ELSEVIER

Speech Communication 34 (2001) 115–126

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Compensation of channel and noise distortions combining normalization and speech enhancement techniques

Xavier Menéndez-Pidal *, Ruxin Chen, Duanpei Wu, Mick Tanaka

SONY US Research Labs, 3300 Zanker Road, SJ1B5, San Jose, CA 95134, USA

Abstract

This paper introduces two techniques to obtain robust speech recognition devices in mismatch conditions (additive noise mismatch and channel mismatch). The first algorithm, adaptive Gaussian attenuation algorithm (AGA), is a speech enhancement technique developed to reduce the effects of additive background noise in a wide range of signal noise ratio (SNR) and noise conditions. The technique is closely related to the classical noise spectral subtraction (SS) scheme, but in the proposed model the mean and variance of noise are used to better attenuate the noise. Information of the SNR is also introduced to provide adaptability at different SNR conditions. The second algorithm, cepstral mean normalization and variance-scaling technique (CMNVS), is an extension of the cepstral mean normalization (CMN) technique to provide robust features to convolutive and additive noise distortions. The requirements of the techniques are also analyzed in the paper. Combining both techniques the relative channel distortion effects were reduced to 90% on the HTIMIT task and the relative additive noise effects were reduced to 77% using the TIMIT database mixed with car noises at different SNR conditions. © 2001 Elsevier Science B.V. All rights reserved.

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 finding reliable algorithms independent of the noise source is still a challenge. The two main sources that can cause acoustic distortion are: (1) presence of additive noise such as car noise, music or background speakers, (2) convolutive distortions due to the use of different microphones, telephone channel and reverberations. One classical approach towards robust recognizers is to identify the adverse effect and compensate for the change introduced by the interfering noise. These techniques can be implemented in the feature front-end to provide fast environment compensation. In this paper, a combination of two low computational cost techniques to improve the portability of a

* Corresponding author. Tel.: +1-408-955-5469; fax: +1-408-955-6848.

E-mail address: xavier@slt.sel.sony.com (X. Menéndez-Pidal).

0167-6393/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(00)00049-2

speech recognizer without decreasing the system accuracy in clean conditions is presented. Section 2 describes the experimental task, databases, training and recognition materials used. Section 3 analyzes the noise attenuation and compensation schemes introduced. The first algorithm, adaptive Gaussian attenuation (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 related with the noise distribution. The AGA algorithm is compared to the classical spectral subtraction (SS) technique providing more noise reduction effects and a task-independent configuration of the algorithm. The second compensation technique called cepstral mean normalization and variance-scaling technique (CMNVS) is later introduced to compensate additive and convolutive distortions. Section 4 summarizes the experimental results obtained with a recognizer and Section 5 analyzes an acoustic assessment test performed to compare the acoustic quality of SS versus AGA.

2. Task description

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 four Gaussian mixtures per state on average (Chen et al., 1998; Lee and Hon, 1989). 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 (Menéndez-Pidal et al., 1999; Milner, 1996). Fig. 1 shows the front-end block diagram and different improvements introduced in our experiments. In the test experiments continuous phoneme accuracy was estimated to analyze the improvements of each technique using a Beam Search of 75 active states.

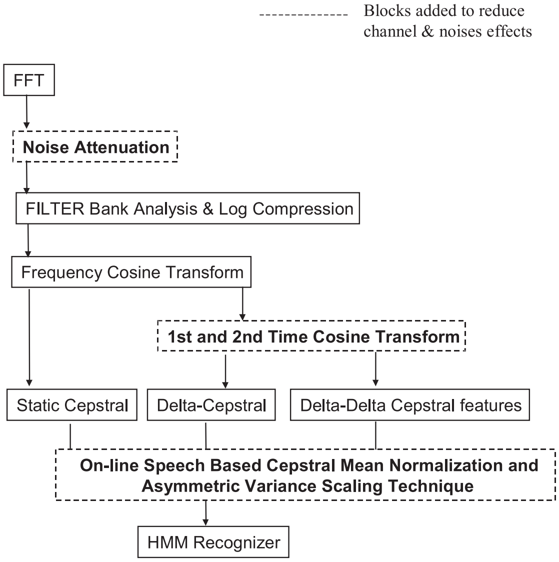
Blocks added to reduce channel & noises effects
FFT
Noise Attenuation
FILTER Bank Analysis & Log Compression
Frequency Cosine Transform
1st and 2nd Time Cosine Transform
Static Cepstral
Delta-Cepstral
Delta-Delta Cepstral features
On-line Speech Based Cepstral Mean Normalization and Asymmetric Variance Scaling Technique
HMM Recognizer

Fig. 1. Feature extraction front-end used to minimize noise and channel distortions.

2.2. Database description

The experiments were carried out using two American English continuous speech databases: TIMIT and HTIMIT, to analyze channel and severe additive noise compensation. The system was trained using clean speech provided by the TIMIT database and the testing set was performed over the TIMIT and HTIMIT databases. Microphone compensation was done using 10 different microphones provided in the HTIMIT database. A total of 384 files per microphone produced by 24 male and 24 female speakers were used during the tests. Additive noise compensation was measured mixing the TIMIT testing set (384 files and 48 speakers) with 11 different car noises. Experiments training and testing in matched conditions with normalized Mel frequency cepstral coefficients (MFCC) were also done to measure the maximum performance of the system.

2.2.1. Channel and middle noise compensation task

To analyze the effects of 10 microphones' distortions, the HTIMIT database was used (Reynolds, 1997). The HTIMIT database is a playback of the original TIMIT through four carbon telephone microphones (cb1, ..., cb4), four electret telephone microphones (el1, ..., el4), and a portable (cord-less) telephone microphone (pt 1). Two high quality Sennheizer head-mounted microphones (senh, timit) were also used as reference microphones. The HTIMIT task is dominated by linear channel distortions and secondary by stationary and variable additive noise (signal noise ratio (SNR): 32 $\sim$ 19 dB) and non-linear channel effects. A description of the existing microphone distortions was generated using the sweep tone files included in the HTIMIT database. The sweep tone files were used to identify (1) presence of constant tones, (2) presence of background constant or variable noise, and (3) presence of secondary harmonics found in non-linear channels. The type of channel and noise distortions was cataloged using the following qualifiers: L (linear microphone), NL (non-linear microphone), T (presence of harmonic tone noise), CBN (constant background noise) and VBN (variable background noise). Table 1 shows the kind of distortions found for each microphone.

2.2.2. Severe additive noise compensation task

Additive noise degradation and compensation was measured artificially mixing the TIMIT testing set with 4 h of car noise. The car noise was recorded in Japan using five cars at different road, driving, and background music and weather conditions. In this task severe additive noise distortions are predominant. The car noise was mixed with the TIMIT database at real SNR conditions except for the Integra car noise. The SNRs used in those experiments were determined in Japan recording speakers in the moving cars (Iwahashi et al., 1998). To analyze the direct influence of the SNR in the speech recognition accuracy the same Integra car noise was mixed at three different SNR conditions (0, 6, 12 dB). Table 2 shows a description of the noises used.

Table 1
Microphone distortions found in HTIMIT

Microphone	Channel characteristics	Noise characteristics	SNR
cb1	L	T	21.3
cb2	NL	CBN + T	25.2
cb3	NL	VBN	23.4
cb4	NL	CBN + T	24.1
el1	NL	T	27
el2	L	VBN + T	22.9
el3	L	VBN + T	24.2
el4	L	VBN + T	19
pt1	NL	CBN + T	19.2
Senh	L	T	27.2
Timit	L	Clean	31.8

Table 2
Car noise description mixed with TIMIT

Car	Road	Weather, 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	Fine, yes	16.1

3. Noise compensation schemes

3.1. Adaptive Gaussian attenuation

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

$$Y_{k,t} = X_{k,t} + N_{k,t}, \tag{1}$$

where $X_{k,t}$ is the power or magnitude at frame t and frequency k of the original speech, $N_{k,t}$ is the noise power or magnitude, and $Y_{k,t}$ is the power or magnitude of the corrupted speech. When additive noise is present, the Gaussian power distribution of the original speech signal becomes bimodal or non-Gaussian. The typical effects of the additive noise over the clean speech distribution are schematically displayed in Fig. 2 (for more information and real examples see (Gales, 1998)).

The noise reduction model proposed tries to attenuate the noise distribution N_k while preserving the clean speech distribution X_k using a non-linear parametric filter. The technique developed tries to improve the classical SS (Berouti et al.,

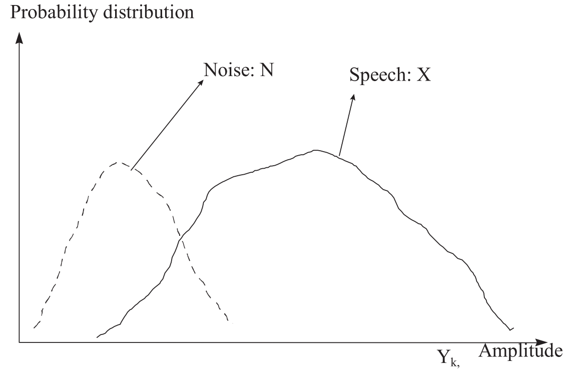

Fig. 2. Hypothetical scenario of the noisy speech distribution.

1979; Boll, 1979; Nolazco and Young, 1993) method to obtain a more reliable noise reduction filter, less dependent on the parameters setup selected, SNR and noise characteristics. 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. In SS the noise attenuation is carried out using a filter as follows:

$$Yat_k = \begin{cases} Y_k - \alpha\mu_k & \text{if } Y_k - \alpha\mu_k > \dfrac{\mu_k}{1+A}, \\ \dfrac{Y_k}{1+A} & \text{otherwise,} \end{cases} \tag{2}$$

where k is the frequency index, Yat_k the attenuated noisy signal, Y_k the incoming noisy speech, and μ_k is the noise power or magnitude average. This filter uses two parameters which need to be optimized experimentally: (1) the overestimation coefficient α, and (2) the attenuation coefficient A called flooring factor in the literature. SS assumes that the variance of the noise is zero and the mean of the noise is only used to attenuate the incoming signal leading to an over simplified model. For example, two noises with identical means but different standard deviations should be attenuated differently. In this case, the input signal corrupted with the noise with higher variance should be more attenuated than the input signal corrupted with the noise with less variance. Lockwood and Boudy (1992) presented an improvement in the SS scheme to overcome this assumption but not really using the variance of the noise. In the proposed noise filtering technique (Gaussian attenuation), the mean and the standard deviation of the noise are introduced to better guide the attenuation process. Using Gaussian attenuation the incoming signal is attenuated as follows:

$$Yat_k = \begin{cases} \dfrac{Y_k}{1 + A\exp - \left(\frac{Y_k - \alpha\mu_k}{\sqrt{2}\delta_k}\right)^2} & \text{if } Y_k \geqslant \alpha\mu_k, \\ Y_k/(1+A) & \text{otherwise,} \end{cases} \tag{3}$$

where δ_k is the standard deviation of the noise power or magnitude, the other variables and parameters are identical to those used in SS. The main difference between SS and the Gaussian attenuation model is the evolution of the attenuation for input signals with high energy. The evolution of the attenuation in this work is referred to the input signal divided by the output signal Y_k/Yat_k. Figs. 3 and 4 display the typical form of the attenuation obtained with SS (Fig. 3) and the Gaussian noise attenuation method (Fig. 4).

In SS, the attenuation for low amplitude input signal ($Y < \alpha\mu$) is related with the attenuation factor A, and the attenuation for high amplitude input signal ($Y > \alpha\mu$) is controlled by the noise mean and the overestimation factor α (see Fig. 3). For example, if the overestimation is doubled, the evolution of the noise attenuation is two times slower from the maximal attenuation $A + 1$ to the minimal attenuation 1. The slope of the attenuation curve in SS has a clear discontinuity near $\alpha\mu$ and makes the SS filter sensitive to the joint optimization of the parameters α and A selected.

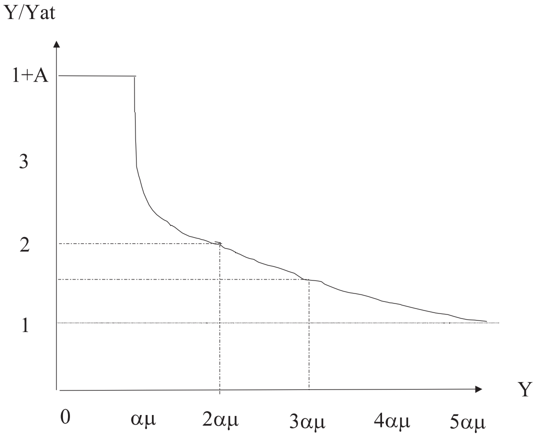

Fig. 3. Attenuation curve of the SS model.

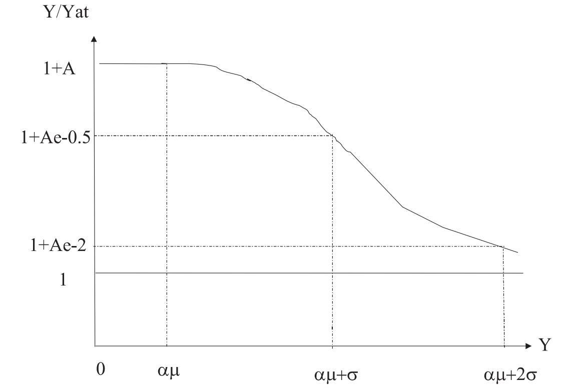

Fig. 4. Attenuation curve of the Gaussian Attenuation model (note that the curve is a half-Gaussian with a width determined by σ, shifted upwards to asymptote to 1 at large Y).

On the other hand, in the Gaussian attenuation model the slope of the attenuation curve has no discontinuities. The influence of the attenuation factor A is progressively eliminated when the input signals have a low probability of belonging to the noise distribution $(Y > \alpha\mu)$. The factor α in the Gaussian model has a minor effect and only determines the power or magnitude value where the signal begins to be attenuated from $1 + A$ to 1 (see Fig. 4).

The algorithm described above is less sensitive to the α overestimation coefficient used. For a large number of noises and SNR conditions the same α was adopted providing a more stable algorithm. The second improvement introduced in the Gaussian attenuation models tries to automatically adapt the attenuation factor A to different SNR conditions. In previous works significant improvements have been reported adapting the attenuation of the incoming input signal to different SNR conditions using different noise attenuation models. In (Schless and Class, 1997) an improved SS implementation with an adaptive flooring and overestimation factor is described. In (Xie and Van Compernolle, 1994, 1996) a multi-layer perceptron (MLP) system was proposed to attenuate the incoming signal. In the previous work the attenuation curves obtained with the MLP are clearly SNR dependent. The shape of the MLP attenuation curves is very similar to the attenuation curves produced with the AGA algorithm proposed here. Also, for some noises like car noise, tones and background music noise not all the frequencies are equally corrupted. In this case, a better solution seems to provide an attenuation coefficient dependent on the frequency. The adaptive attenuation based on the Shannon channel capacity (Verdu, 1998) has been very successful and can be expressed as follows:

$$A_k = \frac{A}{\log_2\left(1 + \frac{Sp_k}{\mu_k}\right)}, \tag{4}$$

where A is the global attenuation and Sp_k and μ_k are the averages of the noisy speech and the noise power or magnitude in the frequency k. The logarithm operator tends to decrease the influence of A for frequencies with high SNR range and maintains significant attenuation for low SNR. The estimation of the means (μ_k, Sp_k) and standard deviations (δ_k) was performed using the following frame synchronous recursive procedures:

$$\begin{aligned} Sp_k(t) &= \beta Sp_k(t-1) + (1-\beta)Y_k(t), \\ \mu_k(t) &= \lambda\mu_k(t-1) + (1-\lambda)Y_k(t), \\ \theta_k(t) &= \lambda\theta_k(t-1) + (1-\lambda)Y_k(t)^2, \\ \sigma_k(t) &= \sqrt{(\theta_k(t) - \mu_k(t)\mu_k(t))}. \end{aligned} \tag{5}$$

where β is the speech forgetting coefficient typically equal to 0.997 and λ is the noise forgetting coefficient typically equal to 0.95.

Finally, this filter can be applied in either the power or magnitude domain as well as, after FFT or after filter bank analysis (i.e. interpreting the k subscript as indexing into FFT bins or the integrated filter bank bins). In our experiments, slightly better but not statistically significant results were obtained by using the magnitude rather than using the power domain. Some noises like music or constant background tones have a very fine harmonic structure and are better attenuated after the FFT analysis. In this case, we use the maximum frequency resolution during the attenuation. If the attenuation is performed after filter bank analysis to simplify the attenuation process it is assumed the noise uniformly corrupt all the FFT-bins for one filter. This assumption can be made for broad band noise like car noise that has a slow frequency evolution. In our experiments, we found the same recognition results performing the attenuation after filter bank or FFT analysis for files corrupted with pure car noise. Some better results were obtained performing the attenuation at FFT level for files corrupted with noise with a clear harmonic structure.

3.2. Cepstral mean normalization and variance-scaling technique

The previous noise attenuation algorithms were also combined with the speech based CMNVS technique to provide microphone independence and a more robust additive noise compensation. 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. Background noise, on the other hand, has two effects on the original signal representation (power, magnitude or log domains). Background noise first shifts the average speech distribution, and also, background noise tends to mask the speech distribution with low amplitude. The noise masking tends to do not effect the portion of the speech signal with high amplitude energy. The elimination of the spectral valleys asymmetrically decreases the dynamic range of the power or magnitude channel values. The decrease in the dynamic variation is propagated later to the cepstral features and differential features (delta and delta–delta cepstral features), which are linear combinations of the log power or magnitude channel values. The average of the cepstral features is also shifted. Some of the asymmetrical masking effects were also translated in the cepstral domain. For example, the shape of cepstral distributions for the coefficients C0 and C1 become very asymmetric in noisy conditions. In the algorithm adopted, using the cepstral mean normalization (CMN) method has compensated linear channel distortions. Furthermore, dynamic range decreases, which have been compensated by variance-scaling of the cepstral and differential features, used in the front-end. In previous works (Viikki and Laurila, 1997; Viikki et al., 1998; Tibrewala and Hermansky, 1997), the normalization was carried out estimating the mean and one variance for each parameter. To provide further accuracy during the adaptation, the use of two asymmetric variances is proposed in this work to better track asymmetric background noise masking effects. The transformation of the cepstral and cepstral-differential features was performed as follows:

$$ns_i = \begin{cases} \dfrac{a_i - x_i}{{}^{\mathrm{l}}v_i}, & x_i < a_i, \\ \dfrac{x_i - a_i}{{}^{\mathrm{r}}v_i}, & x_i > a_i, \end{cases} \tag{6}$$

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, ${}^{\mathrm{r}}v_i$, ${}^{\mathrm{l}}v_i$ are the average, right and left variances of the ith feature. The estimation of the mean and variances for a monophone or triphone recognizer system needs clearly to be done over long span speech segments (3–10 s) to obtain statistical values independent of the speech segment that generated them (Neumeyer et al., 1994; Gauvian et al., 1996). Phonetic based recognizers (like monophone systems) learn a limited contextual information and fluctuations in the front-end due to the surrounding speech units must be omitted. In a phonetic recognizer, if the front-end is normalized with a very short speech sequence, the means and the variances have strong fluctuations, which may lead to recognition errors. While in a typical phonetic recognizer the front-end require 3–10 s of speech to be stabilized, in word models recognition systems the front-end are stabilized with less than 1 s of speech (Viikki and Laurila, 1997; Viikki et al., 1998). This means that the phonetic recognizers can only remove slow time varying distortions (<0.3 Hz) with classical normalization techniques. The estimation of the means and variances should also be done mainly over speech segments and blocked over long noise area, requiring a minimum voice activity device (VAD) or push to talk technique. If very long noise segments are included during the update stage of the means and variances, the statistics tend to reflect only the noise characteristics leading to recognition errors too.

In our implementation, the mean and variances were estimated at speaker level to reduce the speaker variability and to provide a first speaker compensation. During the recognition and the training phase the estimation of those statistics was carried using the same procedure: the on-line normalization mode. This is the only reliable method for real-time applications to avoid introducing delays in the feature extraction stage. Also, to obtain an algorithm adaptive to new slow varying conditions (channel, noise or speaker) a mechanism to forget the past history was introduced during the estimation of the mean and variances. This can be done by (1) using a finite impulse response (FIR) filter (Hanson et al., 1995; Rosenberg et al., 1994; Viikki and Laurila, 1997) or (2) using an infinite impulse response (IIR) filter with a forgetting factor β (Gauvian et al., 1996; Tibrewala and Hermansky, 1997; Viikki et al., 1998). The FIR implementation evaluates the estimates over a finite window length of N frames, and was abandoned due to the computational overload introduced (2 orders of magnitude more expensive than the IIR implementation). For the decomposed right and left variances estimation proposed the IIR filter becomes

$$\begin{aligned} &a_i(t) = \beta a_i(t-1) + (1-\beta)x_i(t), \\ &{}^{\mathrm{l}}v_i(t) = \beta^{\mathrm{l}}v_i(t-1) + (1-\beta)(a_i(t) - x_i(t)), \\ &\quad x_i(t) < a_i(t), \\ &{}^{\mathrm{r}}v_i(t) = \beta^{\mathrm{r}}v_i(t-1) + (1-\beta)(x_i(t) - a_i(t)), \\ &\quad x_i(t) > a_i(t). \end{aligned} \tag{7}$$

A series of experiments showed that values of the speech forgetting factor β between 0.995 and 0.998 (equivalent to a time constant of 2–5 s) provided maximum performance while values below 0.99 strongly degrade the system accuracy. These values provide a good compromise between long-term precision in stable channel and noise conditions and short-term adaptability when the environmental conditions vary.

3.3. Joint implementation of AGA and CMNVS techniques

To avoid introducing discontinuities in the front-end both algorithms (AGA and CMNVS) are continuously applied in speech and noise areas. The algorithms need a speech/noise discrimination process to update the noise $(\mu_k,\ \sigma_k)$ and the speech $(a_i,\ {}^{\mathrm{r}}v_i,\ {}^{\mathrm{l}}v_i,\ SP_k)$ statistics independently. During the speech segments, the speech variables $(a_i,\ {}^{\mathrm{r}}v_i,\ {}^{\mathrm{l}}v_i,\ Sp_k)$ are updated using a unique speech-forgetting factor equal to 0.997 and the last noise statistics are kept fixed. In (Menéndez-Pidal et al., 1999) we reported experiments related to the adaptability of the front-end with different SNR and channel conditions buffering the last estimates between files, requiring 3 s of speech to stabilize the front-end to new environmental noise or channel conditions. On the other hand, in noise areas, the last speech statistics are kept fixed and only the noise statistics $(\mu_k,\ \sigma_k)$ are updated with a noise-forgetting factor equal to 0.95–0.99. Reliable noise statistics can be obtained using 350 ms of noise.

4. Experimental results

Figs. 5–7 analyze the influence of the attenuation coefficient A and the overestimation factor α in SS or AGA. The graphs summarize the global phoneme string accuracy obtained on the HTMIT and TIMIT + car noise tasks, combining the

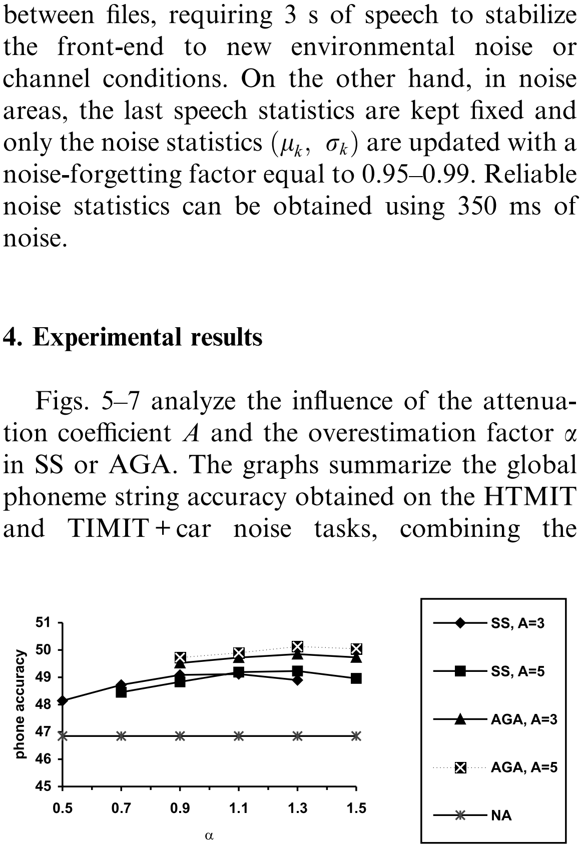

Fig. 5. Phone accuracy versus α overestimation factor in the TIMIT + car noise task.

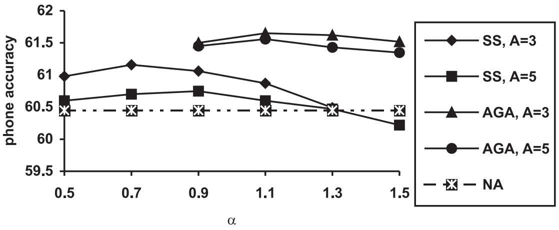

Fig. 6. Phone accuracy versus α overestimation factor in the HTIMIT task.

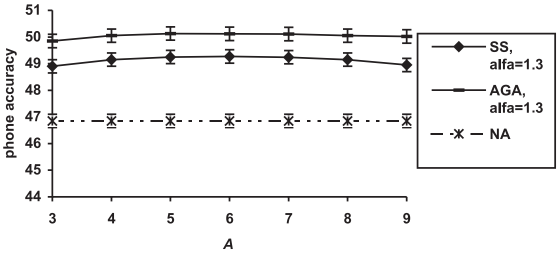

Fig. 7. Phone accuracy versus attenuation factor A in the TIMIT + car noise task.

attenuation algorithms in the magnitude domain with the cepstral MNVS scheme. The results with CMNSV alone with no attenuation (NA) are also provided to measure the final relative improvement introduced by the attenuation algorithms training in matched and unmatched conditions. In SS a complete different algorithm setup was required in the two tasks to obtain phone accuracy gain (see Figs. 5 and 6). For example, using SS the best overestimation factor α and attenuation factor A in the HTIMIT task were $\alpha \sim 0.7$ and $A \leqslant 3$, and in the TIMIT + car noise task the best results were obtained with $1.1 \leqslant \alpha \leqslant 1.3$ and $3 \leqslant A \leqslant 5$, respectively (see Figs. 5 and 6). In SS the optimization of α and A was not independent also, as reported by (Schless and Class, 1998). For example for noises with low SNR ~ 0 db the optimal values are around $A \sim 5$ and $\alpha \sim 1.3$ and for noise with high SNR ~ 20 db the best values are around $A \leqslant 3$ and $\alpha \sim 0.7$.

On the other hand, AGA does not heavily depend on the α and A factors elect, see Figs. 5 and 6, and values of $1.1 \leqslant \alpha \leqslant 1.3$ and $3 \leqslant A \leqslant 5$ provide significant phone error reduction for all the SNR noise conditions ($0 \leqslant \text{SNR} < 30$). For example, the best configuration obtained in the TIMIT + car noise task ($\alpha = 1.3$ and $A = 5$) using AGA was also very efficient in the HTIMIT task. This last configuration almost did not hurt the TIMIT task with free noise conditions (see in Table 3 the cells CMNVS + AGA and CMVS, or MFCC and AGA for the Timit microphone). In all the experiments reported in Figs. 5–7 and Tables 3 and 4, the differences obtained using SS or AGA were statistically significant at 95% confidence level processing 160.000 phones in each experiment of the HTIMIT and the TIMIT + car noise tasks.

Performing the attenuation in the power domain with the AGA algorithm produces nearly the same results using $2.25 \leqslant \alpha \leqslant 2.75$ and $10 \leqslant A \leqslant 30$. The optimal values of α and A in the power and magnitude domains are related approximately by the following equations:

$$\alpha_{\text{magnitude}} = \frac{\alpha_{\text{power}}}{2}, \quad A_{\text{magnitude}} = \sqrt{A_{\text{power}}}. \tag{8}$$

Table 3 compares the maximum compensation obtained in HTIMIT task with different front-end configurations. For this task conventional CMN was compared to the CMNVS algorithm which estimates the mean and the left and right variances (CMNVS). Also AGA (with $\alpha = 1.3$ and $A = 5$)

Table 3
Microphone and noise compensation in HTIMIT task

Mic	MFCC	CMN	CMNVS	MFCC + SS, $A = 3$ $\alpha = 0.7$	MFCC + AGA, $A = 5$ $\alpha = 1.3$	CMNVS + AGA, $A = 5$ $\alpha = 1.3$	MATCH
cb1	56.24	58.7	62.21	57.82	58.25	63.58	63.2
cb2	59.45	63.7	65.36	60.75	61.39	65.67	64.2
cb3	38.28	45.1	50.62	43.13	43.57	51.62	58.8
cb4	41.9	48.0	53.62	47.86	47.92	55.82	59.9
el1	57.23	64.7	65.33	56.12	58.49	65.27	64.3
el2	49.51	59.7	61.09	51.07	50.39	63.05	62.9
el3	50.88	53.3	57.73	52.19	51.82	58.37	60.4
el4	52.42	58.7	61.41	51.53	51.72	62.12	61.9
pt1	36.81	51.6	56.54	44.98	45.72	58.5	61.7
Senh	62.51	62.7	64.68	62.35	63.73	65.39	64.1
Timit	66.56	67.1	66.71	65.52	65.67	66.37	66.71
Average	51.98	57.5	60.45	53.87 ± 0.24	54.42 ± 0.24	61.42 ± 0.23	62.6

Table 4
Noise compensation in TIMIT + car noise task

Car	MFCC	CMNVS	MFCC + SS, $A = 3$ $\alpha = 0.7$	MFCC + AGA, $A = 5$ $\alpha = 1.3$	CMNVS + AGA $A = 5$ $\alpha = 1.3$	MATCH
Est7.5	35.92	48.59	41.32	42.4	49.97	
Imp3.1	25.54	39.59	33.82	34.58	44.46	
Imp1.7	18.66	32.27	27.56	28.01	37.1	
Mar2.5	31.53	45.59	39.35	40.93	49.99	
Mar20	56.7	61.94	59.84	63.04	64.43	
Mar10	51.16	58.5	55.11	56.6	60.68	
Sta0.4	25.19	38.71	31.26	31.98	42.46	
Sta16	49.18	57.91	50.32	52.16	57.72	
Int12	39.62	52.94	48.81	50.7	56.26	60.3
Int6	29.91	44.15	38.77	39.46	48.68	54.6
Int0	20.85	35.17	30.21	29.16	40.04	46.6
Average	34.93	46.85	41.5 ± 0.25	42.63 ± 0.25	50.16 ± 0.25	

provided a positive effect reducing the electrical noises introduced by the poor transducer. Results with conventional Mel-frequency cepstral coefficients (MFCC) and MFCC + SS and MFCC + AGA are also provided. An assessment training and testing in matched conditions (match) was performed using the HTIMIT training corpus available to measure the relative phone error reduction (RPER). In this paper, the RPER is calculated as follows:

RPER = (phone error using compensated MFCC trained in clean condition – phone error using original MFCC trained in clean condition)/(phone error using compensated MFCC (CMNVS–MFCC) trained in noisy condition – phone error using original MFCC trained in clean condition).

All the experiments in matched conditions were done using normalized MFCC (CMNVS) which seems to provide a better accuracy than non-normalized MFCC in clean and noisy conditions. In the clean TIMIT database CMN or CMNVS tend to improve the phone accuracy if a long time constant ($\leqslant 3$ s or $\beta \leqslant 0.997$) is used (see Table 3). In this later case the normalization techniques reduce the speaker variability and have also a positive effect in the phone accuracy.

RPER up to 55% was obtained using CMN, 80% using CMNVS and 90% by adding (CMNVS + AGA).

Table 4 summarizes the system improvements introduced by each technique (SS, AGA, CMNVS and CMNVS + AGA) in the TIMIT database mixed with 11 car noises. In this task also the combination of both techniques CMNVS + AGA (with $\alpha = 1.3$ and $A = 5$) provides very significant and complementary results. An assessment training and testing in matched conditions (match) was performed using the TIMIT database mixed with the Integra car noise at different SNR conditions (0, 6, 12 dB) to estimate the RPER.

RPER up to 77% was obtained combining CMNVS + AGA.

Finally, Table 5 summarizes the sensibility of the CMNVS technique to possible speech detection errors. In the experiment, the speech dependent variables (a_i, ${}^{\mathrm{r}}v_i$, ${}^{1}v_i$) were updated over noisy speech segments (2.8 s on average) and also over car noise areas with variable length (0, 350, 600, 1050, 1800, 3300 ms). In this task, noise provided by the Integra car model was mixed at different SNR (0, 6, 12 dB) and was appended at the beginning of each file of the TIMIT database. In the experiment a real speech detector was not used, and the phonetic labeling information provide in the TIMIT data base was used to update the noisy speech variables (a_i, ${}^{\mathrm{r}}v_i$, ${}^{1}v_i$).

The accuracy of the algorithm remains very stable when the noise areas do not exceed 600 ms but degrades rapidly for noise areas longer than 1 s, which indicates that the speech detection does not

Table 5
Evolution of the phone accuracy updating the speech variables over noisy speech and variable length noise areas appended at the beginning of the TIMIT files

Car\noise	0 ms	350 ms	600 ms	1050 ms	1800 ms	3300 ms
Int12	52.2	52.9	51.8	50.5	48.3	42.2
Int6	44.1	44.1	43.3	37.0	26.3	15.4
Int0	35.0	35.1	34.2	29.9	24.0	18.1
Average	43.8	44.0	43.1	39.1	32.9	25.4

need to be extremely accurate but it is required. If very long noise areas are used to update a_i, ${}^{\mathrm{r}}v_i$, ${}^{1}v_i$, the noisy speech cepstral variables tend to converge to the noise statistics and the algorithm degenerates.

5. Acoustic assessment test

A final acoustic assessment was performed to compare the acoustic quality of the AGA algorithm versus SS. In the experiment, 9 listeners (7 students performing phonetic transcriptions in our facilities and 2 SONY speech engineers) who were not familiar with the present article were asked to evaluate 8 noisy speech files, and also 16 attenuated noisy speech files with SS and AGA. For this task, 4 car noises with the SNR close to 0 dB (Impreza 3.1 dB, Impreza 1.7 dB, Mark2 2.5 dB, Integra 0 dB) were used in order to use α and A factors appropriate for SS. Each listener evaluated 8 different trials using the 4 car noises, and covering a total of 4 TIMIT speakers (2 males and 2 females). In the acoustic test, a total of 72 TIMIT files and 12 speakers were analyzed. The attenuated noisy speech files were produced using the 256-FFT, followed by an attenuation algorithm (SS or AGA) in the magnitude domain. Finally, the inverse 256-FFT was used to reconstruct the audio files. For SS, the values $\alpha = 1.1$ and $A = 3$ were used to attenuate the noise, and in AGA, the values $\alpha = 1.3$ and $A = 5$ was evaluated. In SS the configuration $\alpha = 1.3$ and $A = 5$ was avoided because it introduces too much audible distortion, even if it produces better recognition accuracy with the HMM. The setup in AGA came directly from the HMM recognition scores and was not acoustically tuned up. During the assessment, each listener was asked to reply to the following three questions:

1. Where did you hear more noise? (allowed answers: B, G or “?”)
2. Where did you hear more speech distortions or artifacts? (allowed answers: B, G or “?”)
3. What file did sound more pleasant or natural? (allowed answers: B, G, O or “?”)

where B is the attenuated file with SS, G is the attenuated file with AGA, O is the original noisy file, and “?” means the listener was unsure. The files were randomly presented and the listeners were allowed to hear the different files several times. Table 6 summarizes the total response produced by each listener in the 8 acoustic trials. SS and AGA clearly reduce the additive noise effects, but also in a different way. SS reduces the noise effects more because it uses a common attenuation factor A for all the FFT-bins. On the other hand, the noise effects are more smoothly attenuated using AGA and only the FFT-bins with very low SNR ~ 0 db are heavily attenuated. Globally, AGA seems to introduce less phonetic distortions and less musical tones than SS, leading to a more natural sound.

Table 6
Results obtained in the acoustic assessment tests

Listener	Question 1	Question 2	Question 3
1	G	B	G
2	G	B	O
3	G	B	G
4	G	B	G
5	G	B	O and G
6	G	B	G and B
7	G	B	G
8	G	B	G
9	G	G	B
Total	G	B	G

6. Conclusions

A combination of two low cost front-end techniques has been proposed to obtain a fast and robust speech recognition system. A new cepstral normalization algorithm (CMNVS technique) has been described to minimize non-linear and linear effects introduced by channels and additive noise distortions. Additional improvements were obtained combining the cepstral normalization technique with a new attenuation scheme (AGA) providing a robust front-end in a wide range of SNR conditions ($0\,\mathrm{dB} \leqslant \mathrm{SNR} \leqslant 30\,\mathrm{dB}$). AGA is compared with the SS technique using HMM recognition scores and an acoustic assessment test. The main benefit obtained with AGA is it produces a stable and reliable attenuation algorithm, independent of the SNR. The same algorithmic setup was successfully used in a wide range of SNR and noisy conditions. AGA also statistically improved the HMM recognition scores and acoustically seems to introduce less phonetic distortions and musical tones than SS. The two algorithms (CMNVS and AGA) are really designed for stationary noise conditions and they both require a minimum of 3 s of stationary noisy speech to obtain maximum compensation effects. The two algorithms require also an external speech versus noise discrimination process (VAD, push to talk...) to update the noise and speech variables independently. The real-time (frame synchronous) implementation of the two algorithms (CMNVS and AGA) suited for slow varying (<0.3 Hz) and stationary environmental conditions is also described and used in the paper.

Symbols

Variables

Indexes

k	frequency or filter bank index
t	time or frame index
i	cepstral or differential cepstral coefficient index

Magnitude domain variables

Y_k	magnitude of incoming noisy signal (speech, noise or noisy speech)
Yat_k	magnitude of the attenuated noisy signal
μ_k	average of the noise magnitude
δ_k	standard deviation of the noise magnitude
Sp_k	average of the noisy speech magnitude

Cepstral or deltas-cepstral domain variables

x_i	ith component of the cepstral or differential cepstral coefficient
ns_i	ith component of the normalized and scaled cepstral or delta-cepstral coefficient
a_i	average of ith cepstral coefficient
${}^{\mathrm{r}}v_i$	right variance of the ith cepstral or delta-cepstral coefficient
${}^{\mathrm{l}}v_i$	left variance of the ith cepstral or delta-cepstral coefficient

Coefficients

A	attenuation coefficient
α	overestimation coefficient
β	speech forgetting coefficient
λ	noise forgetting coefficient

References

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by additive noise. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 849–852.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-27 (2), 113–120.

Chen, R., Tanaka, M., Wu, D., Olorenshaw, L., Menéndez-Pidal, X., 1998. Improvements on very large vocabulary word recognition. In: Proceedings of the 8th SONY Research Forum, pp. 3–8.

Gales, M.J.F., 1998. Predictive model-based compensation schemes for robust speech recognition. Speech Communication 25 (1), 49–74.

Gauvian, J.L., Gangaolf, J.J., Lamel, L., 1996. Speech recognition for an information Kiosk. In: Proceedings of the International of Spoken Language Processing, pp. 849–853.

Hanson, B., Applebaum, T., Junqua, J., 1995. Spectral dynamics for speech recognition under adverse conditions. In: Automatic Speech & Speaker Recognition Advanced Topics. Kluwer Academic Publishers, Dordrecht.

Iwahashi, N., Pao, H., Honda, H., Minamino, K., Omote, M., 1998. Stochastic features for robust speech recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 633–636.

Lee, K.F., Hon, H.W., 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Trans. ASSP 37, 1641–1648.

Lockwood, P., Boudy, J., 1992. Experiments with NSS, HMM, & projection for robust SR in cars. Speech Communication 11, 215–228.

Menéndez-Pidal, X., Chen, R., Wu, D., Tanaka, M., 1999. Compensation of channel and noise distortions combining normalization & speech enhancement techniques. In: Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp. 101–105.

Milner, B., 1996. Inclusion of temporal information into feature for speech recognition. In: Proceedings of the International of Spoken Language Processing, pp. 256–259.

Neumeyer, L., Digalakis, V., Weintraub, M., 1994. Training issues and channel equalization techniques for construction of telephone acoustic models using high-quality speech corpus. IEEE Trans. Speech Audio Process. 2 (4), 590–597.

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

Reynolds, D.H., 1997. HTIMIT and LLHDB: speech corpora for the study of hand set transducer effects. In: Proceedings of the International Acoustics, Speech and Signal Processing, pp. 1537–1538.

Rosenberg, A., Lee, C.-H, Soong, L., 1994. Cepstral channel normalization techniques for HMM-based speaker verification. In: Proceedings of the International of Spoken Language Processing, pp. 1835–1838.

Schless, V., Class, F., 1998. SNR-dependent flooring and noise overestimation for joint application of spectral subtraction and model combination. In: Proceedings of the International of Spoken Language Processing, pp. 1495–1498.

Tibrewala, S., Hermansky, H., 1997. Multi-band & adaptation approaches to robust speech recognition. In: Proc. Eurospeech'97, pp. 2619–2622.

Verdu, S., 1998. Fifty years of Shannon theory. IEEE Trans. Information Theory 44 (6), 2057–2078.

Viikki, O., Laurila, O., 1997. Noise Robust HMM-based speech recognition using segmental cepstral feature vector normalization. In: ESCA-NATO Workshop in Robust Speech Recognition for unknown communication channels, pp. 107–110.

Viikki, O., Bye, D., Laurila, K., 1998. A recursive feature vector normalization approach for robust speech recognition in noise. In: Proceedings of the International Acoustics, Speech and Signal Processing, pp. 733–737.

Xie, F., Van Compernolle, D., 1994. A family of MLP based nonlinear spectral estimators for noise reduction. In: Proceedings of the International Acoustics, Speech and Signal Processing, pp. 53–56.

Xie, F., Von Campernolle, D., 1996. Speech enhancement by spectral magnitude estimation, a unifying approach. Speech Communication 19 (2), 89–104.