SENSITIVITY ANALYSIS OF NOISE ROBUSTNESS METHODS

Luca Brayda[1,2], *Luca Rigazio*[1], *Robert Boman*[1] *and Jean-Claude Junqua*[1]

[1] Panasonic Speech Technology Laboratory
3888 State Street, Suite 202,
Santa Barbara, CA 93105, U.S.A
{rigazio, bboman, jcj}@research.panasonic.com

[2] Institut Eurécom
2229 Route des Crêtes, B.P. 193
06904 Sophia-Antipolis Cedex, France
brayda@eurecom.fr

ABSTRACT

This work addresses the problem of noise robustness from the standpoint of the sensitivity to noise estimation errors. Since the noise is usually estimated in the power-spectral domain, we show that the implied error in the cepstral domain has interesting properties. These properties allow us to compare two key methods used in noise robust speech recognition: spectral subtraction and parallel model combination. We show that parallel model combination has an advantage over spectral subtraction because it is less sensitive to noise estimation errors. Experimental results on the Aurora2 database confirm our theoretical findings, with parallel model combination clearly outperforming spectral subtraction and other well-known signal-based robustness methods. Our Aurora2 results with parallel model combination, a basic MFCC front-end and a simple noise estimation are close to the best results obtained on this database with very complex signal processing schemes.

1. INTRODUCTION

This work addresses the problem of noise robustness in Automatic Speech Recognition (ASR), and presents a theoretical investigation of two main approaches used to deal with additive noise: spectral subtraction (SS), which operates in the signal domain, and parallel model combination (PMC), which operates in the acoustic model domain. We propose a square-error analysis of spectral subtraction [1] and parallel model combination [2] for cepstral features to investigate the best domain to operate in when dealing with additive noise. The square-error analysis is interesting because it is closely related to the likelihood measure used in HMM-based ASR and because an analytic solution can be found for cepstral features. The starting point of this research is the fact that any noise robustness method that relies on noise estimation is prone to be sensitive to noise estimation errors. In particular, most methods estimate the noise in the power spectral domain, but the final feature space used for pattern recognition is the cepstral domain. Hence, cepstral domain estimation errors will most directly influence ASR performance. We will study the effects that a noise estimation error in the power spectral domain has on a cepstral domain feature. Also we will show that, because of the nonlinearity of the log-compression, the implied error in the cepstral domain has interesting properties that explain the superiority of PMC over spectral subtraction. Finally, experimental results on the ETSI Aurora2 database [3] show that PMC is superior to spectral subtraction, thus confirming our theoretical findings.

2. BACKGROUND

The ongoing discussion about noise robustness methods has yet to provide a clear answer as to whether or not model-based approaches are superior to signal-based approaches. The (approximate) maximum likelihood solution devised in PMC [2] would hint at a theoretical advantage of PMC over methods not based on maximum likelihood. However, there is no theoretical study that compares PMC to methods operating in the signal domain. Nor we are aware of experimental work that would clarify this matter. For instance, the best Aurora2 systems have complex signal-based robustness components [4, 5]. One can argue that this is because the Aurora2 evaluation requires models to be trained in both clean and noisy conditions (noisy-condition models are not suitable for PMC, which is designed for clean-condition acoustic models). However, it has been recently shown that a modified version of PMC can outperform signal-based methods even for noisy-condition models [6]. For all these reasons we felt compelled in looking for a comparative theoretical investigation of PMC and spectral subtraction.

3. NOISE SENSITIVITY ANALYSIS

Let $C(X) = F\log(X)$ be the cepstral operator, where $X = S+N$ is the power spectral vector of the noisy speech input [1], F is a linear transformation such as the DCT and $\log(X)$ is the vector of the element logarithms $\log(x_k)$. For simplicity we also assume that F is an orthonormal matrix, i.e. $F^TF = I$. Let $S, \hat{S}, N, \hat{N}$ be respectively the clean speech, the estimated clean speech, the noise, and the estimated noise power spectra. Also define $s, \hat{s}, n, \hat{n}$ as one component of $S, \hat{S}, N, \hat{N}$ respectively. The basic form of spectral subtraction analyzed here obtains an estimate of the clean features by removing the estimated noise from the noisy speech input in the power spectral domain, while guaranteeing the positivity of the estimated clean speech power spectrum:

[1] In this study we ignore the effect of the cross correlation term $|S||N|$ in the power spectrum.

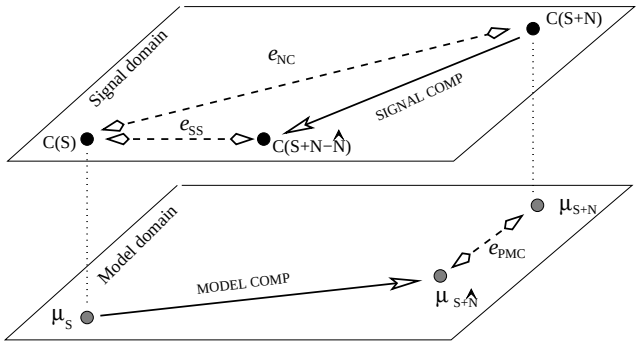

Fig. 1. Compensation spaces and corresponding estimation errors.

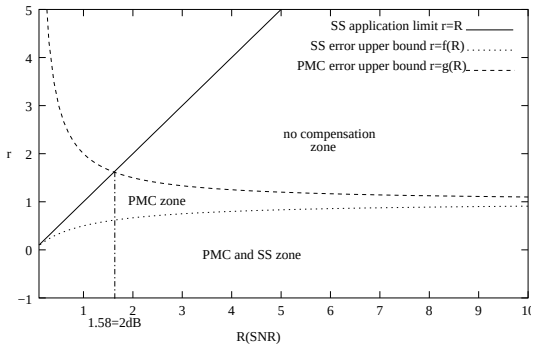

Fig. 2. Minimum square-error regions in the $[R, r]$ plane.

$$\begin{cases} \hat{s} = x - \hat{n} \text{ if } x - \hat{n} > \alpha\hat{n} \\ \hat{s} = \alpha\hat{n} \text{ otherwise} \end{cases} \quad (1)$$

Notice that the estimation error affecting recognition is related to the distance between the unknown clean cepstrum and the estimated clean cepstrum. For simplicity we will assume the flooring constant $\alpha = 0$ in our formulation. It can be shown that this assumption does not restrict the generality of our conclusions as the following formulation can be extended to any value of α.

In the case of PMC the noisy model mean $\mu_X = E\{C(X)\}$ is computed from the clean model mean in the cepstral domain as:

$$\mu_{S+\hat{N}} = C(C^{-1}(\mu_S) + \hat{N}). \quad (2)$$

Here the estimation error affecting recognition is related to the distance between the unknown noisy speech model mean and the estimated noisy speech model mean.

Figure 1 shows how estimation errors are related to compensation methods in both signal and model space.

Based on the previous observations we can define the square-error for no compensation, spectral subtraction and PMC as follows:

$$E^2_{NC} = ||C(S+N) - C(S)||^2, \quad (3)$$

$$E^2_{SS} = ||C(S) - C(\hat{S})||^2, \quad (4)$$

$$E^2_{PMC} = ||C(S+N) - C(S+\hat{N})||^2, \quad (5)$$

In the definition of E_{PMC} we have used $C(S+\hat{N})$ for $\mu_{S+\hat{N}}$ for notational convenience. Notice that all the errors can be expressed as:

$$\begin{aligned} E^2 &= ||C(A) - C(B)||^2 = \quad (6) \\ &= ||F\log(A) - F\log(B)||^2 = \\ &= ||F\log(A/B)||^2 = ||\log(A/B)||^2 = \\ &= \sum_k \log^2(a_k/b_k) = \sum_k e_k^2, \end{aligned}$$

where A/B is the vector of the components ratios a_k/b_k and e_k^2 is the square-error of the component k. Thanks to this we can restrict our analysis to each term e_k^2 of the sum in (6), thus reducing the study to a one-dimensional problem. Specifically, this is allowed by the property of the logarithm that transforms sums into products; this makes the following formulation difficult to extend to generalized cepstrum based on other compression functions, such as the root-cepstrum [7], because no closed form solution can be found.

Let $\varepsilon = \hat{n} - n$ be the noise estimation error for one vector component. Notice that to respect the positivity of $\hat{s}, \hat{n}$ we have the constraint $-n < \varepsilon < s$. We define the quantities $r = \varepsilon/n$, which represents the relative noise estimation error, and $R = s/n$, which represents the signal to noise ratio (SNR). This allows us to rewrite the component square-errors as:

$$e^2_{NC} = \log^2\left(1 + \frac{1}{R}\right), \quad (7)$$

$$e^2_{SS} = \log^2\left(1 - \frac{r}{R}\right), \quad (8)$$

$$e^2_{PMC} = \log^2\left(1 + \frac{r}{R+1}\right), \quad (9)$$

under the constraints:

$$\begin{cases} R > 0 \\ -1 < r < R, \end{cases} \quad (10)$$

Notice that the constraint related to the existence of the logarithm ($r < R$) concerns only spectral subtraction. We can now compare the square-error for spectral subtraction and for the no-compensation baseline by solving the following inequality:

$$e^2_{SS} < e^2_{NC}. \quad (11)$$

When (11) is satisfied, spectral subtraction will perform better than no compensation. We get the condition:

$$r < \frac{R}{R+1} = f(R). \quad (12)$$

Intuitively in (12), the lower the noise estimation error is, the more likely spectral subtraction is to be better than no compensation. Then we compare PMC with no compensation by solving:

$$e^2_{PMC} < e^2_{NC}. \quad (13)$$

We get the condition:

$$r < \frac{R+1}{R} = g(R). \quad (14)$$

Compensation	A	B	C	Avg.
NONE	38.38	40.00	39.01	39.15
SS	44.07	44.99	47.00	45.02
PMC	80.68	78.82	77.57	79.32

Table 1. Comparison of spectral subtraction and PMC digit accuracies on Aurora2 (clean training) for MFCC static coefficients.

As before, the lower the noise estimation error is, the more likely PMC is to be better than the baseline. However notice that $g = 1/f$ and that $f < 1$, which implies $g > f$ always. This is important because it implies that the region in which it is worth using PMC includes the region in which it is worth using spectral subtraction. Indeed if we compare PMC with spectral subtraction square-errors by solving:

$$e_{PMC}^2 < e_{SS}^2, \tag{15}$$

we have that (15) is always satisfied. This means that PMC is always better than spectral subtraction.

Figure 2 shows the regions of the $[R, r]$ plane where the different compensation methods perform best in a square-error sense. Three main regions are obtained: below $f(R)$ it is worthwhile using spectral subtraction and below $g(R)$ it is worthwhile using PMC. Above these curves, both spectral subtraction and PMC are useless, because the resulting square-error is larger than the initial square-error obtained with no compensation. Notice that spectral subtraction cannot be applied for $r > R$ (because the argument of the logarithm will be negative) but PMC can be applied even for $r > R$. Also notice that in the spectral subtraction zone, which is included in the PMC zone, the PMC square-error is always lower than the spectral subtraction square-error. This can be summarized in the following statements:

- PMC is always better than spectral subtraction.
- The spectral subtraction region decreases when the SNR decreases.
- The PMC region increases when the SNR decreases.
- For SNRs lower than 2dB the PMC region increases very quickly whereas the spectral subtraction region decreases very quickly
- For large noise over-estimation errors no compensation method is effective.
- If the noise is under-estimated both spectral subtraction and PMC are more likely to outperform the baseline.

We will see that these theoretical results are confirmed by our experiments.

4. EXPERIMENTAL RESULTS

Experiments are conducted on the ETSI Aurora2 database [3], used for standardization of the front-end analysis for distributed speech recognition in noisy environments. It consists of US-English digits in presence of synthetically added noise and channel distortion. Performance is measured on three test sets containing noises already seen in training (set A), different from those seen in training (set B) and different from those seen in training plus an additional channel distortion (set C).

SNR	SS	PMC
clean	97.45	97.67
20 dB	78.12	94.47
15 dB	63.06	91.38
10 dB	44.51	85.42
5 dB	25.79	73.38
0 dB	13.63	51.93
-5 dB	8.97	24.82
Avg.	45.02	79.32

Table 2. SNR breakdown of spectral subtraction and PMC results for static coefficients.

Our baseline system is based on standard MFCC features. The speech signal is represented with 13 MFCC coefficients computed over a window of 20ms with a frame rate of 100Hz. Depending on the experiment, first and second derivatives are added. Cepstral mean normalization (CMN) is not used when comparing spectral subtraction and PMC. Both spectral subtraction and PMC rely on the same noise estimate computed as the average filter-bank energy over the first 250ms of each utterance. For spectral subtraction we use a flooring constant $\alpha = 0.3$. Also neither a voice activity detector nor frame dropping are used. Acoustic models are based on whole-word models, with 8 states and 8 gaussians per state, trained in clean conditions. To validate our theoretical findings we start by testing the performance of static cepstrum only with no compensation, with spectral subtraction and with PMC (Table 1). Results are reported in digit accuracy. A small improvement over the baseline can be observed for spectral subtraction, whereas with PMC the performance is drastically improved. Also notice that PMC is superior to spectral subtraction at all SNR, and it is particularly good at low SNR (Table 2). This confirms our theoretical finding that PMC is always superior to spectral subtraction, but also that the region in which PMC performs better than spectral subtraction widens as the SNR decreases (see Figure 2). Also notice that the result obtained with PMC (79.32%) is very high for a very simple front-end based on static features only.

To improve the significance of our study, we test the performance with dynamic coefficients. Table 3 shows results for static MFCC plus first derivatives: both baseline and spectral subtraction results improve but are still far from PMC. A small degradation is observed for PMC (compared to the PMC results with static features only) if only the static features are compensated (PMC(s)) but a clear improvement is observed when both static and dynamic features are compensated (PMC(s,d)) as indicated in [2]. Finally table 4 reports the performance with first and second derivatives MFCC with PMC applied to all coefficients. The compensation of the second derivatives is performed by disregarding all terms that depend on second order statistics, which are difficult to estimate, as indicated in [6]. We notice that results for both the sim-

Compensation	A	B	C	Avg.
NONE	47.29	42.57	59.58	47.86
SS	58.38	54.58	69.94	59.17
PMC(s)	77.83	79.38	76.67	78.22
PMC(s,d)	83.54	83.86	83.05	83.57

Table 3. Comparison of spectral subtraction and PMC results for static plus first derivatives.

ple back-end (8 gaussians per mixture) and the complex back-end (16 gaussians per mixture) are close to the state of the art [4, 5]. This is surprising, since the noise estimation and the front-end used are very simple, and because neither a voice activity detector nor frame-dropping is used.

Compensation	A	B	C	Avg.
PMC(s,d,dd), 8G	86.90	86.29	86.33	86.54
PMC(s,d,dd), 16G	87.81	86.55	87.72	87.29

Table 4. PMC results for static plus first and second derivatives, for 8 and 16 gaussians per mixture.

5. DISCUSSION

Strictly speaking, our noise sensitivity analysis applies only to spectral subtraction and PMC for static cepstral features based on the log-compression function. Indeed the logarithm makes it possible to compare cepstral distances in the $[R, r]$ plane, thanks to a closed form solution which would be difficult to derive for other compression functions. Our theoretical and experimental results show that PMC is superior to spectral subtraction in dealing with additive noise. However, it is our intuition that a general form of the analysis may be derived to show that operating in the model space provides lower sensitivity to noise estimation error than operating in the signal space, independently of the specific method used. This is unfortunately far from being formally proved. However, we tested other well-known techniques for signal-based compensation, in the same settings used to compare spectral subtraction and PMC, to see if we could find some signal-based method that could outperform PMC.

Table 5 reports results for minimum-square-error filtering (MMSE) [8], and for codebook dependent cepstral normalization (CDCN) [9]. To model the clean speech in CDCN we use a 64 gaussian components mixture model trained on the same Aurora2 clean training set. We also combine the previous methods with cepstral mean normalization (CMN) (which always provides good improvements but cannot be used with PMC). The best results for signal-based methods are obtained when combining CDCN with CMN. However notice that CDCN/CMN results still lag behind PMC results by 12% absolute. Also notice that CDCN is an hybrid model and signal compensation method. This is just another indication that model-based methods have a performance advantage over signal-based methods.

Compensation	NONE	CMN
NONE	47.86	69.12
MMSE	53.17	67.19
SS	62.96	71.48
CDCN	72.90	74.15

Table 5. Average results over test conditions for different signal-based techniques for static plus first and second derivatives.

6. CONCLUSION

We presented a noise sensitivity analysis of spectral subtraction and PMC. Results from this analysis indicate that PMC is always superior to spectral subtraction because it is less sensitive to noise estimation errors. This is a direct consequence of the properties of the log-compression used in the cepstrum computation. Experimental results on Aurora2 confirm our theoretical findings. Our Aurora2 results with PMC using simple MFCC front-end and noise estimates are close to the best results obtained on this database with very complex signal processing schemes.

7. REFERENCES

[1] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. ASSP 27*, pp. 113–120, 1979.

[2] M.J.F. Gales, *Model-based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, Cambridge University, 1995.

[3] D. Pearce, "Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front-Ends," in *AVIOS*, 2000.

[4] A. Adami et al., "Qualcomm-ICSI-OGI Features for ASR," in *Proc. of ICSLP*, 2002, pp. 21–24.

[5] D. Macho et al., "Evaluation of a Noise-robust Front-end on Aurora Databases," in *Proc. of ICSLP*, 2002, pp. 17–20.

[6] L. Rigazio, P. Nguyen, D. Kryze, and J.C. Junqua, "Large Vocabulary Noise Robustness on Aurora4," in *Proc. of Eurospeech*, 2003, pp. 345–348.

[7] P. Alexandre and P. Lockwood, "Root Cepstral Analysis: A Unified View," *Speech Communication*, vol. 3, pp. 277–288, July 1993.

[8] Y. Ephraim and D. Malah, "Speech Enhancement using a Minimum Mean-Square Error Short-time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics,Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

[9] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," in *Proc. of ICASSP*, 1990.