Individual On-Line Variance Adaptation of Frequency Filtered Parameters for Robust ASR

Jesús Vicente-Peña[1], *Fernando Díaz-de-María*[1], *Bastiaan Kleijn*[2]

[1]Department of Signal Processing and Communications
Universidad Carlos III de Madrid, Leganés, Spain
{jvicente, fdiaz}@tsc.uc3m.es
[2]Sound and Image Processing Lab.
KTH (Royal Institute of Technology), Stockholm, Sweden
bastiaan.kleijn@ee.kth.se

Abstract

In this paper we address the problem of robust speech recognition. We propose a new method based on the individual variance adaptation of frequency filtered parameters to reduce the deleterious effects of additive narrow-band noise. The method can be interpreted as a spectral weighting that assigns increased importance to the most reliable spectral components, typically the spectral peaks. The experiments confirm that the suggested method results in significantly improved recognition rates for additive narrow-band noise.

Index Terms: additive noise ASR, frequency filtered parameters, spectral subtraction, model variance adaptation, spectral weighting.

1. Introduction

State-of-the-art Automatic Speech Recognition (ASR) systems can achieve high recognition rates in distortion-free environments. However, as is well known, in real environments various distortions lead to a significant degradation in recognition performance. To counter this degradation, much effort has been spent to improve the robustness of ASR systems, particularly against the effect of convolutive and additive distortions. A classic review can be found in [1].

In this paper we focus on the reduction of the effect of additive distortion. Perhaps the best-known method to reduce the effects of additive distortions is spectral subtraction [2]. It consists of subtracting an estimate of the noise spectrum from the noisy speech spectrum, achieving good results for stationary and wide-band additive noises.

Other techniques try to reduce the noise sensitivity of the statistics of the parameters. Segmental feature vector normalization [3], also called Mean Variance Normalization (MVN), normalizes the mean and variance of the parameters to zero and one, respectively. Histogram normalization [4] or histogram equalization [5] aim at normalizing all moments of the distribution.

Instead of pursuing a robust parameterization, other proposals aim at adapting the acoustic unit models to the acoustic environment. One of the most popular methods is PMC (Parallel Model Combination) [6], which adapts the mean and variance of the models to compensate for the effect of additive distortion. In contrast to the techniques discussed in this paper, PMC was conceived to work either sporadically when significant changes in the noise statistics are detected or simply off-line.

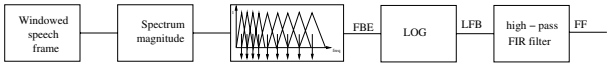

Figure 1: Frequency Filtered Parameterization Scheme.

In this paper we use a parameterization called Frequency Filtered (FF) [7], which achieves results similar to the well-known Mel Frequency Cepstral Coefficients (MFCCs) but operates in the log spectrum domain. Figure 1 summarizes by means of a block diagram the steps involved in FF computation. As can be seen in this figure, the substitution of the DCT (Discrete Cosine Transform) by a high-pass filter is the main difference with respect to the MFCC parameterization. The high-pass filter used in our experiments involves only two log filter bank energies: $z - z^{-1}$ [7].

The main motivation for using the FF parameters is that they have a simple relation to the log spectrum, facilitating simple analytical approximations for their mean and variance that incorporate the effect of additive noise. As we show in Section 2, the use of spectral subtraction can compensate for the bias in the FF parameters resulting from additive noise but it can not compensate for the bias in the variance. Therefore, we propose a new method to perform on-line variance adaptation of the FF parameters.

We show that the new variance adaptation approach can be interpreted as a spectral weighting that explains the strong recognition-rate improvements observed for narrow-band additive noise. The weighting consists of two terms. The first deemphasizes the least reliable components of the feature vector. The second penalizes those models that, due to the variance adaptation (widening), become competitive far from their normal range of activity (around their means). The method has similarities to methods based on missing features [8] where the input time frequency features are usually classified as reliable or unreliable features and recognition is based on the reliable ones. However, in our method the need for a classifier is avoided by the usage of spectral weighting.

To summarize, we propose a method for mean and variance adaptation for the FF parameterization. An approximate analytical expression is found for the mean and variance of the noisy FF parameters. The mean is adapted by spectral subtraction and a novel

procedure is proposed to adapt the variance. In contrast to MVN, which just normalizes mean and variance of the feature vectors, the proposed approach adapts the models depending on the noise spectrum. Compared to PMC, the proposed variance adaptation is much more dynamic. As is show in section 4, the new variance adaptation significantly improves recognition rates over spectral subtraction in the presence of narrow-band additive noise.

This paper is organized as follows. Section 2 explains the proposed method to compensate for the mean and variance of the noisy FF parameters. Section 3 gives an interpretation of the proposed method as a spectral weighting. Experiments and results are presented in section 4. Finally, conclusions are outlined in Section 5.

2. The proposed method: spectral subtraction and variance adaptation of the frequency filtered (FF) parameters

One of the long-standing problems in the field of ASR is the mismatch between training and testing conditions. We use spectral subtraction followed by a simple and novel variance compensation of the FF parameters for adaptive compensation of the model parameters. In this section we develop the required formulation to obtain the mean and variance of the noisy FF parameters for each frequency band, in order to proceed with the corresponding adaptation.

We assume that the speech signal is contaminated with additive and uncorrelated noise. This implies that the noise and speech components are also additive in the magnitude spectrum domain. Since the filter bank energies are just a linear combination of the frequency magnitude spectrum components, the additive property is retained also in that domain. Therefore, we have that

$$FBE_{kn} = FBE_k + n_k, \tag{1}$$

where k denotes the kth filter bank component, FBE_{kn} the Filter Bank Energy of the noisy speech; FBE_k the Filter Bank Energy of the clean speech and n_k the additive noise. Let us furthermore assume that each noise component, n_k, is a random variable with a Gaussian distribution with mean μ_{n_k} and variance $\sigma^2_{n_k}$. We make the reasonable approximation that the noisy filter bank energies are uncorrelated with each other.

Spectral subtraction removes the noise spectrum estimate from the noisy speech spectrum. As a result, the noisy speech filter bank energies after applying spectral subtraction are given by

$$FBE^{ss}_{kn} = FBE_{kn} - \mu_{n_k} = FBE_k + n_k - \mu_{n_k}. \tag{2}$$

For the sake of clarity, we drop the super-index ss in the remainder of this paper. We refer to the filter bank energies after the spectral subtraction as FBE_{kn}.

Next, the log filter bank energies are computed:

$$LFB_{kn} = \log(FBE_{kn}) = \log(FBE_k + n_k - \mu_{n_k}). \tag{3}$$

The first order Taylor series expansion of the log operator around a certain point a is used to obtain a linear approximation of the log filter bank energies. Therefore, the noisy and clean terms are written as

$$LFB_{kn} \approx \log(a) + \frac{FBE_k}{a} - 1 + \frac{n_k - \mu_{n_k}}{a} \tag{4}$$

$$LFB_k \approx \log(a) + \frac{FBE_k}{a} - 1. \tag{5}$$

Combining equations (4) and (5) we then obtain

$$LFB_{kn} \approx LFB_k + \frac{n_k - \mu_{n_k}}{a}. \tag{6}$$

To make this approximation accurate, the point a should be close to $FBE_k + n_k - \mu_{n_k}$ and FBE_k. The former value is just the spectral subtraction-based estimation of the latter, which is unknown. Therefore, we select $a = FBE_k + n_k - \mu_{n_k}$. It is worth noting that (see eq. (6)) the amount of noise at the k'th log filter bank energy is inversely proportional to a, i.e., to FBE_k. As a result, the high-energy bands (spectral peaks) are less sensitive than the low-energy bands (spectral valleys). This fact is due to the log operator: for high energies, where the derivative of the log is quite small, the ear is less sensitive to changes in the power spectrum than for low energies, where the derivative of the log is higher.

Once we have written FBE_{kn} as a function of FBE_k, it is easy to write the noisy FF parameters as a function of the clean FF parameters to compute the new mean and variance. The FF coefficient for the kth log filter bank energy is then

$$\begin{aligned} FF_{kn} &= LFB_{(k+1)n} - LFB_{(k-1)n} \approx \\ &\approx LFB_{k+1} + \frac{n_{k+1} - \mu_{n_{k+1}}}{a} - \\ &- LFB_{k-1} - \frac{n_{k-1} - \mu_{n_{k-1}}}{b}, \end{aligned} \tag{7}$$

with $a = FBE_{k+1} + n_{k+1} - \mu_{n_{k+1}}$ and $b = FBE_{k-1} + n_{k-1} - \mu_{n_{k-1}}$. Given that $FF_k = LFB_{(k+1)n} - LFB_{(k-1)n}$, equation (7) can be rewritten as

$$FF_{kn} \approx FF_k + \frac{n_{k+1} - \mu_{n_{k+1}}}{a} - \frac{n_{k-1} - \mu_{n_{k-1}}}{b}. \tag{8}$$

Assuming that the clean FF parameters are uncorrelated to the corresponding noise component and the noise components are uncorrelated among each other, we can compute the mean and the variance of noisy FF parameters:

$$\mu_{FF_{kn}} = E\{FF_{kn}\} = \mu_{FF_k} \tag{9}$$

$$\begin{aligned} \sigma^2_{FF_{kn}} &= E\{(FF_{kn} - \mu_{FF_{kn}})^2)\} = \\ &= \sigma^2_{FF_k} + \frac{\sigma^2_{n_{k+1}}}{a^2} + \frac{\sigma^2_{n_{k-1}}}{b^2}, \end{aligned} \tag{10}$$

where $\mu_{FF_{kn}}$ and $\sigma^2_{FF_{kn}}$ are the mean and variance of the noisy FF parameters; μ_{FF_k} and $\sigma^2_{FF_n}$ are the mean and the variance of the clean FF parameters; and $\mu_{n_{k-1}}$, $\sigma^2_{n_{k-1}}$ and $\mu_{n_{k+1}}$, $\sigma^2_{n_{k-1}}$ are the mean and the variance of the $(k-1)$'th and $(k+1)$'th noise components, respectively.

It is worth noting that spectral subtraction is an effective method with respect to the mean, since $\mu_{FF_{kn}} = \mu_{FF_k}$. However, the variance of the noisy FF parameters is no longer equal to the clean variance and, as a result, variance adaptation is needed to suitably represent the noisy FF features. Since μ_{FF_k} and $\sigma^2_{FF_k}$ can be estimated easily from the trained Hidden Markov Models (HMMs), we only have to estimate $\sigma^2_{n_k}$ for every k.

Once we know how to compensate the static FF parameters it is straightforward to compensate the delta parameters.

3. Interpretation of variance adaptation as a spectral weighting method

A speech recognizer estimates the acoustic unit that has been uttered by computing the maximum likelihood along all the possible

acoustic models (see [9] for more details):

$$\lambda = \arg_i \max_i \Bigg(\log(a^i_{x_o x_1}) + \sum_{t=1}^{T} \Bigg[-\log\sqrt{(2\pi)^N |\mathbf{\Sigma}^i_{x_t}|} - \\ -\frac{1}{2}(\mathbf{FF}_t - \boldsymbol{\mu}^i_{x_t})^T (\mathbf{\Sigma}^i_{x_t})^{-1} (\mathbf{FF}_t - \boldsymbol{\mu}^i_{x_t}) + \\ + \log(a^i_{x_t x_{t+1}}) \Bigg] + \log(P(\lambda_i)) \Bigg), \quad (11)$$

where λ_i refers to the i^{th} acoustic model; $a^i_{x_t x_{t+1}}$ is the transition probability between the states x_t and x_{t+1} for the model i; N is the number of components of the feature vector; $\mathbf{\Sigma}^i_{x_t}$ and $\boldsymbol{\mu}^i_{x_t}$ are the covariance matrix and the mean vector for the model i and state x_t; $\mathbf{FF}_t$ is the observation vector (***Frecuency Filtered***) at the instant t; and finally, $P(\lambda_i)$ is the probability of the model λ_i. Here we consider only single-Gaussian models but the expressions are easily generalized to the mixture-Gaussian models case.

If we focus on the terms in equation (11) that have to do with the emission probability in each state, suppressing the time and model indexes and considering diagonal covariance matrices we obtain:

$$S = -\frac{1}{2} \left\{ \sum_{k=1}^{N} \log(2\pi\sigma^2_{FF_k}) + \sum_{k=1}^{N} \frac{(FF_k - \mu_{FF_k})^2}{\sigma^2_{FF_k}} \right\} \quad (12)$$

Adapting $\sigma^2_{FF_k}$ in this last equation as previously determined (eq. (10)) and rewriting the result in a more convenient way

$$S = -\frac{1}{2} \left\{ \sum_{k=1}^{N} \log \left(2\pi \frac{\sigma^2_{FF_k} + \frac{\sigma^2_{n_{k+1}}}{a^2} + \frac{\sigma^2_{n_{k-1}}}{b^2}}{\sigma^2_{FF_k}} \sigma^2_{FF_k} \right) + \\ + \sum_{k=1}^{N} \frac{\sigma^2_{FF_k}}{\sigma^2_{FF_k} + \frac{\sigma^2_{n_{k+1}}}{a^2} + \frac{\sigma^2_{n_{k-1}}}{b^2}} \frac{(FF_{kn} - \mu_{FF_k})^2}{\sigma^2_{FF_k}} \right\} \quad (13)$$

Next we introduce the notation

$$w_k = \frac{\sigma^2_{FF_k}}{\left(\sigma^2_{FF_k} + \frac{\sigma^2_{n_{k+1}}}{a^2} + \frac{\sigma^2_{n_{k-1}}}{b^2} \right)}. \quad (14)$$

and rewrite equation (13) as

$$S = -\frac{1}{2} \left\{ \sum_{k=1}^{N} \log\left(2\pi\sigma^2_{FF_k}\right) + \\ + \sum_{k=1}^{N} w_k \frac{(FF_{kn} - \mu_{FF_k})^2}{\sigma^2_{FF_k}} - \sum_{k=1}^{N} \log w_k \right\}. \quad (15)$$

Comparing this last equation with the original for clean features (eq. (12)), two differences become evident:

- The term
$$\sum_{k=1}^{N} \frac{(FF_k - \mu_{FF_k})^2}{\sigma^2_{FF_k}} \quad (16)$$
for clean features turns into
$$\sum_{k=1}^{N} w_k \frac{(FF_{kn} - \mu_{FF_k})^2}{\sigma^2_{FF_k}} \quad (17)$$
for noise features. This term is a normalized Euclidean distance that indicates how far or close is the observation from the model represented by $(\mu_{FF_k}, \sigma^2_{FF_k})$.
We can see the weights w_k given by eq. (14) as a measure of the noise level in our input features. Therefore, when the noisy term in our variance, $\sigma^2_{n_{k+1}}/a^2 + \sigma^2_{n_{k-1}}/b^2$, is relatively small, we find weights close to one. In contrast, the weights are close to zero when the noisy term is relatively large. Normally, weights close to one come from high-energy regions in the log-filter bank energy domain and, therefore, eq. (17) is dominated by the spectral peaks instead of by the valleys. It is also important to note that the weights depend on the variance of the model in such a way that models with larger variances are less sensitive to noise distortions.
- The second difference consists in the addition of the term
$$-\sum_{k=1}^{N} \log w_k. \quad (18)$$
The problem when we weight the Euclidean distance as we do in eq. (17) is that it is close to zero if most of the weights are low. A distance close to zero indicates a perfect matching between the current model and the observation. The term defined by eq. (18) adds a penalty for low weights. It is worth noting that this term vanishes when all weights are equal to one, that is, when there is no noise.

4. Experiments and Results

4.1. Database

The database employed in our experiments was the well-known Resource Management RM1 Database [10], which has a vocabulary of 991 words. The training corpus consisted of 3990 sentences and the test set contains 1200 sentences, which corresponds to a compilation of the first four official test sets. We used a downsampled version (at 8 KHz) of the database (originally recorded at 16 kHz in clean conditions).

4.2. Recognition System

The back-end was based on HMMs. The HTK toolkit [9] was used to build the system. Context-dependent acoustic models in the form of cross-word triphones were used. A three-state, three-mixture per state model was used to represent each triphone. The models were estimated using clean speech. The standard word-pair grammar was used as the language model. Twelve frequency filtered coefficients plus the log-energy and their delta parameters were used as front-end in the recognizer.

4.3. Noise types

To implement the proposed method, an estimate of the mean and the variance of the noise components in the filter bank domain is needed. To render the results independent from the noise estimation method, we performed experiments with oracle knowledge of the noise statistics. We calculated the mean and variance of the noise component at the filter bank domain using the noise signal that was added to each sentence. In this oracle estimation there is an implicit assumption of stationarity. For this reason, the experiments were limited to noises that could be considered stationary.

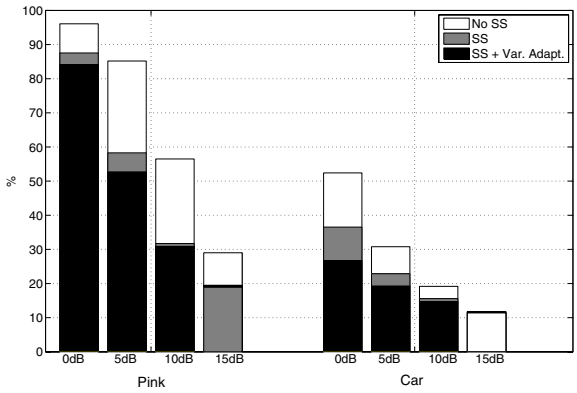

Figure 2: Word Error Rate (WER) for pink and car noises and four SNRs. The white bar is displayed for reference and corresponds to the recognizer without any robust method *(No SS)*; the gray bar shows the results achieved by spectral subtraction *(SS)*; and the black bar represent the results achieved by the proposed method (spectral subtraction plus variance adaptation *(SS + Var. Adapt.)*).

We used white, pink and car noises from the NOISEX-92 [11] database.

4.4. Results

We compared our method (spectral subtraction plus variance adaptation) with spectral subtraction (alone), as described in [12], and with the recognizer without applying any noise compensation method. Figure 2 shows the results for pink and car noises and four Signal to Noise Ratios (SNR).

As can be observed from the Figure 2, variance adaptation is effective in the presence of narrow-band additive noises, such as pink and car noise, at medium and low SNR. For high SNRs, variance adaptation does not result in additional losses.

The good results for car noise deserve special analysis. Car noise is a low-pass noise with a small frequency support. This frequency structure makes that just a few frequency coefficients are affected. The weights computed by eq. (14) are low for these coefficients and close to one for the uncontaminated coefficients. Thus, the method proposed in this paper uses only the reliable components of the feature vector.

The results for white noise (not included in the Figure 2 to avoid distracting the reader from the analysis for narrow-bands noises) tell us that the variance adaptation contribution is not significant.

5. Conclusions

In this paper we addressed the problem of speech recognition in an environment with additive noise. A new variance adaptation method was presented as a way of compensating for the mismatch between training and testing conditions.

The main advantage of our proposal comes from the use of a particular parameterization, the FF parameters, which is closely related to the frequency spectrum. This allows us to avoid the complexity associated with this kind of compensation if other parameterizations, such as cepstra, are used.

Additionally, our method facilitates a novel interpretation of the variance adaptation process as a spectral weighting. From this point of view, the variance adaptation contributes through two separate terms. The first one weights the components of the feature vectors so as to perform recognition only with the most reliable components. The second term penalizes those models that, due to the variance adaptation (widening), become competitive far from their means.

We presented experimental results that show the effectiveness of our method for narrow-band noises.

6. Acknowledgements

This work has been partially supported by Spanish grant UC3M-TEC-05-059.

7. References

[1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, Apr. 1995.

[2] S F Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] Viikki O. and Laurila K., "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 133–147, Aug. 1998.

[4] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 2003, vol. 1, pp. 656–659.

[5] A. de la Torre, A.M. Peinado, J.C. Segura, J.L. Perez-Cordoba, M.C. Benitez, and A.J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, pp. 355–366, May 2005.

[6] M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 5, pp. 352–359, Sept. 1996.

[7] C. Nadeu, D. Macho, and J. Hernando, "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, vol. 34, pp. 93–114, Apr. 2001.

[8] B. Raj and R.M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, Sept. 2005.

[9] *The HTK Book (for HTK Version 3.2.1)*, Cambridge Univ. Press, Cambridge, U.K., 2002.

[10] "NIST, The Resource Management Corpus(RM1)," *Distributed by NIST*, 1992.

[11] A. P. Varga, J. M. Steenneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on Automatic Speech Recognition," in *Tech. Rep. DRA Speech Res. Unit. Malvern, Worcestershire, U. K.*, 1992.

[12] P. Pujol, C. Nadeu, D. Macho, and J. Padrell, "Speech recognition experiments speech recognition experiments with the SPEECON database using several robust front-ends," in *Proc. Int. Conf. on Spoken Language Processing*, Sept. 2004.