Enhancement of Mel log-power spectrum of speech using particle filtering

Ilyas Potamitis[†], *Nikos Fakotakis*[±].

[†]Department of Music Technology and Acoustics
Technological Educational Institute of Crete, Daskalaki-Perivolia, 74100, Rethymno, Crete, Greece
potamitis@wcl.ee.upatras.gr
[±]Department of Electrical Engineering and Computer Science
University of Patras, 26 500 Rio, Greece
fakotaki@wcl.ee.upatras.gr

10.21437/Interspeech.2005-218

Abstract

The subject of this work is a statistical feature enhancement technique for robust speech recognition applied to the log-power domain after the application of the Mel filterbank. The proposed approach makes use of a state space formulation that involves a random walk model for the evolution of the underlying clean features and a non-linear observation model that connects the noisy features with noise and clean speech. The novelty of the proposed approach is that a) both observation and state noise are shown to be heavy-tailed and are subsequently modelled using a mixture of Gaussians, b) a sequential Monte Carlo filter is employed to approximate the posterior probability of clean speech thus avoiding linearization of the non-linear observation model as in the case of algorithms that perform iterative approximations. The efficiency of the approach is illustrated when additive white Gaussian (AWGN) or babble noise is present in low signal-to-noise ratios (SNR).

1. Introduction

Although automatic speech recognition has reached the state of launching commercial products, the real-world environment is still a challenge for the available technology, due to the acoustic mismatch between training and operational conditions. The acoustic disparity is mainly due to different transmission channels used to access the recognition system and the variety of environmental conditions in which the communication takes place.

Sequential Monte Carlo sampling techniques (a.k.a particle filters) have been introduced to speech enhancement to estimate the parameters of a time-varying autoregressive process [1] or to sequentially estimate time-varying noise [2] among others. In this work, we demonstrate the applicability of particle filters for estimating the posterior probability of clean speech features. The features are in the log-spectral domain after the application of the Mel-filterbank and prior to the discrete cosine transform that maps them to cepstral coefficients. Recently, a series of successful feature domain enhancement techniques based on the same non-linear distortion model have been proposed [2]-[6]. The present work adopts this distortion model and introduces the following three extensions: a) the underlying speech log spectral vectors (i.e., the state vectors) are not assumed to be uncorrelated from frame to frame but they are modelled to evolve as a random walk model, b) the residual error in the log spectral domain from modeling the power spectrum of noisy speech as the sum of the power spectra of noise and speech (i.e., the observation error) as well as the state error are shown to be heavy-tailed and are modelled with a mixture of Gaussians instead of a single Gaussian as in [4]-[6], c) the non-linear model is not linearized to allow for the calculation of clean features; rather, a simulation based method is employed to approximate and track the posterior probability distribution of the log-spectral vectors of speech using random samples with associated weights. Finally, we obtain the minimum mean square error (MMSE) estimate of the log-spectra of clean speech.

2. State Space Modeling

Assuming noise is uncorrelated to speech and the channel is linear, time invariant and independent of the signal level, we can derive the power spectral representation of the time domain signal using the short time Fourier transform. This work does not take into account the channel effect, therefore:

$$|Y_k|^2 = |X_k|^2 + |N_k|^2 + 2|X_k||N_k|\cos\theta_k \qquad (1)$$

where $\cos\theta_k$ is the angle between X_k and N_k and k denotes the frequency band. After applying the Mel-scale filters which inflict a linear transform on the power spectrum (1) becomes:

$$|\hat{Y}_m|^2 = |\hat{X}_m|^2 + |\hat{N}_m|^2 + 2\lambda_m|\hat{X}_m||\hat{N}_m| \qquad (2)$$

where, m=1,2,…M denotes the filter bank channel and λ_m is a scalar value bounded between (-1, 1) (see e.g., [5]). After applying the logarithm on the filter bank energies we need to define the following vectors:

$$\mathbf{y} = \begin{bmatrix} \log|\hat{Y}_1|^2 \\ \log|\hat{Y}_2|^2 \\ \cdots \\ \log|\hat{Y}_M|^2 \end{bmatrix}, \ \mathbf{x} = \begin{bmatrix} \log|\hat{X}_1|^2 \\ \log|\hat{X}_2|^2 \\ \cdots \\ \log|\hat{X}_M|^2 \end{bmatrix}, \ \mathbf{n} = \begin{bmatrix} \log|\hat{N}_1|^2 \\ \log|\hat{N}_2|^2 \\ \cdots \\ \log|\hat{N}_M|^2 \end{bmatrix}.$$

The application of the logarithm leads to a non-linear equation that describes the distortion of the Mel log-power spectrum due to noise [2]-[6] for every frame t.

$$\mathbf{y}_t = \mathbf{x}_t + g(\mathbf{n}_t - \mathbf{x}_t) + \mathbf{w}_t, \text{ with } g(\mathbf{z}) = \log(\mathbf{1} + \exp\mathbf{z}) \qquad (3)$$

The measurement error can be easily shown to be $\mathbf{w}_t = \lambda/\cosh((\mathbf{n}_t - \mathbf{x}_t)/2)$ (see [4]-[6]) and is usually modeled as a zero mean Gaussian to simplify the mathematical formulation.

However, in section 2.1 we demonstrate that this error actually follows a heavy-tail distribution and is better modeled with a mixture of zero-mean Gaussians. The inference framework based on simulation techniques allows a more complex error model without the complication that a closed form solution or an approximate linearization of (3) would impose. Furthermore, the evolution of the clean features from frame to frame is allowed to have a correlation through a discrete-time random walk model:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{v}_t \quad (4)$$

where t is the frame index. In section 2.1 we demonstrate that the process noise $\mathbf{v}$ also possesses a heavy-tail distribution. Equations 3 and 4 introduce a state-space formulation where both measurement and process errors are modeled as mixtures of zero-mean Gaussians.

2.1. Modeling the Error Distributions

In order to investigate the empirical distribution of the residual error of (3) and (4), namely $\mathbf{w}$ and $\mathbf{v}$, we used a number of clean speech recordings from SpeechDat database (~ 30 min, silence part removed) which was subsequently corrupted with white Gaussian noise. Since $\mathbf{x}$, $\mathbf{y}$, $\mathbf{n}$ were made available for this corpora we calculated the distribution over all t of the errors $\mathbf{x}_t$-$\mathbf{x}_{t-1}$, and $\mathbf{y}_t$-($\mathbf{x}_t$+g($\mathbf{n}_t$-$\mathbf{x}_t$)). In Fig. 1a and Fig. 2a we depict the result of fitting a single Gaussian and a mixture of Gaussians to the normalized histogram of the process and measurement errors of a single filter bank output. One can observe the sub-optimality of the single Gaussian modeling approach in modeling the heavy-tailed empirical pdf. All log-power bands hold similar patterns for both error types. Subsequently, we derive QQ-plots of process and measurement error observations. The QQ-plot is an empirical plot of the ordered quantiles of a data set versus the quantiles of a standard Normal distribution. The purpose of the quantile-quantile plot is to determine whether the sample is drawn from a Gaussian distribution. If a normal QQ-plot is fairly linear this is an indication that the underlying pdf is normal. The examination of the QQ plots as depicted in Fig. 1b and Fig. 2b reveals that the errors are linear around the origin, indicating that the distributions are Gaussians around their mean. However, the direction of the curvature exhibited in Figs. 1b and 2b indicates that the data are non-Gaussian and that the distributions are in fact heavy-tailed. The observations in the tail region could have a considerable influence especially on those algorithms that make use of the covariance matrix which is quadratic [5]-[6]. The state and measurement errors are modelled as a Gaussian mixture with two zero-mean components and fixed variances $\psi_{0,s}$, $\psi_{1,s}$, and $\psi_{0,m}$, $\psi_{1,m}$ respectively. The first component having a small variance aims at capturing the Gaussian nature of the errors, while the component with the large variance is used to model the non-Gaussian nature of the errors. The variances are tuned using training corpora of uncorrupted speech features extracted from the SpeechDat data and a standard form of the expectation-maximization algorithm for training the Gaussian mixtures. Noise power on per-frame basis $\bar{\mathbf{n}}_t$ is estimated as described in section 2.3. The mixture distributions have the form:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = (1-\pi_s)N(\mathbf{x}_t; \mathbf{x}_{t-1}, \psi_{0,s}) + \pi_s N(\mathbf{x}_t; \mathbf{x}_{t-1}, \psi_{1,s}) \quad (5)$$

$$p(\mathbf{y}_t|\mathbf{x}_t) = (1-\pi_m)N(\mathbf{y}_t; \mathbf{x}_t + g(\bar{\mathbf{n}}_t\text{-}\mathbf{x}_t), \psi_{0,m}) + \pi_m N(\mathbf{y}_t; \mathbf{x}_t + g(\bar{\mathbf{n}}_t\text{-}\mathbf{x}_t), \psi_{1,m}) \quad (6)$$

where π_s , π_m denote the mixture weights, therefore, π_s , π_m ϵ

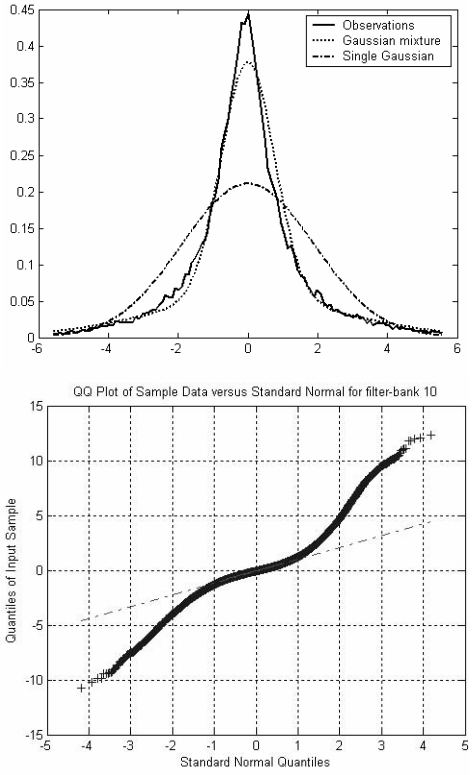

Figure 1: a) Normalized histogram of process error and fitted mixture of Gaussians and single Gaussian pdfs for filter bank 10, b) QQ plot of process error for filter bank 10

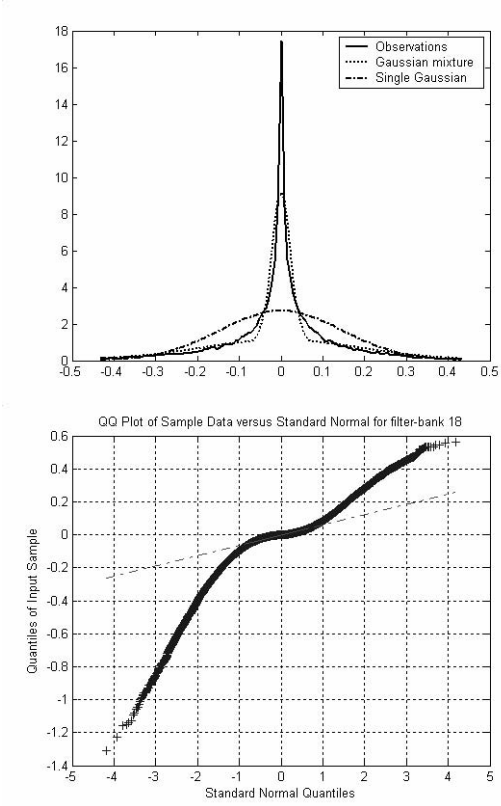

Figure 2: a) Normalized histogram of measurement error and fitted mixture of Gaussians and single Gaussian pdfs for filter bank 18, b) QQ plot of measurement error for filter bank 18

(0, 1). A random variable of this mixture distribution can be generated by first selecting uniformly a sample S from the interval (0, 1). If $S > \pi_s$, then $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is generated by an independent sample from $N(\mathbf{x}_t;\mathbf{x}_{t-1},\psi_{0,s})$; otherwise, the requested variable is a sample from $N(\mathbf{x}_t;\mathbf{x}_{t-1},\psi_{1,s})$.
The initial vector $\mathbf{x}_0$ is drawn from $p(\mathbf{x}_0)$ which follows the pdf of a 256 diagonal covariance components GMM trained with the 23 dimensional uncorrupted feature vectors extracted from the available 30 minutes speech recordings.
In this work we present experimental results for two cases,
a) each spectral band is treated as an independent time-series and a GMM is fitted to each band and,
b) the 23-dimensional log-spectral vector is partitioned to blocks of bands namely bands 1-6, 7-12, 13-18 and 19-23 (i.e., three 6-dimensional and one 5-dimensional block). Subsequently, the enhancement algorithm is applied to each block independently. The particle filter failed to provide reliable results when the GMMs of (5) and (6) model the 23-dimensional log-spectral state, a fact that is attributed to the high dimensionality of the state vector.

2.2. Bayesian Formulation

Based on the state equation (3) and (4), the optimal estimator in the minimum variance sense, is provided by $E[\mathbf{x}_t|\mathbf{y}_{1:t}]$. The problem of assessing this conditional expectation is strictly connected to the computation of $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ which is obtained by a two-step procedure of prediction and update. The prediction step is given by the Chapman-Kolmogorov equation (see e.g., [7]):

$$p\left(\mathbf{x}_t\middle|\mathbf{y}_{1:t-1}\right)=\int p\left(\mathbf{x}_t\middle|\mathbf{x}_{t-1}\right)p\left(\mathbf{x}_{t-1}\middle|\mathbf{y}_{1:t-1}\right)d\mathbf{x}_{t-1} \quad (7a)$$

and the update through the Bayes rule:

$$p\left(\mathbf{x}_t\middle|\mathbf{y}_{1:t}\right)=\frac{p\left(\mathbf{y}_t\middle|\mathbf{x}_t\right)p\left(\mathbf{x}_t\middle|\mathbf{y}_{1:t-1}\right)}{\int p\left(\mathbf{y}_t\middle|\mathbf{x}_t\right)p\left(\mathbf{x}_t\middle|\mathbf{y}_{1:t-1}\right)d\mathbf{x}_t} \quad (7b)$$

The pdf $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ is represented by a set of support points with associated normalized weights (i.e. discrete probability masses) $\left\{\mathbf{x}_t^i, w_t^i\right\}_{i=1}^N$ see e.g. [7]. The posterior density of the log power spectrum state at time t is approximated as

$$p\left(\mathbf{x}_t\middle|\mathbf{y}_{1:t}\right)\simeq\sum_{i=1}^N w_t^i\delta\left(\mathbf{x}_t-\mathbf{x}_t^i\right)$$

where $w_t^i \propto w_{t-1}^i p\left(\mathbf{y}_t\middle|\mathbf{x}_t^i\right)$. Subsequently the weights are normalized so that $\sum_i w_t^i=1,\ i=1, 2,..., N$. The optimal estimate of state $\mathbf{x}_t$ in the MMSE sense is calculated by:

$$\hat{\mathbf{x}}_t=\int_{\mathbf{x}_t}\mathbf{x}_t p\left(\mathbf{x}_t\middle|\mathbf{y}_{1:t}\right)d\mathbf{x}_t\simeq\sum_{i=1}^N w_t^i\mathbf{x}_t^i$$

To reduce the effect of degeneracy, a resampling procedure is applied in order to eliminate particles that have small weights and to concentrate on particles with large weights [7].
A lowest threshold is applied to the sampling procedure to restrict probagation of particles below a certain log-power value, since very low values achieved in the silence parts are found to affect the cepstral mean normalisation procedure. This threshold is tuned based on the mean of the minimum log-power value observed in the uncorrupted recordings of the speech database used to train the recognition engine. The particle filtering procedure in the framework of the log-power spectrum enhancement of speech is described in Table 1.

Draw $\mathbf{x}_0 \sim p(\mathbf{x}_0)$
FOR i=1 to N particles
- Draw $\mathbf{x}_t^i \sim p\left(\mathbf{x}_t\middle|\mathbf{x}_{t-1}^i\right)$ according to (5)
- $\mathbf{x}_t^i=\max\left(thres,\ \mathbf{x}_t^i\right)$ (Mel log-power threshold)
- Assign each particle a weight $w_t^i=p\left(\mathbf{y}_t\middle|\mathbf{x}_t^i\right)$ using (6)

END FOR
- Normalise the weights so that $w_t^i=\frac{p\left(\mathbf{y}_t\middle|\mathbf{x}_t^i\right)}{\sum_i^N p\left(\mathbf{y}_t\middle|\mathbf{x}_t^i\right)}$
- Use w_t^i to resample with replacement the particles $\left\{\mathbf{x}_t^i\right\}$
- Output $\mathbf{x}_t$ as the mean value of the resampled set

Table 1: Generic particle filter for log-spectrum enhancement

2.3. Noise Estimation

The proposed framework requires the estimation of Mel-log noise power. We made use of a statistical voice activity detector (VAD) as described in [8]. This VAD makes the assumption that the DFT coefficients of speech and noise are asymptotically independent, circular, zero mean Gaussian random processes. The noise estimation task proceeds with the consideration that in each frame speech can be absent (hypothesis H_0) or present (hypothesis H_1). The likelihood ratio for the k frequency band of linear spectrum and frame t is based on the probability density functions conditioned on H_0 and H_1 namely $p(X_{t,k}|H_0)$ and $p(X_{t,k}|H_1)$ and is given by

$$L_{t,k}\triangleq\frac{p\left(X_{t,k}\middle|H_1\right)}{p\left(X_{t,k}\middle|H_0\right)}=\frac{1}{1+\xi_{t,k}}\exp\left(\frac{\gamma_{t,k}\xi_{t,k}}{1+\xi_{t,k}}\right)$$

where $\xi_{t,k}=a\frac{X_{t-1,k}^2}{N_{t-1,k}}+(1-a)\max\left(0,\gamma_{t,k}-1\right)$ and a=0.98. The $M=50$ initial frames of each recording are assumed noise only and $N_{1,k}=(1/M)\sum_{t=1}^M\left|N_{t,k}\right|^2$. For the function of VAD and assuming that noise power is slowly varying we set $\gamma_{t,k}=Y_{t,k}^2/N_{t-1,k}$. The term $X_{t,k}$ can be calculated by applying the Ephraim-Malah gain function [8] but for the need of VAD we found simple spectral subtraction equally efficient while much faster. The VAD decision on per-frame basis is based on the mean of the likelihood ratios for the individual frequency bands, which is given by:

$$\text{if } \frac{1}{D}\sum_{k=1}^D\log L_{t,k}<\eta \text{ accept } H_0 \text{ otherwise accept } H_1$$

where D is the number of spectral bands and η is the log-likelihood threshold derived from the initial noise-only segments. When a frame is labeled as noise-only the spectral power vector is passed unaltered to the noise vector. During speech presence the noise variance is updated using the following expression:

$$N_{t,k}=\beta N_{t-1,k}+(1-\beta)\left(Y_{t,k}-X_{t,k}\right). \quad (8)$$

where β=0.8. Subsequently the noise vector is passed through the Mel-filterbank to provide the estimation of the log noise power $\bar{\mathbf{n}}_t$ that is used in (6).

3. Speech Recognition Experiments

For the evaluation of the proposed algorithm we used a speech recognition module built with HTK Hidden Markov Models toolkit. The basic recognition units are tied state, context dependent triphones of five states each. In order to train the reference recogniser we used the SpeechDat-II database of utterances and their associated transcriptions [9]. We made use of the utterances taken from the C3 credit card number corpus and L1-L3 spelled words corpus of the SpeechDat database in order to train our system excluding the testing set comprised of 600 randomly chosen recordings. Each input speech signal waveform is sampled at 8kHz, band-passed between 300 and 3400 Hz, pre-emphasized by the filter $H(z)=1-0.97z^{-1}$ and subsequently, windowed into frames of 256 points using a Hamming window and 50% frame shift. Each frame processed so far is Fourier transformed using 512 points FFT and the power of the transformation is passed through a set of 23 Mel-spaced triangular band-pass filter-bank channels. Subsequently the enhancement technique is applied to the log-spectral power using 200 particles with systematic resampling as regards the 1-dimensional case and 1800 particles for the case of splitting the 23 log-power bands into three 6-dimensional blocks of bands and one 5-dimensional block. 13-dimensional feature vectors are formed after applying DCT to the log-filter-bank enhanced outputs, which reduces the 23 output channels into 13-dimensional Mel frequency cepstral coefficients. Cepstral mean normalization was applied to deal with the linear channel assumption. The 13 aforementioned coefficients form the final observational vector that is passed to the recognition engine. Deltas and acceleration coefficients are not appended primarily because we want to obtain comparative results on an enhancement technique that functions on the static log-spectrum.

The results in Fig. 3 and Fig. 4 demonstrate that the proposed technique can provide substantially improved recognition results at low SNRs. The treatment of spectral bands as independent time-series surprisingly provides better results than the four-block case that allows to model correlation between spectral components inside each block, but increases the dimensionality of the state vector. The algorithm is observed to be sensitive to noise-estimation errors, a fact that results to a small degradation in almost clean conditions for the four-block case. We are currently investigating ways of incorporating speech and noise variance estimation in a common sampling framework as well as Gibbs sampling and Markov random field priors. Moreover, our approach can be extended by employing more accurate speech models as HMMs that comply naturally with the proposed sampling framework.

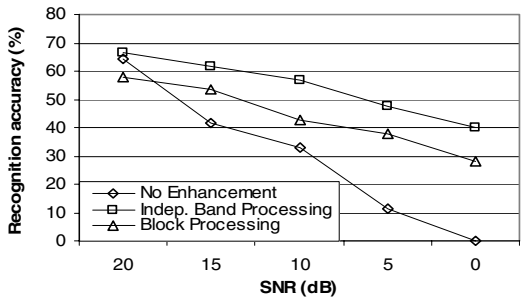

Figure 3: Word Recognition Acc. (%) under Gaussian noise corruption.

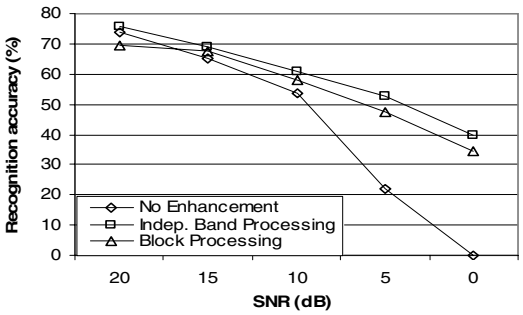

Figure 4: Word Recogn. Acc. (%), Babble noise corruption.

4. Conclusions

This work analyzed the performance of a novel, statistical, speech enhancement approach that functions on the Mel log-power domain and aims primary at robust speech recognition. In our approach, a random walk model and a well-known non-linear speech distortion model in the log-power feature domain form a state space formulation. We proposed the application of a sampling inference technique to derive the MMSE estimation given the noisy feature observations. The results demonstrate that the approximation of the corresponding distributions of the inference framework achieve significant improvement in recognition accuracy even in very low SNRs.

5. References

[1] Vermaak, J., Andrieu, C. and Godsill, S., "Particle methods for Bayesian modeling and enhancement of speech signals", *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 173-185, 2002.

[2] Yao, K. and Lee, Te-W., "Time-varying noise estimation for speech enhancement and recognition using sequential Monte Carlo method", *EURASIP Journal on Applied Signal Processing*, 15, pp. 2366-2384, 2004.

[3] Moreno, P., Raj, B. and Sterrn, R., "A vector Taylor series approach for environment-independent speech recognition", *Proc. ICASSP*, vol. 1, pp. 733-736, 1996.

[4] Frey, B., Deng, L., Acero, A. and Kristjansson, T., "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition", in *Proc. European Conf. Speech Communication*, Aalborg, Denmark, pp. 901-904, 2001.

[5] Deng, L., Droppo, J. and Acero, A., "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features", *IEEE Trans. on Speech and Audio Proc.*, 12(3), pp. 218-232, 2004.

[6] Deng, L., Droppo, J. and Acero, A., "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition", *IEEE Trans. SAP.*, vol. 11, no. 6, pp. 568-580, 2003.

[7] Arulampalam, M., Maskell, S., Gordon, N. and Clapp, T., "A tutorial on particle filters for online non-linear, non-Gaussian Bayesian tracking", *IEEE Transactions on Signal Processing,* vol. 50, no. 2, pp. 174-188, 2002.

[8] Sohn J., Kim N. and Sung W., "A statistical model-based voice activity detection", *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, 1999.

[9] Van den Heuvel H., Moreno, A., Omologo, M., Richard, G., Sanders, E., "Annotation in the SpeechDat projects", *Intern. Journal of Speech Techn.,* 4(2), p. 127-143, 2001.