An Approach to Iterative Speech Feature Enhancement and Recognition

Stefan Windmann, Reinhold Haeb-Umbach

Department of Communication Engineering, University of Paderborn, Paderborn, Germany
windmann@nt.uni-paderborn.de, haeb@nt.uni-paderborn.de

Abstract

In this paper we propose a novel iterative speech feature enhancement and recognition architecture for noisy speech recognition. It consists of model-based feature enhancement employing Switching Linear Dynamical Models (SLDM), a hidden Markov Model (HMM) decoder and a state mapper, which maps HMM to SLDM states. To consistently adhere to a Bayesian paradigm, posteriors are exchanged between these processing blocks. By introducing the feedback from the recognizer to the enhancement stage, enhancement can exploit both the SLDMs ability to model short-term dependencies and the HMMs ability to model long-term dependencies present in the speech data. Experiments have been conducted on the Aurora II database, which demonstrate that significant word accuracy improvements are obtained at low signal-to-noise ratios.
Index Terms: speech recognition, speech feature enhancement, SLDM

1. Introduction

Robust speech recognition in noisy environments remains a tough research challenge while at the same time being of great practical importance. Provisions for noise robustness have been attempted at various stages of the recognizer. Among the most promising approaches are model-based speech feature enhancement techniques, of which the Vector Taylor Series approach is one of the best known examples [1]. More recently, Droppo and Acero have proposed the use of a Switching Linear Dynamical Model (SLDM) for the enhancement of noisy speech features in the cepstral domain, where the Gaussian Pseudo Bayesian algorithm of order one (GPB1) is used for inference [2]. Similar methods have been proposed by [3] and [4]. In these approaches the enhancement of the feature vectors and the actual speech recognition were considered to be two strictly separated stages. In a recent work of Faubel and Wölfel this strict separation was removed [5]. They proposed to track the noise with a particle filter, while the distribution of the speech was determined by the current phoneme in the recognizer.

Rosti and Gales have investigated whether the HMM in the recognizer can be replaced by a SLDM, where the inference in the SLDM is either performed probabilistically by Gibbs sampling or deterministically by GPB or other approaches [6]. Overall they came to the conclusion that the use of an HMM in the recognizer is more suitable than the application of an SLDM.

While a well-known strength of the HMM is the modeling of non-linear long-term dependencies, a SLDM is more suitable for modeling short-term dependencies as it does not suffer from the conditional independence assumption used in HMMs. On the other hand finding the optimal state sequence is computationally intractable for the SLDM. It therefore seems promising to look for approaches, where the two, SLDM for feature enhancement and HMM for recognition, can benefit from each other's complementary modeling strength. While it was shown before that the HMM recognizer improves by SLDM-based feature enhancement, we show in this paper that the SLDM feature enhancement can in turn benefit from the HMMs ability to model long-term dependencies. We propose an enhancement and recognition architecture, where posterior probabilities are iteratively exchanged among the components. This approach is reminiscent of iterative receiver architectures used in telecommunications.

This paper is organized as follows. In the next section we describe the baseline SLDM proposed in [2]. In section 3 the iterative speech feature enhancement and recognition is considered. Section 4 is concerned with noise estimation. We present experimental results in section 5 and finish with some conclusions in section 6.

2. Baseline

The front-end processing in our system is very similar to the Switching Linear Dynamic Model (SLDM) proposed in [2]. For this reason it is described only briefly in the following.

The feature enhancement takes place in the cepstral domain. The clean speech feature vector with 13 correlated components is denoted by $\mathbf{x}_t$. It's dynamics are modeled with a SLDM according to the state equation

$$\mathbf{x}_t = \mathbf{A}_{s_t}\mathbf{x}_{t-1} + \mathbf{b}_{s_t} + \mathbf{v}_t, \mathbf{v}_t \propto \mathcal{N}(0, \mathbf{C}_{s_t}) \qquad (1)$$

where the model parameters $\mathbf{A}_{s_t}$, $\mathbf{b}_{s_t}$ and $\mathbf{C}_{s_t}$ depend on the discrete state s_t. In [2] a time dependence among the continuous $\mathbf{x}_t$ is assumed, but not among the discrete state variables s_t. This can be expressed either by the graphical model in figure 1 or by the equations

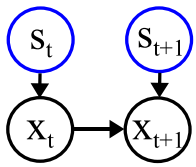

Figure 1: *Graphical model underlying the SLDM*

$$p(\mathbf{x}_t, s_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \mathbf{A}_{s_t}\mathbf{x}_{t-1} + b_{s_t}, C_{s_t})P(s_t)$$
$$p(\mathbf{x}_1^T, s_1^T) = p(\mathbf{x}_1, s_1)\prod_{t=2}^{T} p(\mathbf{x}_t, s_t|\mathbf{x}_{t-1}) \qquad (2)$$

For training standard EM techniques can be used. The Zero Variance Model (ZVM) which was introduced in [7] is applied as observation model. In this approach for each SLDM state s

an SNR variable $\mathbf{r}_s = \mathbf{x}_s - \mathbf{n}$ is iteratively enhanced in order to obtain the updated feature vector $\mathbf{x}_s$. For simplicity the frame index t is omitted here. The clean speech vector $\mathbf{x}_s$, the noise $\mathbf{n}$ and the SNR variable $\mathbf{r}_s$ are assumed to be Gaussians with means μ_s^x, μ^n, μ_s^r respectively and covariance matrices σ_s^x, σ^n, σ_s^r. The moments of $\mathbf{r}_s$ are estimated from the observation vector $\mathbf{y}$ according to the formulas

$$\begin{aligned}(\sigma_s^r)^{-1} &= (\mathbf{F}_s^0 - I)^T(\sigma_s^x)^{-1}(\mathbf{F}_s^0 - I) + (\mathbf{F}_s^0)^T(\sigma^n)^{-1}\mathbf{F}_s^0 \\ \mu_s^r &= \sigma_s^r(\mathbf{F}_s^0 - I)^T(\sigma_s^x)^{-1}(y - \mathbf{f}_s^0 \\ &+ \mathbf{F}_s^0\mathbf{r}_s^0 - \mu_s^x) + \sigma_s^r(\mathbf{F}_s^0)^T(\sigma^n)^{-1}(y - \mathbf{f}_s^0 + \mathbf{F}_s^0\mathbf{r}_s^0 - \mu^n))\end{aligned} \tag{3}$$

where the vector $\mathbf{f}_s^0$ and the matrix $\mathbf{F}_s^0$ represent the first two terms in the Taylor series expansion of $\mathbf{f}(\mathbf{r}) = \mathbf{C}\ln(e^{\mathbf{Dr}} + 1)$ around the state-conditional point $\mathbf{r} = \mathbf{r}_0$:

$$\begin{aligned}\mathbf{f}_s^0 &= \mathbf{C}\ln(e^{\mathbf{D}\mathbf{r}_s^0} + 1) \\ \mathbf{F}_s^0 &= \mathbf{C}\text{diag}(\frac{1}{1 + e^{-\mathbf{D}\mathbf{r}_s^0}})\mathbf{D}\end{aligned} \tag{4}$$

with the discrete cosine transform (DCT) matrix $\mathbf{C}$ and its pseudo-inverse matrix $\mathbf{D} = \mathbf{C}^{-1}$. The parameters of the conditional posterior for a state s are computed to be

$$\begin{aligned}\text{E}[\mathbf{x}|\mathbf{y}, s] &\approx \mathbf{y} - \ln(e^{\mu_s^r} + 1) + \mu_s^r \\ \text{Var}[\mathbf{x}|\mathbf{y}, s] &\approx \sigma_s^r\end{aligned} \tag{5}$$

The posterior estimation of $p(\mathbf{x}_t|\mathbf{y}_1^t)$ under the noisy speech feature sequence $\mathbf{y}_1^t = \mathbf{y}_1, \ldots, \mathbf{y}_t$ with the SLDM is computationally intractable because for M different values of the state variable s_t there are M^t possible state sequences for t frames of speech so that a suitable approximation is required. A very common approximation which is applied in [2] is the generalized pseudo-Bayesian (GPB) algorithm. The idea is to collapse the posterior to M^r Gaussian components for a GPB of order r so that the inference complexity is reduced from M^t to M^r. For each frame of data, three steps are performed in order: collapse, predict and observe. In the collapse step the M^r Gaussian components which represent the distributions $p(\mathbf{x}_{t-1}|\mathbf{y}_1^{t-1}, s_{t-r+1}^{t-1})$ are collapsed to M^{r-1} components. This is achieved by the marginalization

$$p(\mathbf{x}_{t-1}|\mathbf{y}_1^{t-1}, s_{t-r+1}^{t-1}) \approx \sum_{s_{t-r}} p(\mathbf{x}_{t-1}|\mathbf{y}_1^{t-1}, s_{t-r}^{t-1})P(s_{t-r}). \tag{6}$$

In the prediction step the remaining hypotheses are branched out M times, once for each possible state s_t.

$$\begin{aligned}&p(\mathbf{x}_t|\mathbf{y}_1^{t-1}, s_{t-r+1}^t) \\ &= \int p(\mathbf{x}_t - 1|\mathbf{y}_1^{t-1}, s_{t-r+1}^{t-1})p(\mathbf{x}_t|\mathbf{x}_{t-1}, s_t)d\mathbf{x}_{t-1}\end{aligned} \tag{7}$$

Finally the current observation $\mathbf{y}_t$ is incorporated in the observe step. For this purpose the observation model is applied to perform the measurement update $p(\mathbf{x}_t|\mathbf{y}_1^{t-1}, s_{t-r+1}^t) \rightarrow p(\mathbf{x}_t|\mathbf{y}_1^t, s_{t-r+1}^t)$. In [2] the prior distribution $P(s_t)$ for the hidden variables comes from the output of the observe step.

3. Iterative Feature Enhancement and Recognition

It is computationally intractable to optimize the model probability $P(s_t|\mathbf{y}_1^T)$ conditioned on the measurements $\mathbf{y}_1 \ldots \mathbf{y}_T$ of a complete utterance in the approach described in section 2. In contrast, the state probabilities $P(q_t|\mathbf{y}_1^T)$ of the HMM states $q_t, t = 1, \ldots, T$, of the recognizer can be calculated by a forward-backward algorithm. Due to the left-to-right HMM topology with a restricted set of transitions, the HMM is able to model long-term dependencies, i.e. $P(q_t|\mathbf{y}_1^T)$ is influenced by $P(q_\tau|\mathbf{y}_1^T)$ even for large $|t - \tau|$. If a long-span language model is used the "dependency time" is further extended. While the HMM is more suitable to model nonlinear long-term dependencies, LDMs are more appropriate to represent linear short-term correlations, see eq. (2).

In order to exploit the long-term dependencies captured by the HMM in the enhancement stage we calculate the SLDM probabilities $P(s_t|\mathbf{y}_1^T)$ as a function of the HMM probabilities $P(q_t|\mathbf{y}_1^T)$, thus closing an iteration loop. The resulting system for iterative speech feature enhancement and recognition is depicted in figure 2.

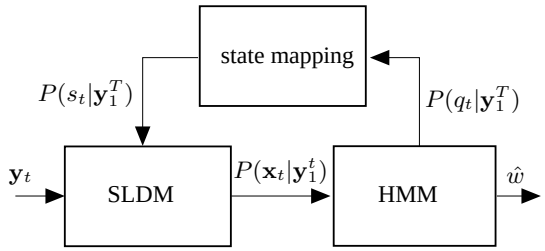

Figure 2: *Iterative speech feature enhancement and recognition*

First, the noisy speech feature vectors are enhanced at the SLDM stage, which delivers the posterior probability $p(\mathbf{x}_t|\mathbf{y}_1^t)$ of the clean feature $\mathbf{x}_t$. In the first iteration there exists no HMM output, so the upper part of figure 2 is left out. The feature enhancement reduces to the one used in [2], where the state probabilities $P(s_t)$ come out of the observe step. The HMM decoder provides the state posterior probabilities $P(q_t|\mathbf{y}_1^T)$, $t = 1, \ldots, T$ by the forward-backward algorithm. Alternatively, a Viterbi decoder can be employed where a single best state sequence is calculated. An approximate posterior may also be computed if the decoder outputs an N-best list.

In order to employ uncertainty decoding the HMM observation probability $p(\mathbf{x}_t|q_t)$ is replaced by [10]

$$\int p(\mathbf{x}_t|q_t)\frac{p(\mathbf{x}_t|\mathbf{y}_1^t)}{p(\mathbf{x}_t)}d\mathbf{x}_t. \tag{8}$$

In case of $\mathbf{y}_t = \mathbf{x}_t$ (absence of noise) the posterior $p(\mathbf{x}_t|\mathbf{y}_1^t)$ becomes a Dirac delta impulse, and (8) reduces to the ordinary observation probability $p(\mathbf{x}_t|q_t)$ (the term $p(\mathbf{x}_t)$ in the denominator can then be dropped as it is a constant). In case of a completely unreliable observations $p(\mathbf{x}_t|\mathbf{y}_1^t)$ tends to the prior $p(\mathbf{x}_t)$, as the observed feature becomes uninformative. Thus, the observation probability tends to unity, which is equivalent to a marginalization of that feature.

Since the posterior $p(\mathbf{x}_t|\mathbf{y}_1^t)$ is modeled as a Gaussian, the integral (8) can be solved analytically. Note that unlike many other formulations of uncertainty decoding the form in (8) allows to employ posteriors which depend on the whole observed feature vector sequence.

The purpose of the state mapping module is to map the HMM state posteriors to SLDM state posteriors. Obviously there is no direct equivalence between HMM and SLDM states, because the HMM states correspond to stationary regions of the speech signal while the SLDM states can be related to regions

with the same dynamics. However, due to the left-to-right topology of the HMM model the HMM states are assigned to certain positions in a word. It is clear that e.g. for HMM states at the beginning of a word SLDM states with dynamic models for rising flanks are more probable, while for HMM states at the end of a word decreasing flanks can be expected. A problem with this approach is the great number of HMM states in realistic speech recognition systems. It is unrealistic to employ an equal number of LDMs due to limited training data and computation time. For this reason we use only a small number of SLDM states and learn the conditional probability $P(s_t|q_t)$ from clean speech training data as follows. First we write

$$\hat{P}(s|q) = \frac{\hat{P}(s,q)}{\hat{P}(q)}. \tag{9}$$

The numerator can be estimated from the posterior probabilities of the states:

$$\hat{P}(s,q) = \sum_t P(s_t = s|q_t = q, \mathbf{X})P(q_t = q|\mathbf{X}). \tag{10}$$

where the summation is over all frames of the data base and $\mathbf{X}$ denotes all feature vectors. The first term can be approximated as follows

$$P(s_t|q_t, \mathbf{X}) \approx P(s_t|x_t, x_{t-1}) \propto p(x_t|x_{t-1}, s_t)P(s_t) \tag{11}$$

The term $P(q_t = q|\mathbf{X})$ in (10) is given by the Baum-Welch HMM training, or may be approximated by a hard decision

$$P(q_t = q|\mathbf{X}) \approx \delta(q_t) = \begin{cases} 1, \text{ if state } q_t \text{ is active in frame } t \\ 0, \text{ else} \end{cases} \tag{12}$$

In a similar way we obtain the denominator of (9):

$$\hat{P}(q) = \sum_t P(q_t = q|\mathbf{X}). \tag{13}$$

The posterior of the SLDM state to be used in the second iteration of the speech feature enhancement is now obtained as follows:

$$P(s_t|\mathbf{y}_1^T) = \sum_{q_t} p(s_t|q_t)p(q_t|\mathbf{y}_1^T) \tag{14}$$

where $p(q_t|\mathbf{y}_1^T)$ can be obtained from a true forward-backward recognizer. For a Viterbi-based HMM decoder, (14) simplifies to

$$P(s_t|\mathbf{y}_1^T) = P(s_t|\hat{q}_t) \tag{15}$$

where $\hat{q}_t$ is the optimal HMM state at time t found by the Viterbi decoder. To give alternative state hypotheses a noticeable contribution to the state mapping we used a heuristic, which is explained for the example of an HMM decoder, which outputs an N-best list. Let $(\mathbf{q}_1^T)^{(n)}$ denote the best state sequence of the n-th best sentence. We then replaced (14) by

$$P(s_t|\mathbf{y}_1^T) \propto \sum_{n=1}^{N} P(s_t|q_t)(p(\mathbf{y}_1^T|(\mathbf{q}_1^T)^{(n)}))^{1/T} \tag{16}$$

Now the loop is closed and the second iteration of feature enhancement and subsequent recognition can start. More than two iterations, however, did not prove to be effective.

4. Feedback of HMM State Posteriors for Noise Estimation

The HMM state probabilities $P(q_t|\mathbf{y}_1^T)$ which are used in section 3 for the estimation of the SLDM probabilities $P(s_t|\mathbf{y}_1^T)$ can also be applied for the calculation of a soft VAD variable

$$P_{sil} = \sum_{q_t \epsilon Q_{sil}} P(q_t|\mathbf{y}_1^T), \tag{17}$$

where Q_{sil} denotes the HMM states which correspond to silence. The noise estimation is initialized with the first and last 10 frames of an utterance. Then, for each frame the noise estimate is updated by

$$\mu_{n_t} = (1 - \alpha P_{sil})\mu_{n_{t-1}} + \alpha P_{sil}\tilde{\mu}_{n_t}, \tag{18}$$

where α is a weighting coefficient which determines the influence of the current noise estimate $\tilde{\mu}_{n_t}$. It has been selected in informal experiments to be 0.05. Note, that no update of the noise estimate $\mu_{n_{t-1}}$ is carried out if $P_{sil} = 0$. The instantaneous noise estimate $\tilde{\mu}_{n_t}$ is computed by $\tilde{\mu}_{n_t}^{(sil)} = \mathbf{y}_t$ in speech pauses, and by $\tilde{\mu}_{n_t}^{(speech)} = \mathbf{y}_t - \ln(1 + e^{-\mu_{r_t}})$ during speech activity [7]. The latter can be estimated after the measurement update with the SLDM, eq. (3). With the soft silence indicator (17) we thus obtain

$$\begin{aligned} \tilde{\mu}_{n_t} &= P_{sil}\tilde{\mu}_{n_t}^{(sil)} + (1 - P_{sil})\tilde{\mu}_{s_t}^{(speech)} \\ &= \mathbf{y}_t - (1 - P_{sil}) \ln(1 + e^{-\mu_{r_t}}) \end{aligned} \tag{19}$$

which is the instantaneous noise estimate employed in eq. (18).

5. Experimental results

The experiments were performed on test set A of the AURORA2 database [11] with clean speech training data. To be able to compare our results with [7], from which we adopted the observation model, we modified the ETSI standard frontend extraction [12] in the same manner as there by replacing the energy feature with c_0 and using the magnitude squared power spectral density rather than the spectral magnitude as the input of the Mel-frequency filter-bank. In the first set of experiments the noise was estimated from the first and last 10 frames of each sentence. The overall recognition accuracy was averaged over all noise conditions at SNR levels between 0dB and 20dB. The speech recognition with the described standard frontend (SFE) yields an overall recognition accuracy of 60.37% (table 1). The

Table 1: Word accuracy on testset A of the AURORA2 database at different input SNR levels

	SFE	SLDM	SLDM-FB1	SLDM-FB2	SLDM-FB$_{opt}$
Clean	99.67%	99.52%	99.69%	99.6%	99.61%
20dB	99.29%	98.01%	98.04%	98.11%	98.63%
15dB	87.83%	95.6%	95.22%	95.49%	96.99%
10dB	67%	89.28%	88.66%	89.3%	92.44%
5dB	36.44%	73.74%	74.59%	75.79%	82.73%
0dB	14.32%	42.7%	47.37%	48.37%	59.42%
-5dB	6.58%	15.09%	19.29%	19.3%	29.5%
Avg.	60.37%	79.87%	80.77%	81.41%	86.04%

baseline SLDM which is described in section 2 leads to an accuracy of 79.87%, if $M = 16$ LDMs are used. An overall accuracy of 80.77% is achieved with the feedback of the HMM state probabilities (SLDM-FB1). The improvement is larger for low SNR levels, while the accuracy at higher SNR levels is hardly increased. By coupling back more than one hypothesis according to eq. (16) a performance of 81.41% could be achieved (SLDM-FB2). A multi-hypotheses approach where the state probabilities $P(q_t)$ are calculated with the forward-backward algorithm, eq. (14), seems to be more adequate. However so far in our experiments the probabilities $P(q_t)$ were almost one for one state and zero for all the other states at time instance t.

In order to obtain an upper bound on the performance achievable with our state mapping approach, the HMM states were calculated from clean speech data and afterwards mapped to SLDM states (SLDM-FB$_{opt}$). Table 1 shows that with the same noisy input data but with optimal state mapping resulting from the correct HMM state sequence a significant improvement of the recognition accuracy can be obtained.

In table 2 the results are given for the different noise types: the average recognition accuracies for the SFE without feature enhancement, the baseline SLDM and the proposed method SLDM-FB2 at different noise conditions. SLDM-FB2 yields improvements compared to SLDM at three noise conditions. At babble noise the recognition rate is slightly decreased. Further

Table 2: Word accuracy on testset A of the AURORA 2 database at different noise conditions

	Subway	Babble	Car	Exhib.	Avg.
SFE	68.06%	45.87%	58.34%	64.76%	60.37%
SLDM	80.19%	72.56%	84.28%	82.43%	79.87%
SLDM-FB2	82.25%	72.15%	87.21%	84.03%	81.41%
SLDM-FBN	82%	76.27%	85.3%	84.29%	81.96%

the results for the noise estimation method (SLDM-FBN) described in section 4 are depicted in table 2. An overall performance of 81.96% is achieved compared to 81.41% with noise estimation from the first and last 10 frames (SLDM-FB1). However the results strongly depend on the background noise. While for the nonstationary babble noise an improvement of 4.12% is obtained, the recognition rates for the car noise are even decreased by 1.91% and the results for the subway and exhibition noise are only slightly changed.

6. Conclusions

In this paper speech feature enhancement with an SLDM and speech recognition with an HMM decoder were combined in a new way. By employing a state mapping to feedback posterior probabilities from the recognizer to the enhancement stage, feature enhancement can benefit from the ability of HMMs to model long-term dependencies in the data. With this approach significant word accuracy gains were achieved at low signal-to-noise power ratios. Further, the HMM state posterior probabilities were shown to be useful for noise estimation under nonstationary background conditions.

7. Acknowledgements

The research is (partly) supported by the DFG Research Training Group GK-693 of the Paderborn Institute for Scientific Computation (PaSCo). We want to thank J. Droppo for valuable comments at an early stage of the work. We also would like to thank Valentin Ion for many useful discussions.

8. References

[1] Moreno, P., Raj, B. and Stern, R., "A Vector Taylor Series Approach for Environment-Independent Speech Recognition", in *Proc. ICASSP*, Atlanta, 1996.

[2] Droppo, J. and Acero, A.: "Noise robust speech recognition with a switching linear dynamic model", In *Proc. ICASSP*, Montreal, Canada, 953-956, 2004.

[3] Kim, N. S., Lim, W., Stern, R. M.: "Feature compensation based on switching linear dynamic model", *IEEE Signal Processing Letters*, vol. 12, no. 6, June 2005.

[4] Deng, J., Bouchard, M., Yeap, T. H.: "Speech Feature Estimation Under The Presence Of Noise With A Switching Linear Dynamic Model", in *Proc. ICASSP*, Toulouse, France, pp. 497-500, vol. 1, May 2006.

[5] Faubel, F. and Wölfel, M.: "Coupling Particle Filters with Automatic Speech Recognition for Speech Feature Enhancement", In *Proc. Interspeech*, Pittsburgh, Pennsylvania, 2006.

[6] Rosti, A-V. I. and Gales, M. J. F.: "Switching linear dynamical systems for speech recognition". *Tech. Rep. CUED/F-. INFENG/TR.461*, Cambridge University, 2003.

[7] Droppo, J., Deng, L. and Acero, A.: "A comparison of three non-linear observation models for noisy speech features", In *Proc. Eurospeech*, Geneva, Switzerland, 681-684, 2003.

[8] Pavlvic, V., Rehg, J. M. and MacCormick, J.: "Learning Switching Linear Models of Human Motion", In *Proc. Neural Information Processing Systems*, 981-987, 2000.

[9] Saul, L., Jordan, M.I.: "Exploiting tractable substructures in intractable networks", In Touretzky, D. S., Mozer, M. C. and Hasselmo, M. E., editors: *Advances in Neural Information Processing Systems*, Vol. 8, 486-492. The MIT Press, 1996.

[10] Ion, V., Haeb-Umbach, R.: "Improved Source Modeling and Predictive Classification for Channel Robust Speech Recognition", in *Proc. Interspeech*, Pittsburgh, 2006.

[11] Pearce, D., Hirsch, H.-G.: "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions", In *Proc. ISCA ITRW Workshop ASR2000*, Paris, France, 2000.

[12] ETSI, "ES 201 108 V1.1.2, Standard front-end feature extraction algorithm", Techn. rep., April 2000.