NOVEL MODEL COMPENSATION FOR FEATURES BASED ON SNR-DEPENDENT NON-UNIFORM SPECTRAL COMPRESSION

Geng-xin NING†, *Shu-hung LEUNG*‡, *Kam-keung CHU*‡ *and Gang WEI* †

†School of Electronic and Information Engineering, South China University of Technology
‡Department of Electronic Engineering, City University of Hong Kong
Tat Chee Avenue, Kowloon Tong, Kowloon, Hong Kong
Tel: (852) 27887784 Fax: (852) 27887791 E-mail: EEEUGSHL@cityu.edu.hk

ABSTRACT

This paper proposes a novel model compensation method for a robust feature extraction technique based on SNR-dependent Nonuniform Spectral Compression (SNSC). The SNSC method is a spectral transformation which resembles human's intensity-to-loudness conversion and de-emphasizes the contributions from noisy spectral components to features. In this paper, we propose a new compressed mismatch function which models the effect of the noise onto the clean speech in the Log-spectral domain together with the SNSC. Based on this mismatch function, a new model compensation procedure is derived. The procedure needs a compensated model of no compression to start with. It is shown that the new model compensation using the Vector Taylor Series method (VTS) for the compensated uncompressed model, remarkable recognition performances at low signal-to-noise ratio (SNR) can be obtained for different additive noises at the expense of slight increase in the computational complexity in comparison with the VTS.

1. INTRODUCTION

In recent years, the problem of achieving robust speech recognition in noisy environments has aroused many interests. However, drastic degradation of performance occurs when recognizers operate under noisy environments owing to the mismatch between the trained models and testing features. Solutions to this problem can be generally divided into three categories: inherently robust feature representation, speech enhancement scheme, and model-based compensation. Details of these approaches are reviewed in [2] and the references therein.

Our previous work [1] presents a new noise-resistant features by employing SNR-dependent Non-uniform Spectral Compression scheme (SNSC), which compresses the noisy spectral components of the corrupted speech signal with an SNR-dependent root value. It is shown that the SNSC-derived MFCC features outperform the conventional MFCC features and cubic-root compressed features [1] substantially. In SNSC, the compressed speech spectrum in the Linear domain, $\ddot{Y}_k$, is obtained as:

$$\ddot{Y}_k = Y_k^{\alpha_k}, 0 \leq \alpha_k \leq 1 \qquad (1)$$

where Y_k is the k^{th} spectral component of uncompressed spectrum, and α_k is the compression root which is a function of band SNR for the k^{th} filter band. However, since α_k is SNR-dependent, the power spectrum of the noise is required in the training session for finding α_k for a particular noise type and a specified SNR. However, the noise type and SNR of the testing environment are generally unknown. Thus the models estimated by training in this way have limited usage because they could only be applied for a recognition task under the same SNR and noise environment. To solve this drawback, this paper presents a novel compensation procedure to compensate the clean-data-trained models for SNSC-features in various noisy environments of different SNR. In this scheme, it needs a compensated model of no compression, which can be obtained from any known model compensation methods. In this paper, we adopt the Vector Taylor Series (VTS) method [3] for the compensated uncompressed model. This new recognition scheme provides noticeable performance at low SNR and improvement over those systems compensated by the VTS method. We call our scheme as the Model Compensation for SNR Non-uniform Spectral Compression (MC-SNSC).

The structure of this paper is as follows. Firstly, the SNSC method is briefly reviewed in Section 2. In Section 3, we will introduce the MC-SNSC model compensation. Experimental results along with discussion are then presented in Section 4. Our conclusion on this study will be given in the final section.

2. SNR-DEPENDENT NON-UNIFORM SPECTRAL COMPRESSION (SNSC)

Motivated by the partial masking effect imposed by the background noise on perceived loudness, SNSC gives a small compression root value to de-emphasize those speech components that are of low SNR. The procedures to obtain SNSC-derived MFCC features is presented in figure 1. After obtaining the output energies of Mel-scale band-pass filters, spectral compression is carried out as in (1). The core of the SNSC method lies in defining the compression function α_k, which depends on the band SNR. With a slight modification from [1][1], α_k is defined as:

$$\begin{aligned} \alpha_k &= (1 - A_0)(1 - e^{-[\log(\frac{Y_k}{\hat{N}_k}) - \beta]/\gamma}) \\ &\quad \cdot u(\log(\frac{Y_k}{N_k}) - \beta) + A_0 \end{aligned} \qquad (2)$$

where A_0 is the floor compression root, β is the cutoff band SNR to function as the just-audible threshold, γ is the gain to control the

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7001607).

[1]The definition of SNR in this work is different from the work [1].

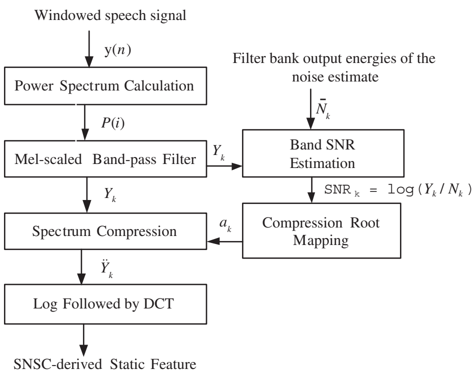

Fig. 1. Procedure of the SNSC scheme

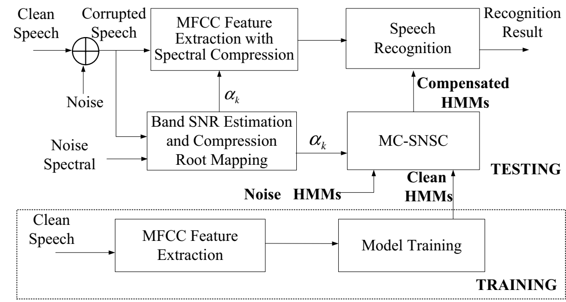

Fig. 2. Processing stages for MC-SNSC

steepness of compression function, and $u(\cdot)$ is a unit step function. This equation can be explained with the partial masking effect of background noise on perceived loudness as discussed in [1]. α_k is dependent on the band SNR, $\log(Y_k/N_k)$. For small band SNR, equation (2) yields a small value to α_k.

In this paper, we assume that the background noise is additive and independent from the speech, and each Mel-band signal is also independent from each other. The mismatch function Y_k of the k^{th} Mel-filter is modelled as the sum of the noise energy N_k and the clean speech energy X_k in the Linear-spectral domain and expressed as

$$Y_k = X_k + N_k \tag{3}$$

If define the variables in the Log-spectral domain as $X_k^l = \log(X_k)$ and $N_k^l = \log(N_k)$, then the mismatch function in the Log-spectral domain is expressed as

$$Y_k^l = \log(e^{X_k^l} + e^{N_k^l}) \tag{4}$$

Thus the compressed mismatch function for the SNSC in the Log-spectral domain is expressed as

$$\begin{aligned} \ddot{Y}_k^l &= \alpha_k Y_k^l && (5)\\ &= \alpha_k \log(e^{X_k^l} + e^{N_k^l}) && (6) \end{aligned}$$

where

$$\alpha_k = (1 - A_0)(1 - e^{-(Y_k^l - N_k^l - \beta)/\gamma}) \cdot u(Y_k^l - N_k^l - \beta) + A_0 \tag{7}$$

The compressed mismatch function as described in equation (5) is an expression relating the effect of the noise onto the clean speech with the SNSC.

In this paper the notation for the description of domain is defined as follows. The superscript ^{l} means the Log-spectral domain. When parameters have no superscript, they are in the Linear-spectral domain. The model parameters of the background noise model and the noise-corrupted speech model are capped with ˜ and ˆ, respectively.

3. MODEL COMPENSATION FOR SNSC (MC-SNSC)

Figure 2 shows the general framework of the MC-SNSC recognition system. During the feature extraction in the testing phase, the SNSC scheme as described in (1) is used to compress the output energy of each filter band. The clean model is combined with the noise model together with the SNSC to construct the corrupted speech model to recognize the SNSC-features.

In the MC-SNSC scheme, we let the first order time derivative of the static features be the delta features. Using (5) and (7), the delta feature of the compressed mismatch function can be expressed as

$$\begin{aligned} \Delta\ddot{Y}_k^l &= \Delta(\alpha_k Y_k^l) = \frac{d(\alpha_k Y_k^l)}{dt} \\ &= A_0 \frac{\partial Y_k^l}{\partial t} + (1 - A_0)\{[(1 - e^{-(Y_k^l - N_k^l - \beta)/\gamma})\frac{\partial Y_k^l}{\partial t} \\ &\quad + \frac{1}{\gamma} e^{-(Y_k^l - N_k^l - \beta)/\gamma} Y_k^l (\frac{\partial Y_k^l}{\partial t} - \frac{\partial N_k^l}{\partial t})] \\ &\quad \cdot u(Y_k^l - N_k^l - \beta) + (1 - e^{-(Y_k^l - N_k^l - \beta)/\gamma}) \\ &\quad \cdot Y_k^l \cdot \delta(Y_k^l - N_k^l - \beta) \cdot (\frac{\partial Y_k^l}{\partial t} - \frac{\partial N_k^l}{\partial t})\} \end{aligned} \tag{8}$$

where $\delta(\cdot)$ is the Dirac-delta function.

By employing the the compressed mismatch functions, (5) and (8), and Gaussian-Hermite numerical integration, we derive a procedure of computing the parameters of the MC-SNSC compensated model in the following subsections.

3.1. Mean Compensation

Using the compressed mismatch function described in (5), the mean of the static SNSC-feature in the Log-spectral domain is given by

$$\begin{aligned} \hat{\mu}_{\ddot{Y}_k}^l &= \hat{\mu}_{(\alpha Y)_k}^l = E\{\alpha_k Y_k^l\} \\ &= (1 - A_0) \cdot (E\{Y_k^l \cdot u(Y_k^l - N_k^l - \beta)\} \\ &\quad - E\{e^{-(Y_k^l - N_k^l - \beta)/\gamma} Y_k^l \cdot u(Y_k^l - N_k^l - \beta)\}) \\ &\quad + A_0 E\{Y_k^l\} \end{aligned} \tag{9}$$

We define

$$g(\gamma) = E\{e^{-(Y_k^l - N_k^l - \beta)/\gamma} Y_k^l \cdot u(Y_k^l - N_k^l - \beta)\} \tag{10}$$

Then

$$\hat{\mu}_{\ddot{Y}_k}^l = (1 - A_0) \cdot [g(\infty) - g(\gamma)] + A_0 \cdot \hat{\mu}_{Y_k}^l \tag{11}$$

where

$$g(\gamma) = e^{\frac{\Phi + \Psi/(2\gamma)}{\gamma}} \cdot [\frac{\hat{\Sigma}_{Y_{kk}}^l}{\sqrt{2\pi\Psi}} e^{-\frac{[\Phi + \Psi/(2\gamma)]^2}{2\Psi}} + \Omega \cdot Sum(\gamma)] \tag{12}$$

$$Sum(\gamma) \cong \frac{1}{2} - \frac{1}{2\sqrt{\pi}} \cdot \sum_{i=1}^{n} \omega_i \cdot erf(\frac{\sqrt{\tilde{\Sigma}^l_{N_{kk}}}}{\sqrt{\hat{\Sigma}^l_{Y_{kk}}}} t_i + \frac{\Phi + \Psi/\gamma}{\sqrt{2\hat{\Sigma}^l_{Y_{kk}}}}) \quad (13)$$

where $\Phi = \tilde{\mu}^l_{N_k} - \hat{\mu}^l_{Y_k} + \beta$, $\Psi = \tilde{\Sigma}^l_{N_{kk}} + \hat{\Sigma}^l_{Y_{kk}}$, and $\Omega = \hat{\mu}^l_{Y_k} - \frac{1}{\gamma}\hat{\Sigma}^l_{Y_{kk}}$. $erf(\cdot)$ is the error function. t_i and ω_i are respectively the abscissas and the weights of the n^{th} order Hermite integration [4].

We assume the background noise to be stationary. The mean of the delta features of the noise can be approximately set equal to zero. Using the compressed mismatch function for delta features in (8), the mean of the delta SNSC-features is obtained as

$$\begin{aligned}
\hat{\mu}^l_{\Delta\ddot{Y}_k} &= \hat{\mu}^l_{\Delta(\alpha Y)_k} = E\{\Delta\alpha_k Y^l_k\} \\
&= A_0\hat{\mu}^l_{\Delta Y_k} + (1-A_0)E\{[1 - e^{-(Y^l_l - N^l_k - \beta)/\gamma} \\
&\quad + \frac{1}{\gamma} e^{-(Y^l_k - N^l_k - \beta)/\gamma} Y^l_k] \\
&\quad \cdot u(Y^l_k - N^l_k - \beta)\}\hat{\mu}^l_{\Delta Y_k} \qquad (14)
\end{aligned}$$

We define

$$h(\gamma) = E\{e^{-(Y^l_k - N^l_k - \beta)/\gamma} \cdot u(Y^l_k - N^l_k - \beta)\} \qquad (15)$$

Then

$$\hat{\mu}^l_{\Delta\ddot{Y}_k} = \{A_0 + (1-A_0)[h(\infty) - h(\gamma) + \frac{1}{\gamma}g(\gamma)]\}\hat{\mu}^l_{\Delta Y_k} \quad (16)$$

where

$$h(\gamma) = e^{\frac{\Phi + \Psi/(2\gamma)}{\gamma}} \cdot Sum(\gamma) \qquad (17)$$

3.2. Variance Compensation

The covariances of the compressed mismatch function in the Log-spectral domain are given by

$$\begin{aligned}
\hat{\Sigma}^l_{\ddot{Y}_{kl}} &= \hat{\Sigma}^l_{(\alpha Y)_{kl}} \qquad (18) \\
&= E\{(\alpha_k Y^l_k) \cdot (\alpha_l Y^l_l)\} - E\{\alpha_k Y^l_k\}E\{\alpha_l Y^l_l\}
\end{aligned}$$

The diagonal elements of the covariance matrix can be calculated as

$$\begin{aligned}
\hat{\Sigma}^l_{\ddot{Y}_{kk}} &= \hat{\Sigma}^l_{(\alpha Y)_{kk}} \\
&= (1-A_0)^2 E\{Y^{l^2}_k \cdot u(Y^l_k - N^l_k - \beta)\} - 2(1-A_0) \\
&\quad \cdot E\{e^{-(Y^l_k - N^l_k - \beta)/\gamma} Y^{l^2}_k \cdot u(Y^l_2 - N^l_2 - \beta)\} + \\
&\quad (1-A_0)^2 E\{e^{-2(Y^l_k - N^l_k - \beta)/\gamma} Y^{l^2}_k u(Y^l_k - N^l_k - \beta)\} \\
&\quad + A_0^2 E\{Y^l_k \cdot u(Y^l_k - N^l_k - \beta)\} - \hat{\mu}^{l^2}_{\ddot{Y}_k} \qquad (19)
\end{aligned}$$

We define

$$f(\gamma) = E\{Y^{l^2}_k e^{-(Y^l_k - N^l_k - \beta)/\gamma} \cdot u(Y^l_k - N^l_k - \beta)\} \quad (20)$$

Then

$$\begin{aligned}
\hat{\Sigma}^l_{\ddot{Y}_{kk}} &= (1-A_0^2)f(\infty) - 2(1-A_0)f(\gamma) + (1-A_0)^2 \\
&\quad \cdot f(\gamma/2) + A_0^2(\hat{\mu}^{l^2}_{Y_k} + \hat{\Sigma}^l_{Y_{kk}}) - \hat{\mu}^{l^2}_{\ddot{Y}_k} \qquad (21)
\end{aligned}$$

where

$$\begin{aligned}
f(\gamma) &= e^{\frac{\Phi + \Psi/(2\gamma)}{\gamma}} \cdot [\frac{\hat{\Sigma}^l_{Y_{kk}}}{\sqrt{2\pi\Psi}}(\hat{\Sigma}^l_{Y_{kk}}\Phi/\Psi + 2\hat{\mu}^l_{Y_k} - \hat{\Sigma}^l_{Y_{kk}}/\gamma) \\
&\quad \cdot e^{-\frac{(\Phi + \Psi/\gamma)^2}{2\Psi}} + (\hat{\Sigma}^l_{Y_{kk}} + \Omega^2)Sum(\gamma)] \qquad (22)
\end{aligned}$$

For off-diagonal elements of the covariance matrix, the computations involve a set of two dimensional Gaussian-Hermite numerical integrals. To reduce the computational complexity, the off-diagonal elements are approximated as

$$\hat{\Sigma}^l_{\ddot{Y}_{lk}} = \hat{\Sigma}^l_{(\alpha Y)_{lk}} \approx \lambda_{lk} E\{\alpha_l\}E\{\alpha_k\}\hat{\Sigma}^l_{Y_{lk}} \qquad (23)$$

where

$$\lambda_{lk} = \lambda_{kl} = \sqrt{\rho_{kk}\rho_{ll}} \qquad (24)$$

and

$$\rho_{kk} = \hat{\Sigma}^l_{\ddot{Y}_{kk}} / \hat{\Sigma}^l_{Y_{kk}} \qquad (25)$$

λ_{lk} is the scaling factor to make the off-diagonal elements keeping up with the corresponding diagonal elements.

Similar to the approach for estimating the static variance of the compressed mismatch function, the diagonal elements of the covariance matrix of the delta model can be calculated by

$$\begin{aligned}
\hat{\Sigma}^l_{\Delta\ddot{Y}_{kk}} &= [A_0^2 + (1-A_0^2)h(\infty) - 2(1-A_0)h(\gamma) \\
&\quad + (1-A_0)^2 h(\gamma/2) + \frac{2(1-A_0)}{\gamma}g(\gamma) \qquad (26) \\
&\quad - \frac{2(1-A_0)^2}{\gamma}g(\gamma/2) + \frac{(1-A_0)^2}{\gamma^2}f(\gamma)] \\
&\quad \cdot (\hat{\Sigma}_{\Delta Y_{kk}} + \hat{\mu}^{l^2}_{\Delta Y_k}) + \frac{(1-A_0)^2}{\gamma^2}f(\gamma/2)\tilde{\Sigma}^{l^2}_{\Delta N_{kk}} \\
&\quad - \hat{\mu}^{l^2}_{\Delta\ddot{Y}_k}
\end{aligned}$$

and the off-diagonal elements as

$$\hat{\Sigma}^l_{\Delta\ddot{Y}_{lk}} = \hat{\Sigma}^l_{\Delta(\alpha Y)_{lk}} \approx \lambda'_{lk} E\{\alpha_l\}E\{\alpha_k\}\hat{\Sigma}^l_{\Delta Y_{lk}} \qquad (27)$$

where

$$\lambda'_{lk} = \lambda'_{kl} = \sqrt{\rho'_{kk}\rho'_{ll}} \qquad (28)$$

and

$$\rho'_{kk} = \hat{\Sigma}^l_{\Delta\ddot{Y}_{kk}} / \hat{\Sigma}^l_{\Delta Y_{kk}} \qquad (29)$$

3.3. Solution for Non-compressed Corrupted Model

The above model compensation equations show that a non compressed corrupted speech models ($\hat{\mu}^l_{Y_k}$,$\hat{\Sigma}^l_{Y_{kl}}$) and ($\hat{\mu}^l_{\Delta Y_k}$,$\hat{\Sigma}^l_{\Delta Y_{kl}}$) are required in the MC-SNSC model compensation process, which can be obtained from any model compensation methods. We employ the VTS [3] method to obtain the non-compressed corrupted speech models.

4. EXPERIMENTAL RESULTS

Isolated digits from the AURORA Project Database 2.0 [5] is used for the evaluation of the proposed MC-SNSC approach. There are 2412 utterances for clean training, with 110 speakers uttering each digits twice. For the testing, data corrupted with Car noise, Exhibition noise and Suburban Train noise from the test set of the AURORA are used (around 300 utterances for each noise and each SNR). For the matched case, the noisy training data were generated by adding the noise to the clean training data at various global SNR. The length of the analysis frame (windowed by Hamming weights) is 25ms, and the frame rate is 10ms/frame. The feature vector is composed of static feature and its first derivative, each has 13 cepstral coefficients.

A word based HMM with six states and four Gaussian densities per state is used as the recognizer. In the training mode, we train

Table 1. Recognition Rate (%) of various Model Compensation Methods For Car Noise ($A_0 = 0.7, \gamma = 1, \beta = -0.2$).

SNR/dB	Mismatched	Matched	VTS-1	MC-SNSC
Clean	98.57	98.57	98.57	98.57
20	84.68	98.39	96.28	96.93
15	71.33	97.74	95.05	96.02
10	40.74	96.09	92.74	94.81
5	27.93	95.65	88.79	89.91
0	20.10	87.14	70.74	80.60
-5	11.01	51.56	36.93	58.15
Average between -5 and 5dB	19.68	78.12	65.49	76.22

Table 2. Recognition Rate (%) of various Model Compensation Methods for Exhibition Noise ($A_0 = 0.85, \gamma = 1, \beta = 0$).

SNR/dB	Mismatched	Matched	VTS-1	MC-SNSC
Clean	99.62	99.62	99.62	99.62
20	64.46	96.76	95.39	95.03
15	39.12	94.97	94.01	94.21
10	20.70	94.63	91.26	90.28
5	16.11	89.23	84.54	87.30
0	9.82	78.43	61.03	69.32
-5	9.49	51.75	33.11	47.41
Average between -5 and 5dB	19.68	73.14	59.56	68.01

the system with the clean speech utterances to produce clean models except for matched case. In the testing, three other speech recognition methods are carried out for the sake of comparison with our MC-SNSC. These three methods are the mismatched case without model compensation, matched case, and the first order VTS method (denoted as VTS-1).

Experimental results for Car noise, Exhibition noise and Suburban Train noise are shown in Table 1 to 3, respectively. The results show that the VTS-1 and our MC-SNSC method can achieve good performance for the three additive noises at low SNR. However, the MC-SNSC provides better accuracy than the VTS-1 at very low SNR and gives a performance near to the matched case. Looking at the recognition rate "averaged between -5dB to 5dB" for Car noise, the MC-SNSC has accuracy of 11% (in absolute percentage) more than the VTS-1. At SNR=-5dB, even MC-SNSC gives a better performance than the matched case. For Exhibition and Suburban Train noise, the performance of using the MC-SNSC method is also better than that of the VTS-1 at low SNR.

Table 4 lists the number of multiplications, divisions, logarithm and exponential operations for each technique to update the parameters of a single mixture density, where N and M are the dimensions of features in the cepstral domain and the Log-spectral domain, respectively. It is shown that the computational complexity of MC-SNSC is comparable to that of VTS-1 [3].

These experimental results reveal that the new model compensation for the SNSC scheme can deal with different types of additive noise and yield remarkable recognition performance, which are attributed to the noise-resistant feature extraction (SNSC scheme) [1] and pertinent model compensation.

Table 3. Recognition Rate (%) of various Model Compensation Methods for Suburban Train Noise ($A_0 = 0.85, \gamma = 1, \beta = -0.2$)

SNR (dB)	Mismatch Case	Matched Case	VTS-1	MC-SNSC
Clean	99.29	99.29	99.29	99.29
20	68.97	99.38	98.36	98.62
15	43.54	98.28	97.67	96.78
10	22.74	98.21	94.04	94.64
5	12.63	96.22	89.14	91.33
0	8.50	89.05	79.04	84.11
-5	9.09	66.68	54.41	64.07
Average between -5 and 5dB	10.07	83.98	74.20	79.84

Table 4. Computation complexity of the mentioned compensation approach

Method	Total	M=23, N=13
VTS	2MN(2M+N+3) +12M^2+13M	43723
MC-SNSC	2MN(2M+N+3) +16M^2+(3n+69)M	47403 (n=4)

5. CONCLUSION

A novel MC-SNSC technique that provides a model compensation scheme for the the robust SNSC-feature is presented in this paper. The compressed mismatch function that models the effect of additive noise on the spectrally compressed speech features is proposed and the corresponding model compensation is developed. Experimental results show that the MC-SNSC can cope with different kinds of noises and provide a remarkable performance on the recognition accuracy especially at low SNR. Nevertheless, the computational complexity of the new MC-SNSC scheme is comparable to that of the VTS method.

6. REFERENCES

[1] K. K. Chu and S. H. Leung,"SNR-dependent Non-uniform Spectral Compression for Noisy Speech Recognition," *Proc. ICASSP 2004*, pp.973-976, 2004.

[2] Y. Gong,"Speech recognition in noisy environments: A survey," *Speech Communication*, vol.16, pp.261-291, 1995.

[3] P. J. Moreno, B. Raj, and R.M. Stern,"A vector Taylor series approach for environment-independent speech recognition", *ICASSP-96*, vol.2, pp.733-736, May, 1996.

[4] M. Abramowitz and I. A. Stegun, *HANDBOOK OF MATHEMATICAL FUNCTIONS with Formulas, Graphs, and Mathematical tables*, Dover Publications Inc., New York, 1972.

[5] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions" *IACA ITRW ASR 2000*, paris, France, Sept. 2000.