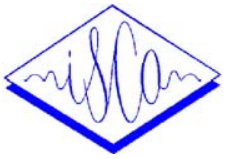


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

FRONT-END IMPROVEMENTS TO REDUCE STATIONARY & VARIABLE CHANNEL AND NOISE DISTORTIONS IN CONTINUOUS SPEECH RECOGNITION TASKS

Xavier Menéndez-Pidal, Ruxin Chen, Duanpei Wu, Mick Tanaka
SONY US Research Labs, 3300 Zanker Road, SJ1B5, San Jose, CA 95134, USA
TEL: +1-408-955-5469, FAX: +1-408-955-6848, EMAIL: xavier@slt.sel.sony.com

ABSTRACT

This paper introduces our actual work in front-end techniques to obtain robust speech recognition devices in mismatch conditions (additive noise mismatch and channel mismatch). Two algorithms have been combined to compensate the distortions due to different channel characteristics and additive noise: 1) A Cepstral Mean Normalization and Variance Scaling technique (MNVS) and 2) An Adaptive Gaussian Attenuation algorithm (AGA). Combining both techniques the channel distortion effects were reduced to 90% on the HTIMIT task and the additive noise effects were reduced to 80% on the TIMIT task corrupted with additive car noise.

Keywords: robust speech recognition, noise and channel compensation

1. INTRODUCTION

To improve the portability of a speech recognizer in very different noisy environments is still a difficult problem to solve. While an actual speech recognizer can provide very good performance in laboratory conditions, in real scenarios such as: hands-free car applications, the presence of interfering noises can drastically reduce the accuracy of a recognizer. Automatic speech recognizers tend to degrade in performance when there is a mismatch between training and testing conditions and to found reliable algorithms independent of the noise source is still a challenge. In this paper a combination of two different techniques to improve the portability of a speech recognizer without decreasing the system accuracy in clean conditions is presented. The first technique to be introduced is a noise attenuation scheme called: Adaptive Gaussian Attenuation (AGA). AGA blocks the noise signal and tends to preserve the speech signal. In AGA the mean and the standard deviation of the noise are used to attenuate the incoming signal following an evolution inverse to the noise distribution. The AGA algorithm is compared with the classical Spectral Subtraction (SS) technique providing more noise reduction effects and a configuration of the algorithm task independent. In the second technique the classical Speech Cepstral Mean Normalization algorithm has been improved using also Cepstral Variance Scaling to reduce both additive and convolutive distortion effects.

2. THE TASK

2.1 Recognizer and Front-End Description

The recognition system used in the experiments was based on a classical 350 Context Dependent phones HMMs using 4 Gaussian mixtures per state on average. In the feature analysis a pre-emphasis coefficient of 0.97, a Hamming window of 25 ms, a frame shift of 10 ms, and 16 Mel-scale filters covering from 80 Hz to 3800 Hz were used. Thirteen Cepstral features (C0-C12) were calculated in combinations with 13 delta and 13 delta-delta Cepstral features estimated with a 9 points Time Domain Cosine Transform [1]. In the test experiments continuous phoneme accuracy was estimated to analyze the improvements of each technique using a Beam Search of 75 active states.

2.2 Data Base Description

The experiments were carried out using two American English continuous speech data bases: TIMIT and HTIMIT, to analyze channel and severe additive noise compensation. The system was trained using clean speech provided by the TIMIT data base and the testing set was performed over the TIMIT and HTIMIT data bases. A total of 384 files per microphone produced by 24 males and 24 females speakers was used during the tests. The maximum performance expected for the system was measured training and testing in matched conditions with Mel Frequency Cepstral Coefficients (MFCC).

2.1.1 Channel & Middle Noise Compensation Task

To analyze the effects of 10 microphones' distortions, the HTIMIT data base was used. The

HTIMIT database is a playback of the original TIMIT through 4 carbon telephone microphones (cb1… cb4), 4 electret telephone microphones (el1… el4), and a portable (cord-less) telephone microphone (pt1). Two high quality Sennheizer head-mounted microphones (senh, timit) were also used as reference microphones. A description of the existing microphone distortions can be found in [1]. The HTIMIT task is dominated by linear channel distortions and secondary by stationary & variable additive noise (SNR: 30~20 dB) and non-linear channel effects.

2.1.2 Severe Additive Noise Compensation Task

Additive noise degradation and compensation was measured mixing the TIMIT testing set with 4 hours of car noise. The noise was produced by 5 different cars at different SNRs, road, driving, background music and whether conditions. In this task severe additive noise distortions are predominant. The car noise was mixed with the TIMIT data base at real SNR conditions except for the Integra car noise. Table 1 shows a description of the noises used.

Car	Road	whether, music	SNR
Estima	Town	fine, yes	7.5
Integra	high way	fine, no	12
Integra	high way	fine, no	6
Integra	high way	fine, no	0
Mark2	Idle	fine, no	20
Mark2	mid way	fine, no	10
Mark2	high way	fine, no	2.5
Impreza	high way	fine, no	3.1
Impreza	high way	rain, no	1.7
Starlet	high way	fine, yes	0.4
Starlet	Idle	idle, yes	16.1

TABLE 1. Car noise description mixed with TIMIT.

3. NOISE COMPENSATION SCHEMES

3.1 Adaptive Gaussian Attenuation (AGA)

In the FFT domain the noisy speech signal is distorted as follows:

$$Y_{k,n} = X_{k,n} + N_{k,n}$$

Where $X_{k,n}$ is the power or magnitude energy at frame n and frequency k of the original speech, $N_{k,n}$ is the noise energy and, $Y_{k,n}$ is the energy of the corrupted speech. In our experiments, slightly better results were obtained by using the magnitude energy rather than using the power domain. To better attenuate background music and constant tones present in HTIMIT, the noise compensation was performed over the FFT-bins rather than after the filter bank analysis. To attenuate the noise energy distribution, it is assumed that the low amplitude energy components are dominated by the interfering noise while the high amplitude energy components are dominated by the speech signal. The effects of the additive noise over the clean speech energy distribution are displayed in Fig. 3.

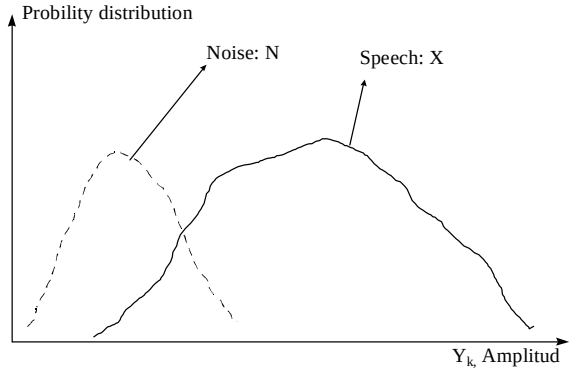


Figure 3. Noise effects on the original clean speech distribution

The noise attenuation model proposed tries to attenuate the noise distribution N_k while preserving the clean speech distribution X_k using a non-linear filter. In the non-linear filter, the mean and the variance of the noise are introduced to guide the attenuation of the incoming signal as follows:

$$Yat_k = \frac{Y_k}{1 + A\exp\left(-\left(\frac{Y_k - \alpha\mu_k}{\sqrt{2}\delta_k}\right)^2\right)} \quad if\ Y_k \geq \alpha\mu_k$$

$$Yat_k = \frac{Y_k}{1+A} \quad otherwise$$

Where k is frequency index, Yat_k is the attenuated noisy signal, Yk is the incoming noisy speech, and μ_k and δ_k are the average and the standard deviation of the noise energy. Like in Classical SS [3], this non-linear filter uses 2 parameters which need to be optimized experimentally: 1) the overestimation coefficient α and, 2) the attenuation coefficient A called flooring factor in the literature. In this new model, the overestimation constant was also related with the variance of the noise but no significant difference was obtained. Using this non-linear filter a variable noise attenuation is performed varying from 1+A (maximal attenuation) to 1 (no attenuation). Similar behavior is also ob-

tained with SS which can be expressed as follows:

$$Yat_k = Y_k - \alpha\mu_k \quad if\ Y_k - \alpha\mu_k > \frac{\mu_k}{1+A}$$

$$Yat_k = \frac{Y_k}{1+A} \quad otherwise$$

The main difference between SS and the Gaussian attenuation model is the evolution of the attenuation in the high energy values. Using SS the evolution of the attenuation is related with the noise mean and the overestimation factor. For example, if the overestimation is doubled, the evolution of the noise attenuation is two time slower from the maximal attenuation A+1 to the minimal attenuation 1. On the other hand, in the Gaussian model the evolution of the attenuation is only related to the spread or standard deviation of the noise. The α factor in the Gaussian model determines the energy value where the signal begins to be attenuated from 1+A to 1 but, does not affect the evolution of the attenuation. Compared to SS, the previous Gaussian model was less sensitive to the overestimation coefficient used. Nevertheless, in SS and in the Gaussian model the election of the attenuation factor A is dependent on the SNR conditions. For very noisy conditions, better results were obtained using a higher A value and for low noise conditions better results were obtained with a lower A attenuation factor. Also, the noise does not always corrupt all the frequencies equally and an attenuation coefficient dependent of the frequency seems a better solution. To overcome those problems an adaptive A attenuation factor dependent of the frequency k and the SNR was introduced in the Gaussian model. The adaptive attenuation based on the Shannon channel capacity provided to be very successful and can be expressed as follows:

$$A_k = \frac{A}{\log_2(1+\frac{\alpha_2\mu_k}{Sp_k})}$$

Where A is the global attenuation coefficient, α_2 is an optional overestimation factor, Sp_k and μ_k are the averages of the noisy speech energy and the noise energy in the frequency k.

3.2 Cepstral MNVS

The previous noise attenuation algorithms were also combined with the Speech based Cepstral Mean Normalization and Variance Scaling technique (MNVS) analyzed in [1, 2] to provide microphone independence and a more robust additive noise adaptability. Linear channel distortions due to microphone or channel characteristics introduce a constant shift in the Cepstral features which can be eliminated by subtracting from the Cepstral domain a long term average estimated in Speech segments. On the other hand additive background noise tends to eliminate the spectral valleys decreasing the dynamic variations of the Cepstral and Differential features. In the Cepstral scaling algorithm two independent left and right Laplacian variances (^rv_i, ^lv_i) were used to better compensated the asymmetric noise masking effects. The transformation of the Cepstral and Cepstral-differential features was performed as follows:

$$ns_i = \frac{a_i - x_i}{l_{v_i}} \quad x_i < a_i$$

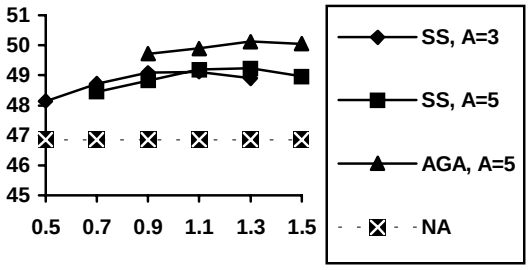
$$ns_i = \frac{x_i - a_i}{r_{v_i}} \quad x_i > a_i$$

Where x_i is the ith component of the original feature vector and ns_i is the normalized and scaled one. The statistics a_i, ^rv_i, ^lv_i, are the average, right and left variances of the ith feature. In the AGA and MNVS algorithms, the estimation of the means and variances was performed using a frame synchronous recursive procedure introduced in [1, 2]. The estimation of the means and variances is based on an IIR filter using a forgetting factor. The algorithms need also a speech, noise discrimination process to update noise and speech statistics independently. A global Speech forgetting factor equal to 0.997 was used to update a_i, ^rv_i, ^lv_i, Sp_k. The noise statistics μ_k, δ_k were updated with a noise forgetting factor equal to 0.95.

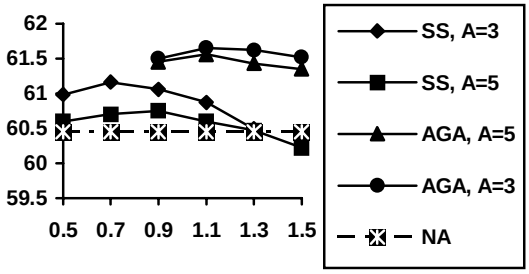
4. EXPERIMENTAL RESULTS

Graphs 1, 2, 3, 4 summarize the global phoneme string accuracy obtained on the two tasks: HTMIT and TIMIT + Car Noise. The Graphs analyze the influence of the attenuation coefficient A or the overestimation factor α in SS or AGA models. The attenuation algorithms were combined with the Cepstral MNVS scheme. The results with MNSV alone with no attenuation (NA) are also provided. In all the experiments α_2 was equal to 1.

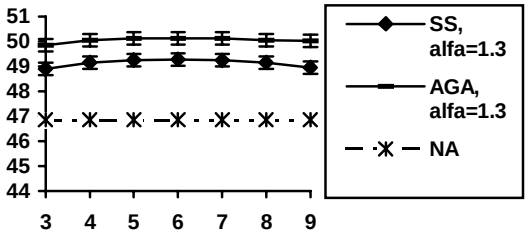
Graph 1:
Ph. acc. vs alfa in Car task



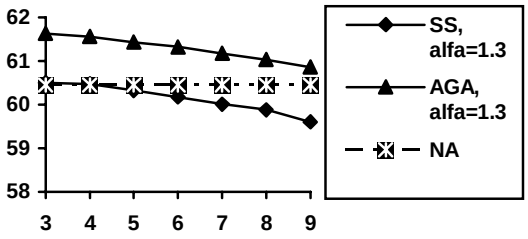
Graph 2:
Ph. acc. vs alfa in Mic task



Graph 3:
Ph. acc. vs Aenuation in Car task



Graph 4:
Ph. acc. vs Atnuation in Mic task



The new AGA algorithm significantly improved the system accuracy in optimal conditions and provided to be much less task and SNR-dependent than Spectral Subtraction. In combinations with MNVS Non-linear SS [4] did not provided any benefit in AGA or SS. The last two Tables summarize the independent system improvements introduced by each technique in the Car and Microphone tasks. The final configuration adopted uses α=1.3, α_2=1.5, A=5 in the Magnitude domain.

Mic	mfcc	mnvs	Aga	mnvs+aga	match
cb1	56.24	62.21	58.25	63.58	63.2
cb2	59.45	65.36	61.39	65.67	64.2
cb3	38.28	50.62	43.57	51.62	58.8
cb4	41.9	53.62	47.92	55.82	59.9
el1	57.23	65.33	58.49	65.27	64.3
el2	49.51	61.09	50.39	63.05	62.9
el3	50.88	57.73	51.82	58.37	60.4
el4	52.42	61.41	51.72	62.12	61.9
pt1	36.81	56.54	45.72	58.5	61.7
senh	62.51	64.68	63.73	65.39	64.1
timit	66.56	66.71	65.67	66.37	66.71
Ave.	51.98	60.45	54.42	61.42±0.23	62.6

Car	mfcc	mnvs	Aga	mnvs+aga	match
Est7.5	35.92	48.59	42.4	49.97	
Imp3.1	25.54	39.59	34.58	44.46	
Imp1.7	18.66	32.27	28.01	37.1	
Mar2.5	31.53	45.59	40.93	49.99	
Mar20	56.7	61.94	63.04	64.43	
Mar10	51.16	58.5	56.6	60.68	
Sta0.4	25.19	38.71	31.98	42.46	
Sta16	49.18	57.91	52.16	57.72	
Int12	39.62	52.94	50.7	56.26	60.3
Int6	29.91	44.15	39.46	48.68	54.6
Int0	20.85	35.17	29.16	40.04	46.6
Ave.	34.93	46.85	42.63	50.16±0.25	

In both tasks MNVS was the most effective compensation algorithm but AGA significantly helped the system.

5. CONCLUSION

Two low-cost algorithms tested over 22 different distortions have been presented to provide on-line task independent robust speech recognition.

6. REFERENCE

[1] X. Menéndez-Pidal, R. Chen, D. Wu, M. Tanaka, "Compensation Of Channel And Noise Distortions Combining Normalizations & Speech Enhancement Techniques", Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, 99.

[2] S. Tibrewala, H. Hermansky, "Multi-band & adaptation approaches to robust Speech Recognition", pp. 2619-2622, Eurospeech'97.

[3] J. Nolazco, S. Young, "Adapting a HMM-based Recognizer for Noisy Speech Enhanced by Spectral Subtraction", TR. 123, CUED-Cambridge Univ., 1993.

[4] P. Lockwood, J. Boundy, "Experiments with NSS, HMM, & projection for Robust SR in cars", Speech Communicaton, Vol. 11, pp 215-228, 1992.