INTEGRATION OF THE PREDICTED WALK MODEL ESTIMATE INTO THE PARTICLE FILTER FRAMEWORK

Matthias Wölfel

Institut für theoretische Informatik, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
wolfel@ira.uka.de

ABSTRACT

Distortion robustness is one of the most significant problems in automatic speech recognition. While a lot of research in speech feature enhancement in automatic recognition has focused on stationary distortions, most of the observed distortions are non-stationary. To cope with the non-stationary behavior, just recently, various particle filter approaches have been proposed to track the non-stationary distortions on speech features in logarithmic spectral or cepstral domain. Most of those techniques rely on the prediction of the noise evolution model by a linear prediction matrix. The current estimation of the linear prediction matrix, however, needs noise only observations which have to be either given a priori or to be detected by voice activity detection. This makes it impossible to adapt the linear prediction matrix to the dynamics of the noise on speech regions. In this publication we propose to estimate or update the linear prediction matrix directly on the noisy speech observations. This is possible within the particle filter framework by weighting the different noisy estimates (particles) due to their likelihood in the estimation equation of the linear prediction matrix.

Speech recognition experiments on actual recordings with different speaker to microphone distances confirm the soundness of the proposed approach.

Index Terms— speech feature enhancement, particle filter, predicted walk, linear prediction matrix, automatic speech recognition

1. INTRODUCTION

Speech feature enhancement can be formulated as a tracking problem where the clean speech features $\mathbf{x}_k$ have to be estimated, for each frame k, given the observation history of the noisy speech features $\mathbf{y}_{1:k}$. The clean and noisy features are related by the probabilistic relationship $p(\mathbf{x}_{0:k}|\mathbf{y}_{1:k})$. As stated in Julier and Uhlmann [1] the minimum mean square error solution to such a tracking problem consists in finding the conditional mean

$$\mathbb{E}\left\{\mathbf{x}_{0:k}|\mathbf{y}_{1:k}\right\} = \int \mathbf{x}_{0:k} p(\mathbf{x}_{0:k}|\mathbf{y}_{1:k}) d\mathbf{x}_{0:k}.$$

Assuming that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is a Markov process and that the current observation is only dependent on the current state facilitates sequential calculation of the conditional mean, the solution is given by

$$\mathbb{E}\left\{\mathbf{x}_k|\mathbf{y}_{1:k}\right\} = \int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{y}_{1:k}) d\mathbf{x}_k.$$

Introducing the noise $\mathbf{n}_k$ as a hidden variable

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}) = \int p(\mathbf{x}_k, \mathbf{n}_k|\mathbf{y}_{1:k}) d\mathbf{n}_k$$

with the relation $p(\mathbf{x}_k, \mathbf{n}_k|\mathbf{y}_{1:k}) = p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k) p(\mathbf{n}_k|\mathbf{y}_{1:k})$ and a changed integration order we obtain

$$\mathbb{E}\left\{\mathbf{x}_k|\mathbf{y}_{1:k}\right\} = \int \underbrace{\int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k) d\mathbf{x}_k}_{=h_k(\mathbf{n}_k)} p(\mathbf{n}_k|\mathbf{y}_{1:k}) d\mathbf{n}_k. \quad (1)$$

The filtering density $p(\mathbf{n}_k|\mathbf{y}_{1:k})$ keeps track of the probability throughout time which can be realized by sequential updating

$$p(\mathbf{n}_k|\mathbf{y}_{1:k}) = \frac{p(\mathbf{n}_k, \mathbf{y}_k|\mathbf{y}_{1:k-1})}{p(\mathbf{y}_k|\mathbf{y}_{1:k-1})} \quad (2)$$

with

$$p(\mathbf{n}_k, \mathbf{y}_k|\mathbf{y}_{1:k-1}) = \quad p(\mathbf{y}_k|\mathbf{n}_k) \cdot \int p(\mathbf{n}_k|\mathbf{n}_{k-1}) p(\mathbf{n}_{k-1}|\mathbf{y}_{1:k-1}) d\mathbf{n}_{k-1}$$

and

$$p(\mathbf{y}_k|\mathbf{y}_{1:k-1}) = \int p(\mathbf{n}_k, \mathbf{y}_k|\mathbf{y}_{1:k-1}) d\mathbf{n}_k$$

To avoid intractable integration, we aim to construct the empirical density $p(\mathbf{n}_k|\mathbf{y}_{1:k})$ by Monte Carlo sampling [2]. The previous two equations can then be expressed as

$$p(\mathbf{n}_k, \mathbf{y}_k|\mathbf{y}_{1:k-1}) \approx \frac{1}{M} \sum_{m=1}^{M} p(\mathbf{y}_k|\mathbf{n}_k^{(m)}) p(\mathbf{n}_k|\mathbf{n}_{k-1})$$

and

$$p(\mathbf{y}_k|\mathbf{y}_{1:k-1}) \approx \frac{1}{M} \sum_{m=1}^{M} p(\mathbf{y}_k|\mathbf{n}_k^{(m)})$$

where $p(\mathbf{y}_k|\mathbf{n}_k^{(m)})$ represents the corresponding importance weight $\omega_k^{(m)}$ for each sample m, and $p(\mathbf{y}_k|\mathbf{y}_{1:k-1})$ represents a normalization term.

The solution of (2) requires a model of the evolution of noise $p(\mathbf{n}_k|\mathbf{n}_{k-1})$. A simple solution is the so called random walk $\mathbf{n}_{k+1} = \mathbf{n}_k + \epsilon$ where ϵ describes random variations. However, it might result in an unstable prediction, because the random walk process provides a state transition of the target signal using random noise which might be a bad approximation of the true noise. Thus, to improve the estimate of the evolution one aims to use a model which predicts the noise observation $\mathbf{n}_{k+1}$ given the noise observation sequence $\mathbf{n}_{1:k}$. One way to predict the noise has been proposed by [3] where the noise transition has been modeled by an extended Kalman filter which has been augmented with *Polyak averaging* and feedback. If the noise is moving slowly, then the difference between the Polyak average and noise has a small value and thus the parameter range gets small and vice versa. Alternatively it is possible to use a predicted walk which is modeled by a linear prediction matrix [4]. A

major drawback of the linear prediction matrix is that it has to be calculated on noise only regions which have to be either given a prior or to be detected by voice activity detection which might be unreliable in particular for noisy signals. In addition, with the given approach, it is impossible to adapt the linear prediction matrix to the dynamics of the noise in speech regions. To overcome those drawbacks we aim on estimating or updating the linear prediction matrix on noisy speech observations. This can be established within the particle filter framework by weighting the different noise estimates due to their likelihood in the estimation equation of the linear prediction matrix.

2. BRIEF REVIEW OF SPEECH FEATURE ENHANCEMENT BY PARTICLE FILTERS

Different approaches to speech feature enhancement by particle filters exist. We follow Singh and Raj [4] who have proposed to track the noise frame by frame in the logarithmic spectral domain and later on subtract the noise estimates from the contaminated speech signal. An extended algorithm of the original approach as stated by Singh and Raj can be outlined as follows:

1. *Draw noise samples*
   At the start frame $k = 0$, M particles (noise hypotheses) $\mathbf{n}_0^{(m)}$ $(m = 1, ..., M)$ are drawn from the prior noise density $p_{\text{noise}}(\mathbf{n})$ which is modeled as a Gaussian mixture model.
   For frames $k > 0$, M particles $\mathbf{n}_k^{(m)}$ are sampled from the noise transition probability $p(\mathbf{n}_k|\mathbf{n}_{k-1})$ where different models for $p(\mathbf{n}_k|\mathbf{n}_{k-1})$ are described in Section 3.
2. *Evaluate noise samples*
   The importance weight for each particle $\mathbf{n}_k^{(m)}$ is evaluated according to the likelihood
   $$p(\mathbf{y}_k|\mathbf{n}_k^{(m)}) = \frac{p_{\text{speech}}(\mathbf{y}_k + \log(\mathbf{1} - e^{\mathbf{n}_k^{(m)} - \mathbf{y}_k})}{\prod_{d=1}^{D} \left|1 - e^{\hat{\mathbf{n}}_{k,d}^{(m)} - \mathbf{y}_{k,d}}\right|} \quad (3)$$
   where $p_{\text{speech}}(\cdot)$ denotes the prior speech density represented by the Gaussian mixture model which has been trained on clean speech. Thereafter the importance weights are normalized by
   $$\tilde{\omega}_k^{(m)} = \frac{p(\mathbf{y}_k|\mathbf{n}_k^{(m)})}{\sum_{m=1}^{M} p(\mathbf{y}_k|\mathbf{n}_k^{(m)})}.$$
   Note that the likelihood $p(\mathbf{y}_k|\mathbf{n}_k^{(m)})$ can only be evaluated if $n_{k,d}^{(m)} < y_{k,d}$ $\forall$ dimensions $d = 1, 2, \ldots, D$. If the noise exceeds the noisy observation in just a single spectral bin the given noise hypothesis has to be rejected by setting the particle weight to zero. This causes a decimation of the particle population which can be remedied by a *fast acceptance test* [5] that virtually boosts the number of particles by redrawing samples in case of rejection.
3. *Compensate for noise estimates*
   Different methods to compensate for noise densities exist, e.g. one popular method is the vector Taylor series [6]. However, clean speech spectra can be estimated by using the discrete Monte Carlo representation of the continuous filtering density and a direct calculation, the so called *statistical inference approach* (SIA) [5], can be applied
   $$\begin{aligned} h_k^{\text{SIA}}(\mathbf{n}_k) &= \int \mathbf{x}_k \delta_{\mathbf{y}_k + log(\mathbf{1} - e^{\mathbf{n}_k - \mathbf{y}_k})}(\mathbf{x}_k) d\mathbf{x}_k \\ &= \mathbf{y}_k + log(\mathbf{1} - e^{\mathbf{n}_k - \mathbf{y}_k}) \end{aligned} \quad (4)$$
   where $\delta_{\mathbf{n}_k^{(m)}}$ denotes a translated Dirac delta function.
4. *Resample noise*
   The normalized weights are used to resample among the noise hypotheses $\mathbf{n}_k^{(m)}$ $(m = 1, ..., M)$ [7, 8]. This can be regarded as a pruning step where likely hypotheses are multiplied and unlikely ones are removed from the population.

Those steps are repeated with $k \mapsto (k + 1)$ until all time-frames are processed.

Working Domain

Particle filters for speech feature enhancements are typically applied in the logarithmic spectral domain after dimension reduction by mel-filterbanks. Due to the properties of the used spectral estimation method provided by warped minimum variance distortionless response [9], no filterbank is applied and thus the dimension in the logarithmic spectral domain is not reduced. As the operation of a particle filter with high dimensions (in our case 129) would be infeasible or very slow, we decided to work in the logarithmic spectral domain after cepstral truncation to 20 dimensions by applying an inverse Fourier transformation to the cepstral coefficients. In the *truncated* logarithmic spectral domain the relation between the noisy observation $\mathbf{y}$, the clean feature $\mathbf{x}$ and noise $\mathbf{n}$ can be approximated by

$$\mathbf{x} \approx \log(e^{\mathbf{y}} - e^{\mathbf{n}}) = \mathbf{y} + \log(1 - e^{\mathbf{n} - \mathbf{y}}). \quad (5)$$

3. EVOLUTION OF THE NOISE SPECTRA (SAMPLING)

As seen in previous sections particle filter tracking application requires the prediction of the noise $\hat{\mathbf{n}}_k = p(\mathbf{n}_k|\mathbf{n}_{1:k-1})$ given the trajectory of the noise up to time $k - 1$. The noise transition probability $p(\mathbf{n}_k|\mathbf{n}_{1:k-1})$ can be modeled by a dynamic system model which can be classified into *random walk* and *predicted walk*.

In this section we review the random walk model and the predicted walk model represented by a static autoregressive process which has to be estimated on noise only regions. In addition we propose a novel method which can give an instant, and thus dynamic, estimate of the predicted walk model on a frame by frame basis which can be estimated or updated on noise only as well as noisy speech observations. Thus, the proposed approach is able to overcome the problems associated with predicted walk models which have to be determined on noise only regions.

3.1. Random Walk

The simplest way to model the evolution of noise features is a random walk

$$\hat{\mathbf{n}}_k = \mathbf{n}_{k-1} + \varepsilon_k$$

where $\mathbf{n}_k$ denotes the noise spectrum at time k while the ε_k terms are considered to be i.i.d. zero mean Gaussian, i.e. $\varepsilon_k \sim \mathcal{N}(0, \Sigma_{noise})$, where the covariance matrix Σ_{noise} is assumed to be Gaussian.

3.2. Predicted Walk by static autoregressive processes

To consider information about the evolution of the noise, Raj *et al.* [10] proposed and investigated to use a lth-order autoregressive process $\mathbf{A}^{(1:l)}$ to predict the evolution of the noise

$$\begin{aligned} \hat{\mathbf{n}}_k &= \mathbf{A}^{(1)}\mathbf{n}_{k-1} + \mathbf{A}^{(2)}\mathbf{n}_{k-2} + \cdots + \mathbf{A}^{(l)}\mathbf{n}_{k-m} + \varepsilon_k \\ &= \mathbf{A}^{(1:l)}\mathbf{n}_{k-1:k-l} + \varepsilon_k. \end{aligned}$$

Learning the Autoregressive Noise Model

The autoregressive noise model consists of two components that have to be learned for a specific type of noise:

- the *linear prediction transition matrix* $\mathbf{A}^{(1:l)}$ and
- the *covariance matrix* $\mathbf{\Sigma}_{\text{noise}}$ where once again the ε_k terms are considered to be i.i.d. zero mean Gaussians.

Minimization of the prediction error norm results in the following estimate of the linear prediction matrix:

$$\mathbf{A}^{(1:l)} = \mathbb{E}[\mathbf{n}_k \mathbf{N}_{k-1:k-l}^T]\mathbb{E}[\mathbf{n}_{k-1:k-l}\mathbf{N}_{k-1:k-l}^T]^{-1} \quad (6)$$

Those matrices can be derived from the noise data $1, 2, \ldots, K$ as

$$\mathbb{E}[\mathbf{n}_k \mathbf{N}_{k-1:k-l}^T] \;=\; \frac{1}{K}\sum_{k=l}^{K} \mathbf{n}_k \mathbf{N}_{k-1:k-l}^T$$

and

$$\mathbb{E}[\mathbf{n}_{k-1:k-l} \mathbf{N}_{k-1:k-l}^T] \;=\; \frac{1}{K}\sum_{k=l}^{K} \mathbf{n}_{k-1:k-l} \mathbf{N}_{k-1:k-l}^T.$$

Note that it is sufficient to estimate the matrices from pieces of noise as long as the pieces are long enough to contain enough history. In our experiments we have train the linear prediction matrix on 150 seconds of noise only pieces, collected on silent regions among 35 minutes of speech which has been found by voice activity detection.

To learn a linear prediction matrix of model order length l requires $d^2 l$ coefficients to be reliably estimated which can only be established if a huge amount of training data is available. For a reasonable amount of training data only a small reduction in the mean square error can be reached by using higher order models. Thus, a first model order is sufficient for our investigations.

The diagonal covariance can be learned by

$$\sigma_d^2 := \mathbb{E}[(n_{k,d} - \hat{n}_{k,d})^2],$$

where $n_{k,d}$ denotes the dth vector component of the noise $\mathbf{n}_k$ and $\hat{n}_{k,d}$ denotes the dth vector component of the predicted noise $\hat{\mathbf{n}}_k = \mathbf{A}^{(1:l)}\mathbf{N}_{k-1:k-l}$. However, in practice, we have yielded better results by manually increasing the variance with identical values over all dimensions.

3.3. Predicted Walk by dynamic autoregressive processes

In this section, instead of estimating the linear prediction matrix previous to the application of the particle filter based on a priori knowledge of the noise or noise pices found by voice activity detection, we aim for an instantaneous and integrated estimate of the linear prediction matrix. Thus, we have to solve for the minimization of the prediction error norm for each frame k by estimating the linear prediction matrix as

$$\mathbf{A}_k = \mathbf{A}_k^{(1)} = \mathbb{E}[\mathbf{n}_k \mathbf{n}_{k-1}^T]\mathbb{E}[\mathbf{n}_{k-1}\mathbf{n}_{k-1}^T]^{-1}. \quad (7)$$

Instead of deriving the two matrices from noise only frames, as demonstrated in Section 3.2, we estimate the matrices on the current $\mathbf{n}_k^{(m)}$ and previous $\mathbf{n}_{k-1}^{(m)}$ noise estimates for all particles $m = 1, 2, \ldots$,M. To ensure that the prediction estimates which lead to a good noise estimate are emphasized and those predictions who lead to a poor estimate are alleviated, we have to weight the contribution of each particle to the matrices due to their likelihood $p(\mathbf{y}_k|\mathbf{n}_k^{(m)})$ as given in (3). Thus, the matrices can be evaluated for each frame k by using

$$\mathbb{E}[\mathbf{n}_k \mathbf{n}_{k-1}^T] \;=\; \frac{1}{M}\sum_{m=1}^{M} p(\mathbf{y}_k|\mathbf{n}_k^{(m)}) \cdot \mathbf{n}_k \cdot {\mathbf{n}_{k-1}^{(m)}}^T$$

and

$$\mathbb{E}[\mathbf{n}_{k-1} \mathbf{n}_{k-1}^T] \;=\; \frac{1}{M}\sum_{m=1}^{M} p(\mathbf{y}_k|\mathbf{n}_k^{(m)}) \cdot \mathbf{n}_{k-1} \cdot {\mathbf{n}_{k-1}^{(m)}}^T$$

to solve for (7).

A smoothing over previous frames might help to improve the reliability of the estimate. With the introduction of the forgetting factor α we can write the smoothed matrices $\overline{\mathbf{A}}_k$ with

$$\overline{\mathbb{E}}[\mathbf{n}_k \mathbf{n}_{k-1}^T] \;=\; \alpha\mathbb{E}[\mathbf{n}_k \mathbf{n}_{k-1}^T] + (1-\alpha)\overline{\mathbb{E}}[\mathbf{n}_{k-1}\mathbf{n}_{k-2}^T]$$

and

$$\overline{\mathbb{E}}[\mathbf{n}_{k-1} \mathbf{n}_{k-1}^T] \;=\; \alpha\mathbb{E}[\mathbf{n}_{k-1} \mathbf{n}_{k-1}^T] + (1-\alpha)\overline{\mathbb{E}}[\mathbf{n}_{k-2}\mathbf{n}_{k-2}^T].$$

The sample variance can now be calculated according to the normalized weight $\tilde{\omega}_k^{(m)}$ of each particle as

$$\sigma_k^2 := \sum_{m=1}^{M} \tilde{\omega}_k^{(m)}(\mathbf{n}_k^{(m)} - \hat{\mathbf{n}}_k^{(m)})^2 = \sum_{m=1}^{M} \tilde{\omega}_k^{(m)}(\mathbf{n}_k^{(m)} - \mathbf{A}_k\mathbf{n}_{k-1}^{(m)})^2. \quad (8)$$

Note that the subscript d representing each vector component of the noise $\mathbf{n}_k$ at frame k has been neglected to improve readability.

4. EXPERIMENTS

In order to evaluate the performance of the proposed particle filter enhancements under realistic conditions we have recorded 35 minutes of lecture speech with different microphone types and speaker to microphone distances (similar to RT-06s development and evaluation data [11]).

As a speech recognition engine we used the *Janus Recognition Toolkit* (JRTk) with the same setup as described in [12]: The acoustic training material, approximately 100 hours, used for the experiments reported here was taken from the ICSI, NIST, and CMU meeting corpora, as well as the *Translanguage English Database* (TED) and CHIL lecture corpora resulting in a discriminatively trained semi-continuous quint phone systems that contain 16000 distributions over 4000 codebooks, with a maximum of 64 Gaussians per model. The 3-gram language model contains approximately 23,000 words and has a perplexity of 125 on the test corpora. The used warped minimum variance distortionless response cepstral coefficients [9] have been shown to outperform mel frequency cepstral coefficients [13] in combination with and without speech feature enhancement. The particle filter has used 100 particles with the fast acceptance test and a fixed, identical variance for all evaluated sampling techniques, unless stated otherwise.

We evaluated on unadapted (first pass) acoustic models and acoustic models (second pass) which have been unsupervised adapted by *maximum likelihood linear regression* (MLLR), constrained MLLR and *vocal track length normalization* (VTLN). The determined VTLN factors have also been used in the second pass of the particle filter.

Comparing the different word error rates on actual recodings with different speaker to microphone distances, given in Table 1,

Microphone	CTM		Lapel		Table Top		Wall	
Distance	5 cm		20 cm		100–150 cm		300–350 cm	
SNR	24 dB		23 dB		17 dB		10 dB	
Pass	1	2	1	2	1	2	1	2
Particle Filter	Word Error Rate							
no particle filter	11.6%	09.8%	11.7%	09.9%	19.0%	14.6%	45.6%	29.0%
random walk	12.3%	09.8%	12.0%	09.8%	20.7%	14.5%	46.5%	26.3%
predicted walk (static)	11.6%	09.4%	11.6%	09.8%	19.3%	13.9%	43.5%	25.7%
predicted walk (dynamic[1])	11.4%	09.7%	11.6%	09.5%	17.9%	13.1%	44.3%	25.8%
predicted walk (dynamic[2])	11.5%	09.9%	12.0%	09.8%	18.3%	13.7%	43.5%	25.7%

Table 1. Word error rates without particle filter and particle filter enhanced features using different sampling strategies. [1] *fixed variance*, [2] *variance determined on speech frames (8)*

indicates that the proposed approach (dynamic) can reach at least equal performance as compared to the previous approach (static). The proposed approach has several advantages. Namely it can be used in runtime systems, as it gives an instant estimate of the linear prediction matrix and it offers a reliable estimate without the need for silent regions. To determine the variance on the noise hypothesises (dynamic[2]) can not improve the performance over a fixed, in average higher, variance (dynamic[1]) which is consistent to previous experiments and might be explained by the enlarged search space. In addition we played with various smoothing values for α, however, were not able to improve (in average over all speaker to microphone distances) over $\alpha = 0$.

5. CONCLUSIONS

We have introduced an instantaneous and integrated approach to estimate the linear prediction matrix which is used in the predicted walk model. In comparison to the previous estimation of the linear prediction matrix which has to rely on noise only regions, the proposed model performs at least equally well and has several advantages on hand: It can be used in runtime systems, as it gives an instantaneous estimate of the matrix and offers a reliable estimate without the need for noise only regions.

6. REFERENCES

[1] S. Julier and J.K. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions," *tech. rep., RRG, Dept. of Engineering Science, University of Oxford*, Nov. 1996.

[2] W.K. Hastings, "Monte carlo sampling methods using markov chain and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Jan. 1970.

[3] M. Fujimoto and S. Nakamura, "Particle filter and polyak averaging-based non-stationary noise tracking for ASR in noise," *Proc. of ASRU*, 2005.

[4] R. Singh and B. Raj, "Tracking noise via dynamical systems with a continuum of states," *Proc. of ICASSP*, 2003.

[5] F. Faubel and M. Wölfel, "Overcoming the vector tailor series approximation in speech feature enhancement – a particle filter approach," *Proc. of ICASSP*, 2007.

[6] P.J. Moreno, B. Raj, and R.M. Stern, "A vector taylor series approach for environment-independent speech recognition," *Proc. of ICASSP*, 1996.

[7] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," *Proc. of Radar and Signal Processing*, vol. 140, pp. 107–113, Sep. 1993.

[8] A. Doucet, *On Sequential Simulation-Based Methods for Bayesian Filtering, Technical report CUED/F-INFENG/TR 310*, Cambridge University Department of Engineering, 1998.

[9] M. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[10] B. Raj, R. Singh, and R. Stern, "On tracking noise with linear dynamical system models," *Proc. of ICASSP*, 2004.

[11] J.G. Fiscus, J. Ajot, M. Michel, and J.S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," *Proc. of Machine Learning for Multimodal Interaction, S. Renals, S. Bengio, and J.G. Fiscus (Eds.), LNCS vol. 4299, Springer*, pp. 309–322, 2006.

[12] M. Wölfel, S. Stüker, and F. Kraft, "The ISL RT-07 speech-to-text system," *In Proc. of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop (RT-07), Baltimore, USA*, 2007.

[13] F. Faubel and M. Wölfel, "Coupling particle filters with automatic speech recognition for speech feature enhancement," *Proc. of Interspeech*, Sep. 2006.