speech and GEMS measurements. Half an hour each of the rest of the data were corrupted to 5dB, 15dB and 25dB respectively with each of four kinds of noises: Babble noise, Factory noise, the cockpit noise of a Buccaneer jet, and engine noise from a destroyer. While the Buccaneer and Destroyer noises were fairly stationary, the Factory and Babble noises were highly nonstationary. Both the speech and the GEMS recordings were sampled at 16000Hz.

32 dimensional mel-frequency log-spectral vectors were computed from the speech signals. Since GEMS signals represent the excitation to the vocal tract, and have a one-to-one frequency correspondence with the speech signal, they were also represented as 32-dimensional mel-frequency log spectral vectors.

The log-spectral vectors of the noisy speech were compensated with the proposed secondary-sensor based algorithm. 13-dimensional cepstral vectors derived from the obtained denoised log-spectral vectors were used for recognition. As a comparison, the VTS algorithm [6] was also used to denoise the noisy log-spectral vectors. We note that the VTS algorithm is very similar to the proposed GEMS-based algorithms, with the exception that no secondary-sensor data are used. Thus the difference between the recognition performance obtained with VTS and the proposed GEMS-based algorithm shows the improvement obtained from using the secondary sensor to guide the compensation.

The CMU Sphinx-3 continuous density speech recognition system was used for the recognition experiments. 2400 tied states, each modelled by a mixture of 4 Gaussians were trained with the Resource Management database. A simple "flat" unigram language model covering all the words in the test data was used for the experiment, in order to emphasize the effect of the acoustics on recognition performance.

Figure 5 shows the recognition results obtained for the various noise types. In addition to the recognition performance obtained with GEMS-based compensation and VTS, the baseline performance obtained with uncompensated noisy speech is also shown.

6. OBSERVATIONS AND CONCLUSIONS

The recognition results in Figure 5 show that the use of secondary sensor measurements for noise compensation results in significant improvements in recognition performance over that obtained when noise compensation is performed using the noisy speech alone. In particular, greater improvements are observed as the SNR of the signal decreases, providing fewer cues to the speech-only VTS algorithm for effective compensation.

In particular, secondary-sensor based compensation is observed to continue to provide effective compensation even on nonstationary noises, *i.e.* for the factory and babble noises, whereas VTS-based compensation fails on these noise types. Both VTS and the secondary-sensor based algorithm make the implicit assumption that the noise corrupting the speech is pseudo-stationary, i.e. it does not change much over the course of the utterance. However, this assumption is unsuitable for babble noise, as a result of which VTS is unable to compensate for it. Regardless of the impropriety of the assumption, the supervision of the measurements from the secondary sensor is able to guide our proposed algorithm to a reasonable estimate of the parameters of the noise distribution, resulting in effective noise compensation.

Although the algorithm presented in this paper is shown to be highly effective, there remains considerable room for improvement in it. *E.g.* the statistical models used in this paper assume that the speech and GEMS measurements are conditionally independent for any Gaussian. However, the speech and GEMS measurements are highly correlated, since the former are derived from the excitation provided by the latter. Significantly greater improvements may be obtained by using these correlations. This will be one the foci of our future work on this topic.

REFERENCES

[1] Singh, R., Stern, R.M. and Raj, B., "Signal and Feature Compensation Methods for Robust Speech Recognition," CRC Handbook on Noise Reduction in Speech Applications, Gillian Davis, Ed. CRC Press, 2002.

[2] Singh, R., Raj, B. and Stern, R.M., "Model Compensation and Matched Condition Methods for Robust Speech Recognition," CRC Handbook on Noise Reduction in Speech Applications, Gillian Davis, Ed. CRC Press, 2002.

[3] Holzrichter, J.F., Burnett, G.C., Ng., L.C., and Lea, W.A. (1998), "Speech Articulator Measurements using Low Power EM-Wave Sensor", J. Acoustic. Soc. Am. 103(1), 622.

[4] Hershey, J., Kristjansson, T. and Zhang, Z. (2004), "Model-based fusion of bone and air sensors for speech enhancement and robust speech recognition", ISCA ITRW on statistical and perceptual audio processing (SAPA2004), Jeju Korea.

[5] Demiroglu, C. and Anderson, D., "Noise Robust Digit Recognition with Missing Frames," Eurospeech 2003, Geneva

[6] Moreno, P.J. (1996), *Speech Recognition in Noisy Environments*, Ph.D Thesis, ECE Department, Carnegie Mellon University