INTERSPEECH 2009 BRIGHTON

An analytic derivation of a phase-sensitive observation model for noise robust speech recognition

Volker Leutnant and Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, Germany

{leutnant,haeb}@nt.uni-paderborn.de

Abstract

In this paper we present an analytic derivation of the moments of the phase factor between clean speech and noise cepstral or log-mel-spectral feature vectors. The development shows, among others, that the probability density of the phase factor is of sub-Gaussian nature and that it is independent of the noise type and the signal-to-noise ratio, however dependent on the mel filter bank index. Further we show how to compute the contribution of the phase factor to both the mean and the variance of the noisy speech observation likelihood, which relates the speech and noise feature vectors to those of noisy speech. The resulting phase-sensitive observation model is then used in model-based speech feature enhancement, leading to significant improvements in word accuracy on the AURORA2 database.

Index Terms: model-based feature enhancement, phase-sensitive observation model, phase factor distribution

1. Introduction

Model-based speech feature enhancement is a front-end technique, where the noise-free speech cepstral or log-mel-spectral feature vectors are estimated from the noisy observations based on an a priori model of clean speech and noise and an observation model relating the two to the noisy speech feature vectors. This approach to noise robust speech recognition has gained considerable interest in recent years as it approaches the recognition performance of back-end techniques, where the acoustic models of clean speech are modified to reflect the observed noisy speech, while being at the same time computationally less demanding.

As the relationship between clean speech and noise feature vectors, $\mathbf{x}$ and $\mathbf{n}$, and those of noisy speech $\mathbf{y}$ is highly nonlinear, several approximations have been proposed to model the observation probability $p(\mathbf{y}|\mathbf{x},\mathbf{n})$. The most widely used and at the same time the simplest approach is to neglect any phase difference between speech and noise resulting in a Dirac impulse for the aforementioned probability density function at the location $\mathbf{y} = \mathbf{x} + \log(1 + \exp(\mathbf{n} - \mathbf{x}))$ (assuming log-mel-spectral feature vectors). However, it is well-known that a more accurate model is obtained if a phase factor $\boldsymbol{\alpha}$, which results from the unknown phase between the complex speech and noise short-term discrete-time Fourier transform, is taken into account [1, 2, 3, 4]. While in most cases the probability density of the phase factor is assumed to be a zero mean Gaussian whose variance is determined experimentally on stereo training data, Faubel et al. [4] determined the density by Monte Carlo simulations and showed experimentally that it is sub-Gaussian, approaching a Gaussian density only for higher mel filter bank bins. Subsequently, the observation probability $p(\mathbf{y}|\mathbf{x},\mathbf{n})$ can be determined either by Vector Taylor Series approximation up to linear [5] or higher-order terms [3] or by Monte Carlo Integration [4]. An analytic expression can be found in case the phase factor is assumed to be Gaussian distributed [1].

In speech feature enhancement the mean, and in case of uncertainty decoding also the variance, of the posterior $p(\mathbf{x}|\mathbf{y})$ needs to be computed. Since a numerical evaluation of the resulting integrals is computationally very demanding if not almost impossible, the observation probability is approximated by a Gaussian, where the effect of the phase factor is either modelled as a contribution to the mean [4], to the variance [2] or to both mean and variance [3].

In this paper we show how the moments of the phase factor can be computed analytically, rendering stereo training data obsolete. From this derivation we can confirm the experimental observation made by others that the density of the phase factor is sub-Gaussian and that it is independent of the noise type and the signal-to-noise ratio. A Taylor series expansion is then carried out to obtain mean and variance of $p(\mathbf{y}|\mathbf{x},\mathbf{n})$, which is assumed to be Gaussian. Clearly, the phase factor delivers both a contribution to the mean (a bias term) and variance of the observation probability and the best recognition results are obtained if both are accounted for.

The paper is organized as follows. After a short introduction to model-based speech feature enhancement incorporating an a priori model of clean speech which accounts for inter-frame correlations, we consider the phase-sensitive observation model. Next, we show how the moments of the phase factor can be computed analytically, closing with a comparison with empirically determined parameters. Finally, we present recognition results on the AURORA2 database providing experimental evidence of the superiority of the phase-sensitive observation model to its phase-insensitive counterpart.

2. Model-Based Feature Enhancement

Given a sequence of (possibly corrupted) feature vectors $\mathbf{y}_1^T = (\mathbf{y}_1, ..., \mathbf{y}_T)$, a key element of speech feature enhancement is the posterior $p(\mathbf{x}_t|\mathbf{y}_1^T)$ of the clean speech feature $\mathbf{x}_t$ or, in case the noise feature $\mathbf{n}_t$ is modelled as a random variable rather than an unknown parameter, the joint posterior $p(\mathbf{z}_t = (\mathbf{x}_t, \mathbf{n}_t)'|\mathbf{y}_1^T)$. Hence, the effectiveness of a feature enhancement scheme crucially depends on how well it can be determined. Knowledge of the posterior density enables the computation of an optimal estimate with respect to any criterion. For example, the minimum mean squared error (MMSE) estimate equals its mean. Furthermore, a measure of accuracy of the estimate can be obtained from the variance of the posterior. In the Gaussian case the variance of the posterior is even identical to the estimation error variance. Conceptually, the joint posterior can be estimated recursively via the following equations:

$$p(\mathbf{z}_t|\mathbf{y}_1^{t-1}) = \int p(\mathbf{z}_t|\mathbf{z}_{t-1}) p(\mathbf{z}_{t-1}|\mathbf{y}_1^{t-1}) d\mathbf{z}_{t-1} \quad (1)$$

$$p(\mathbf{z}_t|\mathbf{y}_1^{t}) = \frac{p(\mathbf{y}_t|\mathbf{z}_t) p(\mathbf{z}_t|\mathbf{y}_1^{t-1})}{\int p(\mathbf{y}_t|\mathbf{z}_t) p(\mathbf{z}_t|\mathbf{y}_1^{t-1}) d\mathbf{z}_t}, \quad (2)$$

where we restrict ourselves to causal processing, i.e. rather than computing $p(\mathbf{z}_t|\mathbf{y}_1^T)$ we compute $p(\mathbf{z}_t|\mathbf{y}_1^t)$.

From these equations it can be observed that the observation probability or likelihood $p(\mathbf{y}_t|\mathbf{z}_t)$ is a key component for the determination of the posterior. This term should be modelled such that it both reflects the true dependency as accurately as possible and at the same time is computationally tractable to enable an, at least approximate, inference, i.e. recursive computation of the posterior $p(\mathbf{z}_t|\mathbf{y}_1^t)$, $t = 1, \ldots, T$.

In the following we are going to develop such a model incorporating a novel analytical treatment of the phase factor between speech and noise.

3. A Phase-Sensitive Observation Model

We assume that the speech signal is corrupted by additive environmental noise. Let $X_{t,k}$, $N_{t,k}$ and $Y_{t,k}$ denote the complex-valued short-time Discrete Fourier Transform coefficients in the k^{th} frequency bin, $k \in [1, ..., K]$, and at discrete-time frame index t of the clean speech, noise and noisy speech, respectively. Their relationship in the power spectral domain follows to be

$$|Y_k|^2 = |X_k|^2 + |N_k|^2 + 2|X_k||N_k| \cos(\theta_k), \quad (3)$$

with θ_k being the relative phase between the speech and noise short-term discrete-time Fourier transforms. Here and in the following we omit the frame index t for ease of notation. The representation in the i^{th} mel-frequency bin thus becomes

$$\tilde{Y}_i = \tilde{X}_i + \tilde{N}_i + 2\alpha_i \sqrt{\tilde{X}_i \tilde{N}_i}, \quad i \in [1, ..., I] \quad (4)$$

with $\tilde{()}_i = \sum_{k=1}^{K} W_{i,k} |()_k|^2$ and

$$\alpha_i = \frac{\sum_{k=1}^{K} W_{i,k}|X_k||N_k| \cos\theta_k}{\sqrt{\tilde{X}_i \tilde{N}_i}} = \sum_{k=1}^{K} c_{i,k} \cos\theta_k \quad (5)$$

denoting the phase factor of the i^{th} triangular-shaped mel filter with coefficients $W_{i,k}$ as derived in [6]. Translation of (4) to the log-mel domain results in the phase-sensitive environment model as given in [1]

$$y_i = \log\left(e^{x_i} + e^{n_i} + 2\alpha_i e^{\frac{x_i+n_i}{2}}\right) \quad (6)$$

$$= \log\left(e^{x_i} + e^{n_i}\right) + \log\left(1 + 2\alpha_i \frac{e^{\frac{n_i - x_i}{2}}}{1 + e^{n_i - x_i}}\right) \quad (7)$$

$$= \log\left(e^{x_i} + e^{n_i}\right) + \varphi\left(\alpha_i, n_i - x_i\right). \quad (8)$$

A common but problematic approximation of (8) is

$$y_i \approx \log\left(e^{x_i} + e^{n_i}\right), \quad (9)$$

where the phase-dependent term $\varphi(\alpha_i, n_i - x_i)$ is neglected. While the error due to this approximation is rather small when clean speech and noise mix at different levels, i.e. $|n_i - x_i| \gg 0$, the opposite is true when they mix at levels where $x_i \approx n_i$.

Since feature enhancement based on the phase-sensitive observation model presented in (6) can be expected to outperform approaches based on the phase-insensitive model given by (9), a way to incorporate the information provided by the phase-dependent term into model-based feature enhancement is now derived. In order to avoid dealing with the DCT matrix and its pseudo-inverse, the derivation is carried out in the log-mel domain. Generalization to the cepstral domain is straightforward.

The non-linearity of (6) and (9) makes their application to speech feature enhancement rather impractical. A common way to circumvent dealing with this non-linearity directly is to first expand the observation model $\mathbf{y}$ into a Vector Taylor Series and later truncate it to linear terms, only. While this linearization is usually carried out with respect to $\mathbf{x}$ and $\mathbf{n}$ and disregards any terms of order 2 or higher [5], our approach is to expand the observation model with respect to $\mathbf{x}$, $\mathbf{n}$ and $\boldsymbol{\alpha}$, truncate the series to linear terms in $\mathbf{x}$ and $\mathbf{n}$ and use all other terms up to and including 2^{nd}-order terms to model the linearization error.

Denoting the i^{th} component of the expansion vectors of the clean speech, the noise and the phase factor by x_i^0, n_i^0 and α_i^0 respectively, Taylor series expansion of y_i with respect to x_i, n_i and α_i gives

$$\begin{aligned} y_i =& y_i\left(x_i^0, n_i^0, \alpha_i^0\right) + J_x^i\left(x_i - x_i^0\right) + J_n^i\left(n_i - n_i^0\right) \\ &+ J_\alpha^i\left(\alpha_i - \alpha_i^0\right) + \frac{1}{2}H_{xx}^i\left(x_i - x_i^0\right)^2 \\ &+ \frac{1}{2}H_{nn}^i\left(n_i - n_i^0\right)^2 + \frac{1}{2}H_{\alpha\alpha}^i\left(\alpha_i - \alpha_i^0\right)^2 \\ &+ H_{xn}^i\left(x_i - x_i^0\right)\left(n_i - n_i^0\right) \\ &+ H_{x\alpha}^i\left(x_i - x_i^0\right)\left(\alpha_i - \alpha_i^0\right) \\ &+ H_{n\alpha}^i\left(n_i - n_i^0\right)\left(\alpha_i - \alpha_i^0\right) + \mathcal{H}.\mathcal{O}.\mathcal{T}. \quad (10) \\ =& y_i\left(x_i^0, n_i^0, \alpha_i^0\right) + J_x^i\left(x_i - x_i^0\right) + J_n^i\left(n_i - n_i^0\right) \\ &+ \varepsilon_i + \mathcal{H}.\mathcal{O}.\mathcal{T}. \quad (11) \\ =& g(x_i, x_i^0, n_i, n_i^0, \alpha_i^0) + \varepsilon_i + \mathcal{H}.\mathcal{O}.\mathcal{T}. \quad (12) \end{aligned}$$

with

$$J_x^i = \left.\frac{\partial y_i}{\partial x_i}\right|_{\mathbf{z}_i^0}, J_n^i = \left.\frac{\partial y_i}{\partial n_i}\right|_{\mathbf{z}_i^0}, J_\alpha^i = \left.\frac{\partial y_i}{\partial \alpha_i}\right|_{\mathbf{z}_i^0} \quad (13)$$

denoting the elements of the Jacobian matrix and

$$\begin{aligned} H_{xn}^i &= \left.\frac{\partial^2 y_i}{\partial x_i \partial n_i}\right|_{\mathbf{z}_i^0}, H_{x\alpha}^i = \left.\frac{\partial^2 y_i}{\partial x_i \partial \alpha_i}\right|_{\mathbf{z}_i^0}, H_{n\alpha}^i = \left.\frac{\partial^2 y_i}{\partial n_i \partial \alpha_i}\right|_{\mathbf{z}_i^0} \\ H_{xx}^i &= \left.\frac{\partial^2 y_i}{\partial x_i^2}\right|_{\mathbf{z}_i^0}, H_{nn}^i = \left.\frac{\partial^2 y_i}{\partial n_i^2}\right|_{\mathbf{z}_i^0}, H_{\alpha\alpha}^i = \left.\frac{\partial^2 y_i}{\partial \alpha_i^2}\right|_{\mathbf{z}_i^0} \quad (14) \end{aligned}$$

denoting the corresponding elements of the Hessian matrix with respect to x_i, n_i and α_i, all evaluated at the expansion points x_i^0, n_i^0 and α_i^0 (abbreviated as $\mathbf{z}_i^0$). By further neglecting higher order terms ($\mathcal{H}.\mathcal{O}.\mathcal{T}.$) and assuming the linearization error ε_i to be Gaussian, the observation probability $p(y_i|x_i, n_i)$ becomes Gaussian, too, with mean $\mu_{y,i} = g(x_i, x_i^0, n_i, n_i^0, \alpha_i^0) + E[\varepsilon_i]$ and variance $\sigma_{y,i}^2 = E[\varepsilon_i^2] - E[\varepsilon_i^2]^2$.

4. Application to Feature Enhancement

The linearization of (6) is quite sensitive to the choice of the expansion points. A common practice with regard to (2) is to use prior knowledge about the estimation problem, which is provided by $p(\mathbf{z}_t|\mathbf{y}_1^{t-1})$ and $p(\boldsymbol{\alpha})$. While their means $\boldsymbol{\mu}_z = (\boldsymbol{\mu}_x, \boldsymbol{\mu}_n)$ and $\boldsymbol{\mu}_\alpha$ are used as expansion vectors, their covariances $\boldsymbol{\Sigma}_z = (\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_{xn}; \boldsymbol{\Sigma}_{nx}, \boldsymbol{\Sigma}_n)$ and $\boldsymbol{\Sigma}_\alpha$ can be employed to compute the mean and variance of the error term $\boldsymbol{\varepsilon}$ using (10). The first and second central moments of the distribution of the linearization error ε_i are thus given by

$$E[\varepsilon_i] = \frac{1}{2}\left(H_{xx}^i \sigma_{x,i}^2 + H_{nn}^i \sigma_{x,i}^2 + H_{\alpha\alpha}^i \sigma_{\alpha,i}^2\right) \quad (15)$$

and

$$E[\varepsilon_i^2] = \left(J_\alpha^i\right)^2 \sigma_{\alpha,i}^2 + \frac{3}{4}\left(H_{xx}^i\right)^2 \sigma_{x,i}^4 + \frac{3}{4}\left(H_{nn}^i\right)^2 \sigma_{n,i}^4 + \frac{1}{4}\left(H_{\alpha\alpha}^i\right)^2 E\left[(\alpha_i - \mu_{\alpha,i})^4\right] + H_{xn}^i \sigma_{x,i}^2 \sigma_{n,i}^2 + H_{x\alpha}^i \sigma_{x,i}^2 \sigma_{\alpha,i}^2 + H_{n\alpha}^i \sigma_{n,i}^2 \sigma_{\alpha,i}^2, \quad (16)$$

where the speech, the noise and the phase factor are assumed to be uncorrelated. The approximation in (16) assumes the posterior distribution of speech and noise to be jointly Gaussian, allowing us to replace 4^{th}-order moments by three times the square of the 2^{nd}-order moments. While the Gaussian assumption is quite valid for the joint feature vector distribution of $\mathbf{x}$ and $\mathbf{n}$, it does not hold for the phase factor distribution, which is of sub-Gaussian nature [1, 4].

A related approach has been proposed by Stouten et al. [3]. However, they only employed the first two terms in (15) and only the first term in (16) to approximate the mean and the variance, respectively.

5. Moments of the Phase Factor Distribution

The moments of the posterior $p(\mathbf{z}_t|\mathbf{y}_1^{t-1})$ are a byproduct of the enhancement scheme. Thus, the only unknowns remaining are the 2^{nd} and 4^{th} central moments of the phase factor distribution. In contrast to [1], explicit knowledge of the phase factor distribution is not required.

Since stereo training data comprising the noisy observation, the clean speech and the noise only is usually not given, an analytical solution to the required moments of the phase factor is desirable. Recalling (5), the phase factor of the i^{th} mel-frequency bin is given by $\alpha_i = \sum_{k=1}^{K} c_{i,k} \cos\theta_k$, with $c_{i,k}$ collating all terms not depending on θ_k. The relative phases θ_k, $k \in [1, ..., K]$, are now assumed to be statistically independent random variables each drawn from the uniform distribution over $-\pi \leq \theta_k < \pi$. The density of the random variable $u_k = \cos(\theta_k)$ thus can be shown to be

$$p(u_k) = \begin{cases} \frac{1}{\pi} \frac{1}{\sqrt{1-u_k^2}}, & \text{for } |u_k| < 1 \\ 0, & \text{else.} \end{cases} \quad (17)$$

The central moments of $p(u_k)$ can now be obtained by utilizing the characteristic function $\phi_{u_k}(\tau) = E[e^{j\tau u_k}]$ of the random variable u_k, which formally equals (up to a factor of 2π) the inverse Fourier transform of $p(u_k)$.

More precisely, the n^{th}-order derivative of $\phi_{u_k}(\tau)$ with respect to τ evaluated at $\tau = 0$ and divided by j^n gives the n^{th} central moment of $p(u_k)$. From Fourier transform tables, the characteristic function of the random variable u_k can be found to be $\phi_{u_k}(\tau) = J_0(\tau)$, with J_0 denoting the 0^{th}-order Bessel function. The characteristic function $\phi_{\tilde{u}_k}(\tau)$ of the random variable $\tilde{u}_k = c_{i,k} u_k$ is obtained by applying standard Fourier transform rules which yield $\phi_{\tilde{u}_k}(\tau) = J_0(c_{i,k}\tau)$.

Since neighbouring short-term DFT-bins and thus the relative phases are asymptotically independent [7], the probability density of the random variable α_i can be expressed as the convolution of the probability densities of all terms under the sum. Applying standard Fourier transformation rules, again, we find the characteristic function of the random variable α_i to be

$$\phi_{\alpha_i}(\tau) = \prod_{k=1}^{K} J_0(c_{i,k}\tau). \quad (18)$$

Differentiating (18) n times with respect to τ, dividing the outcome by j^n and evaluating the resulting function at $\tau = 0$ finally gives the n^{th} central moment. In particular, we find

$$E[\alpha_i] = \mu_{\alpha,i} = 0 = E[\alpha_i^{2n-1}], \quad n = 1, 2, ... \quad (19)$$

$$E[\alpha_i^2] = \sigma_{\alpha,i}^2 = \frac{1}{2}\sum_{k=1}^{K} c_{i,k}^2 \quad (20)$$

$$E[\alpha_i^4] = 3E[\alpha_i^2]^2 - \frac{3}{8}\sum_{k=1}^{K} c_{i,k}^4 \leq 3E[\alpha_i^2]^2 \quad (21)$$

with (19) confirming the zero mean assumption made in literature and (21) imposingly pointing out the sub-Gaussian nature of the phase factor distribution. Replacing the weights $c_{i,k}$ by their definition introduced in (5), Eq. (21) further indicates that $E[\alpha_i^4]$ approaches $3 \cdot E[\alpha_i^2]^2$ with an increasing number of non-zero mel filter weights contributing to $c_{i,k}$.

Assuming the magnitude spectra of the clean speech $|X_k|$ and the noise $|N_k|$ to be constant over the short-term DFT-bins covered by the non-zero elements of the according triangular-shaped mel filter i finally yields the required central moments to be independent of the clean speech and the noise and thus to be independent of both, noise type and signal-to-noise ratio:

$$E[\alpha_i^2] = \sigma_{\alpha,i}^2 = \frac{1}{2}\sum_{k=1}^{K} W_{i,k}^2 \Bigg/ \left(\sum_{k=1}^{K} W_{i,k}\right)^2 \quad (22)$$

$$E[\alpha_i^4] = 3E[\alpha_i^2]^2 - \frac{3}{8}\sum_{k=1}^{K} W_{i,k}^4 \Bigg/ \left(\sum_{k=1}^{K} W_{i,k}\right)^4. \quad (23)$$

However, the estimated variance depends on the shape of the analysis window. To be more specific, Brillinger [8, Theorem 5.6.4] found that if h is a tapered window of length L, e.g. the commonly used Hamming window, the variance of the averaged periodograms is larger than for the untapered (rectangular) case by a factor of [8, 7]

$$F_h = L\sum_{l=1}^{L} h_l^4 \Bigg/ \left(\sum_{l=1}^{L} h_l^2\right)^2 \geq 1. \quad (24)$$

Hence, strictly speaking, equations (22) and (23) hold for a rectangular analysis window, only. The variance for a tapered window h is thus given by $F_h \cdot E[\alpha_i^2]$. Comparing the resulting variances with the ones determined experimentally, based on the evaluation of (5) with stereo data from the AURORA2 database for different noise types and different SNRs, validates our results, as depicted in Figure 1. The effect on the 4^{th} central moment is considered to be marginal and (23) can be applied once the *corrected* variance has been determined.

6. Experimental Results

The experiments were conducted on the test sets A and B of the AURORA2 database. Training has been carried out on the clean speech. The ETSI standard front-end feature extraction algorithm has been modified by replacing the log-energy feature with c_0 and using the power spectral density rather than the spectral magnitude as the input of the mel filter bank. The baseline feature enhancement is similar to the one described in [5], differing however in the fact that we use GPB-inference of order one in a system comprising switching linear dynamic models (SLDMs) with $M = 16$ individual dynamic models to describe the clean speech trajectory. The noise prior is modelled by a

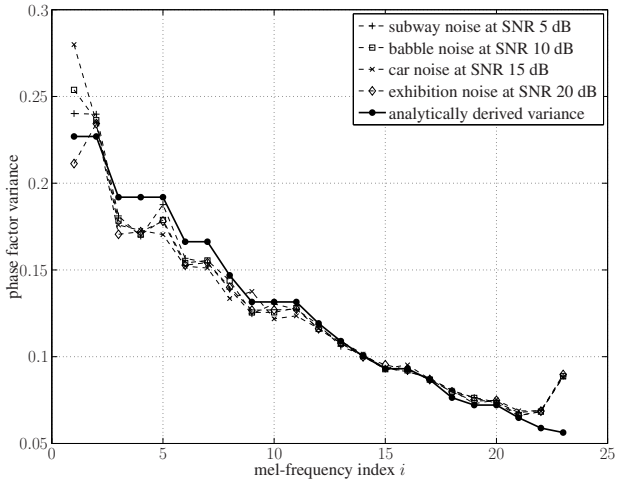

Figure 1: Analytically found and experimentally determined phase factor variances on a subset of the AURORA2 database

single Gaussian trained on the first and last 10 frames of each utterance. In addition, cepstral mean normalization (CMN) was applied to the enhanced features.

We compare the performance of the phase-insensitive observation model (9) with the phase-sensitive model derived in this paper. The iterated extended Kalman filter (IEKF) is used to compute (2), with the number of additional iterations set to 1. To distinguish the observation models, our phase-sensitive observation model is denoted by IEKF-α while the phase-insensitive observation model is just denoted by IEKF. Since the model-based feature enhancement applied in this paper not only delivers a point estimate for the clean speech feature vector but also a measure of its accuracy, namely the variance of its posterior, application of uncertainty decoding (UD) [9], denoted by IEKF-α-UD, is also investigated.

The results on the AURORA2 recognition task can be studied in Table 1. The recognition accuracy is increased by approximately 2.5% absolute on test set A and B when using the phase-sensitive observation model IEKF-α. Further improvements are obtained when uncertainty decoding is applied. Interestingly, the incorporation of the phase factor into the observation model is found to become more important with decreasing SNR values – thus confirming our thesis stated in section 3.

SNR	**test set A**			**test set B**		
[dB]	IEKF	IEKF-α	IEKF-α-UD	IEKF	IEKF-α	IEKF-α-UD
20	98.56	98.72	98.72	98.53	98.53	98.58
15	96.93	97.26	97.37	97.03	97.36	97.38
10	92.70	93.73	94.28	92.39	93.38	93.62
5	79.74	83.07	84.42	78.87	82.40	83.49
0	49.54	56.94	59.32	49.68	56.97	58.50
AVG	83.50	85.95	86.82	83.30	85.73	86.31

Table 1: Averaged recognition accuracies on test set A and B of the AURORA2 database

We also compared our phase-sensitive observation model IEKF-α with other phase-sensitive models in the literature. If only the effect of the phase-factor on the mean of ε_i is taken into account, an approach similar to [4], average recognition accuracies of 84.21% and 83.76% are obtained on test set A and test set B, respectively. If only the effect on the variance of ε_i is considered, similar to [2], the averaged word accuracy is 85.23% on test set A and 85.41% on test set B. And with (16) and (15) approximated according to Stouten et al. [3], we achieved recognition accuracies of 85.30% and 85.03% on test set A and B of the AURORA2 database.

7. Conclusions

In this paper we have first studied an observation model accounting for the relative phase between the complex short-term DFT-bins of the clean speech signal and the noise. While common approaches assume the involved phase factor to be Gaussian distributed and base the estimation of the required moments on available stereo training data, we proved analytically that the phase factor distribution is of sub-Gaussian nature and moreover derived a way to analytically compute all central moments solely based on the used mel filter bank. In doing so, we implicitly showed their independence of noise type and signal-to-noise ratio. Incorporation of the phase-sensitive observation model into a model-based feature enhancement scheme and its application to the AURORA2 recognition task revealed the superiority of the phase-sensitive observation model to its phase-insensitive counterpart and different other phase-sensitive models proposed in the literature.

8. Acknowledgement

This work has been supported by Deutsche Forschungsgemeinschaft (DFG) under contract no. Ha3455/6-1.

9. References

[1] J. Droppo, A. Acero, and L. Deng, "A nonlinear observation model for removing noise from corrupted speech log mel-spectral energies," in *Proc. International Conference on Spoken Language Processing*, Denver, Colorado, 2002.

[2] J. Droppo, L. Deng, and A. Alex, "A comparison of three nonlinear observation models for noisy speech features," in *Proc. Eurospeech*. Geneva, Switzerland: International Speech Communication Association, September 2003, pp. 681–684.

[3] V. Stouten, H. Van hamme, and P. Wambacq, "Effect of phase-sensitive environment model and higher order vts on noisy speech feature enhancement," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, 2005, pp. 433–436.

[4] F. Faubel, J. McDonough, and D. Klakow, "A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-mel domain," in *Proceedings of Interspeech 2008*. Interspeech, September 2008.

[5] J. Deng, M. Bouchard, and T. H. Yeap, "Noisy speech feature estimation on the aurora2 database using a switching linear dynamic model," *Journal of Multimedia (JMM)*, vol. 2, no. 2, pp. 47–52, April 2007.

[6] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of hmm variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, 2005.

[7] R. Martin and T. Lotter, "Optimal recursive smoothing of non-stationary periodograms," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, September 2001.

[8] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston, Inc., 1975.

[9] V. Ion and R. Haeb-Umbach, "A novel uncertainty decoding rule with applications to transmission error robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 1047–1060, 2008.