Discriminant Sub-Space Projection of Spectro-Temporal Speech Features based on Maximizing Mutual Information

Martin Heckmann, Claudius Gläser

Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany
martin.heckmann@honda-ri.de, claudius.glaeser@honda-ri.de

Abstract

We previously developed noise robust Hierarchical Spectro-Temporal (HIST) speech features. The learning of the features was performed in an unsupervised way with unlabeled speech data. In a final stage we deployed Principal Component Analysis (PCA) to reduce the feature dimensions and to diagonalize them. In this paper we investigate if a discriminant projection can further increase the performance. We maximize the mutual information between the features and the phoneme categories using a procedure known as Maximizing Renyi's Mutual Information (MRMI) and also compare it to Linear Discriminant Analysis (LDA). Based on recognition tests in clean and in noise, i. e. in matching and mismatching conditions, we show that the discriminant projections increases recognition scores compared to PCA in matching conditions. However, this improvement does not transfer to the mismatching, i. e. noisy, conditions. We discuss measures to alleviate this problem. Overall MRMI performs better than LDA.

Index Terms: Spectro-temporal, discriminant, mutual information, robust speech recognition, auditory

1. Introduction

Most common speech features as Mel Cepstral Coefficients (MFCCs) and RelAtive SpecTrAl Perceptual Linear Predictive (RASTA-PLP) features use only spectral information. However, from measurements in the mammalian auditory cortex it is known that the mammalian brain jointly uses spectral and temporal information [1]. This is potentially better suited to capture the information conveyed by formant trajectories. Different spectro-temporal feature sets have been introduced to model this [2, 3, 4, 5, 6].

We previously presented Hierarchical Spectro-Temporal (HIST) features [7, 8]. They consist of two layers, the first capturing local spectro-temporal variations and the second integrating them into larger receptive fields (compare Fig. 1). This layout was inspired by a recently proposed system for visual object recognition [9]. At both layers the receptive fields are learned in a data-driven unsupervised way. On the first layer we apply Independent Component Analysis (ICA) and in the second layer Non-Negative Sparse Coding (NNSC). In our original setup we applied Principal Component Analysis (PCA) to orthogonalize the features and reduce their dimensionality followed by a Hidden Markov Model (HMM) for the recognition.

The PCA projects the features into an orthogonal subspace

Figure 1: Overview of the feature extraction framework

such that the mean squared error is minimized. One weakness of PCA in the context of classification is that no information on the class membership of the feature vectors can be taken into account. In contrast to this, Linear Discriminant Analysis (LDA) tries to find a subspace where the L classes are best separated. However, LDA yields only a feature vector of dimensionality $L-1$ and makes the assumption that the features conditioned on the class are normally distributed with equal variances for all classes. A more general approach is to maximize the mutual information between the class labels and the transformed features. One way of obtaining this is using an approach called Maximizing Renyi's Mutual Information (MRMI) [10]. It overcomes the assumptions on the input variances and also allows for projections with more dimensions than classes.

LDA and an approach maximizing the mutual information have been applied to speech recognition before [11, 12]. These approaches have in common that they are based on MFCC features and start from a rather low-dimensional feature vector. In contrast to this we use 150-dimensional spectro-temporal features. The previous experiments reported a moderate improvement due to a discriminative feature projection. However, they only considered situations where the conditions during the learning of the features and there application were similar. In the following we will investigate MRMI, LDA, and PCA in matching and mismatching conditions. To model matching and mismatching conditions we add different types of noise to the speech data either only in the test or in the test and training set.

The rest of the paper is organized as follows. In Section 2 we will briefly describe our Hierarchical Spectro-Temporal (HIST) feature extraction framework. This is followed by a description of the MRMI method in Section 3. The experimental conditions and recognition results will be presented in Section 4. A conclusion and a discussion in Section 5 will close the paper.

2. Hierarchical Spectro Temporal Features

The main building blocks of our hierarchical feature extraction framework are a preprocessing to enhance the formant structure in the spectrograms, a calculation of local and combination features, and a projection of the features into a lower-dimensional space (compare Fig. 1).

2.1. Preprocessing

We apply a Gammatone filter bank to transform the speech signal sampled at 16 kHz into the frequency domain. The filter bank has 128 channels ranging from 80 Hz to 8 kHz and follows the implementation of [13]. From this we obtain spectrograms by rectification and low-pass filtering of the filter bank response. The sampling rate of the spectrograms is then reduced

to 400 Hz.

An enhancement of the formant structure in the signal is obtained by a pre-emphasis of $+6\,\mathrm{dB/oct.}$ and a subsequent filtering along the frequency axis with a Mexican Hat filter. The last step removes the harmonic structure of the spectrograms and forms peaks at the formant locations (see [14] for details).

2.2. First stage: Extraction of local features

In the first layer $Q^{(1)}$ of our hierarchical feature extraction framework local features are extracted via a 2D filtering with a set of $l = 1 \dots N_1$ receptive fields $\boldsymbol{w}_l^{(1)}$, taking the absolute value of the response:

$$q_l^{(1)}(t,f) = \left| \left(\boldsymbol{S} * \boldsymbol{w}_l^{(1)} \right)(t,f) \right|, \tag{1}$$

where the responses $\boldsymbol{q}_l^{(1)}$ of each neuron has the same size as the input spectrogram $\boldsymbol{S}$. The filtering, i. e. convolution, operation is depicted by $*$.

These $N_1 = 8$ receptive fields are learned using Independent Component Analysis (ICA) on 3500 randomly selected local 16×16 patches of the enhanced spectrograms taken from the training set.

For a given point (t, f) in the spectrogram, the activity $q_l^{(1)}(t,f)$ of the l-th neuron reveals how close a local patch of $\boldsymbol{S}$ centered in (t, f) is to the pattern l. For each local patch only the highest correlated patterns are of interest. Therefore, we perform a Winner-Take-Most (WTM) competition which inhibites the response of the less active neurons at the position (t, f) resulting in the activations $r_l^{(1)}(t,f)$ [8]. Furthermore, a nonlinear transformation with a Heaviside step function is applied on all the $\boldsymbol{r}_l^{(1)}$. After smoothing with a 2D Gaussian filter the resolution of the spectrograms $\boldsymbol{s}_l^{(1)}$ is reduced by a factor of four in both frequency and time dimension yielding features $\boldsymbol{c}_l^{(1)}$ with 32 frequency channels and a sampling rate of 100 Hz [8].

2.3. Second stage: Extraction of combination features

Each of the $k = 1 \dots N_2$ combination patterns on the second layer $Q^{(2)}$ of our hierarchy is composed of N_1 receptive fields $\boldsymbol{w}_{l,k}^{(2)}$, i. e. one for each of the neurons in the previous layer. The coefficients of these receptive fields are non-negative and span all frequency channels. Similarly to (1) the activity $q_k^{(2)}(t)$ of the k-th neuron at time t is given by:

$$q_k^{(2)}(t) = \sum_{l=1}^{N_1} \left(\boldsymbol{c}_l^{(1)} * \boldsymbol{w}_{l,k}^{(2)} \right)(t,f). \tag{2}$$

As the combination patterns span the whole frequency range the response of the neurons does not depend on f anymore. This means that, by computing the convolution, the patterns $\boldsymbol{w}_{l,k}^{(2)}$ are only shifted in the time direction. Note that the absolute value is not required in (2) as both the $\boldsymbol{c}_l^{(1)}$ and the $\boldsymbol{w}_{l,k}^{(2)}$ are non-negative.

The combination patterns were also learned in an unsupervised manner using Non-Negative Sparse Coding (NNSC) [15]. Thereby, we cut out patches $\boldsymbol{P}$ of length $\Delta = 40\,\mathrm{ms}$ of the first layer activations $\boldsymbol{c}_l^{(1)}$. From these patches we learn $N_2 = 50$ combination features by minimizing the following cost function [9]:

$$E = \sum_i \|\boldsymbol{P}_i - \sum_{k=1}^{N_2} \alpha_{k,i} \boldsymbol{w}_k^{(2)}\|^2, \tag{3}$$

where $\boldsymbol{P}_i$ is a tensor representing the N_1 layers of the i-th patch, the $\boldsymbol{w}_k^{(2)}$ are N_2 non-negative tensors each of them containing the N_1 receptive fields $\boldsymbol{w}_{l,k}^{(2)}$, the $\alpha_{k,i}$ are nonnegative reconstruction factors, and β is a parameter allowing to control the sparsity of the learned features (see [7, 8] for details).

Overall, this yields $N_2 = 50$ features $q_k^{(2)}(t)$ at a sampling rate of 100 Hz. Delta and double-delta features are computed using a 9th order FIR lowpass and bandpass filter, respectively. When combining the features $q_k^{(2)}(t)$ with their deltas we obtain an $N = 150$ dimensional vector $\boldsymbol{x}$.

3. Subspace Projection via MRMI

As described above we previously applied PCA to project the N-dimensional feature vector $\boldsymbol{x}$ from our hierarchical feature extraction to an M-dimensional subspace. The PCA takes only the variance of the input dimensions into account and hence feature dimensions with low variation but possible high discriminative power will be discarded. In contrast to this, discriminative feature projections span a subspace in which the discriminative power of the dimensions is maximized. In the approach we investigate, namely Maximizing Renyi's Mutual Information (MRMI), the mutual information

$$I(\boldsymbol{Y}; C) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|C) \tag{4}$$

between the features $\boldsymbol{y}$ in the subspace $\mathbb{R}^M$ and the corresponding class labels C is maximized [10]. For a linear feature extraction of form $\boldsymbol{y} = \boldsymbol{R} \cdot \boldsymbol{x}$ one consequently searches for a feature extraction matrix $\boldsymbol{R} \in \mathbb{R}^{M \times N}$ which maximizes (4).

In [10] it was shown that the maximization of the mutual information can be simplified when Shannon's definition of entropy is replaced by Renyi's quadratic entropy $H_2(\boldsymbol{Y})$. Thereby, $H_2(\boldsymbol{Y})$ can be efficiently calculated by relying on Parzen window density estimation using K randomly ordered feature samples $\boldsymbol{y}(k)$:

$$\begin{aligned} I(\boldsymbol{Y}; C) &\cong H_2(\boldsymbol{Y}) - H_2(\boldsymbol{Y}|C) \\ H_2(\boldsymbol{Y}) &\cong -\log \frac{1}{K} \sum_{k=1}^{K} G(\boldsymbol{y}(k) - \boldsymbol{y}(k-1), 2\sigma^2 \boldsymbol{I}) \end{aligned} \tag{5}$$

Here, $G(\boldsymbol{z}, \sigma^2 \boldsymbol{I}) = \exp(-\frac{1}{2} \frac{\boldsymbol{z}^T \boldsymbol{z})}{2\sigma^2})$ is a Gaussian kernel evaluated at $\boldsymbol{z}$, where the kernel is centered at the origin and has a diagonal isotropic covariance matrix.

Consider a training set composed of samples $\boldsymbol{x}_j(k)$ as representatives of class j where $\boldsymbol{y}_j(k) = \boldsymbol{R} \cdot \boldsymbol{x}_j(k)$. Furthermore, let K_j denote the number of samples belonging to class j, K_c the number of classes, and $K_T = \sum_{j=1}^{K_c} K_j$ the length of the overall training set. Then the information-theoretic criterion can be formulated as [10]

$$\begin{aligned} I(\boldsymbol{Y}; C) = &-\log \frac{1}{K_T} \sum_{k=1}^{K_T} G(\boldsymbol{y}(k) - \boldsymbol{y}(k-1), 2\sigma^2) \\ &+ \sum_{j=1}^{K_c} \left(\frac{K_j}{K_T} \log \frac{1}{K_j} \sum_{k=1}^{K_j} G(\boldsymbol{y}_j(k) - \boldsymbol{y}_j(k-1), 2\sigma^2) \right). \end{aligned} \tag{6}$$

Consequently, $\boldsymbol{R}$ can be learned via stochastic gradient ascent on $I(\boldsymbol{Y}; C)$.

To decorrelate the resulting feature dimensions we finally apply PCA on the feature space learned via MRMI. Without loss of generality we assume the features $\boldsymbol{X}$ to be white with zero mean and unit variance. Then the covariance of the class-discriminative feature vectors $\boldsymbol{y}$ is $\mathrm{cov}(\boldsymbol{Y}, \boldsymbol{Y}) = \boldsymbol{R} \cdot \boldsymbol{R}^T$. Let

	white	factory	babble	car
RASTA-PLP	43.1	41.0	35.0	19.5
HIST-PCA$_{20}$	51.3	56.2	82.7	19.4
HIST-LDA$_{20}$	53.0	47.6	68.9	36.5
HIST-MRMI$_{20}$	38.1	38.8	86.8	16.6
HIST-PCA$_{39}$	50.0	58.5	80.5	21.0
HIST-MRMI$_{39}$	34.9	42.0	76.5	18.2
RASTA-PLP+HIST-PCA$_{20}$	32.6	34.8	36.5	14.4
RASTA-PLP+HIST-LDA$_{20}$	34.3	34.7	62.0	19.7
RASTA-PLP+HIST-MRMI$_{20}$	24.0	28.2	61.2	12.0
RASTA-PLP+HIST-PCA$_{39}$	27.9	31.5	58.7	11.0
RASTA-PLP+HIST-MRMI$_{39}$	23.4	32.9	71.7	12.2

Table 1: Word error rates (in %) averaged for an individual type of additional noise over SNR values ranging from $-5\,\mathrm{dB}\ldots$ inf.

$\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_M]$ be the eigenvectors of $\boldsymbol{R} \cdot \boldsymbol{R}^T$. Consequently, the decorrelated class-discriminative feature space can be obtained by

$$\boldsymbol{y} = \boldsymbol{\Psi}^T \cdot \boldsymbol{R} \cdot \boldsymbol{x}. \tag{7}$$

4. Results

We compare the discriminative feature projection MRMI to LDA and PCA on a task very similar to Aurora-2 [16]. To TIDigits [17], a database for speaker independent continuous digit recognition, we added White, Babble, Factory, and Car noise from the Noisex database [18] at Signal to Noise Ratios (SNRs) ranging from $-5\,\mathrm{dB}\ldots$ inf, i. e. we also kept the clean signal. However, we did keep the original sampling rate of 16 kHz of TIDigits but did not add channel distortions when mixing the signals using FaNT [19]. As in Aurora-2 we used HTK [20] to train whole word HMMs containing 16 states without skip transitions and a mixture of 3 Gaussians with a diagonal covariance matrix per state.

Training of the receptive fields of the $Q^{(1)}$ and $Q^{(2)}$ layer of our feature hierarchy was performed on TIDigits. If not stated otherwise the features as well as the HMMs are trained on clean speech data. For the learning of the MRMI projection we used the Timit database as this step required phonetic labels [21]. We identified 21 phonemes necessary to cover the digit sequences in TIDigits and randomly extracted for each of these phonemes 3000 segments of length 10 ms from Timit. Silences and pauses were not included as phonemic categories. The variance σ^2 for the Parzen approximation in (6) was set to 10 and the learning was terminated after 10000 iterations at a learning rate of 0.1. The LDA and PCA matrix was calculated on the identical data. The output dimensionality of the LDA is limited to $L-1$ with $L = 21$ the number of phoneme classes in our database. Therefore we calculated all transformations for $M = 20$, i. e. $L-1$, and MRMI and PCA also for $M = 39$, the dimensionality we used in our previous experiments with PCA [8].

As benchmark we also extracted RASTA-PLP features [22] with 45 dimensions. We have seen in previous experiments that a combination of the HIST and RASTA-PLP features is especially beneficial [7, 8]. Therefore, we chiefly investigated a combination of RASTA-PLP and HIST features where we used projections based on PCA, LDA, as well as MRMI for the HIST features. The combination was obtained via feature concatenation, i. e. resulting in 65 or 84 dimensional vectors.

In Table 1 the Word Error Rates (WERs) for the different noise types averaged over all SNR values are given. The subscripts at PCA, LDA, and MRMI indicate the dimensionality of the output space. Similar to our previous results using only PCA [8] the HIST features, with either projection, perform worse than the RASTA-PLP features in most cases. However, the combination of HIST features and RASTA-PLP features is superior to RASTA-PLP features alone in almost all cases. The exception are features based on the LDA projection with car noise and Babble noise for all feature projections. We have seen before that Babble noise is problematic and could identify a high sensitivity of the HIST features to speech and a subsequent high probability for word insertions in Babble noise [8].

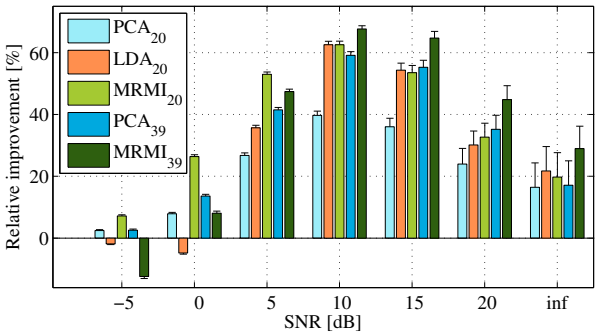

Figure 2: Relative improvements of the combination of RASTA-PLP and HIST features with different projections compared to RASTA-PLP features when factory noise was added to the test set. The bars indicate the 95% confidence interval [23].

Fig. 2 allows to better assess the impact of the different projections. Here the relative improvements compared to RASTA-PLP features when factory noise was added are shown. For low and medium additional noise, i. e. when training and testing conditions match, MRMI and LDA do in fact improve the performance (when compared to features of the same dimensionality). When the noise level is increasing the situation changes. LDA is performing poor with additional noise. Also MRMI$_{39}$ is not performing well at high noise levels. However, MRMI$_{20}$ remains superior to PCA also for high noise levels. When the discriminant projections are learned on clean data there seems to be a tendency of selecting dimensions which achieve good discrimination in clean conditions, but with a high susceptibility to noise (as can be seen by comparing the performance of MRMI$_{39}$ and PCA$_{39}$). This is mitigated by reducing the dimensionality (as can be seen by comparing the performance of MRMI$_{39}$ and MRMI$_{20}$). Which on the other hand comes at the price of reduced performance in clean and low noise conditions.

To further investigate this we also performed tests with noisy data. For doing so we added all four types of noise at SNR levels of 10 and 20 dB also to the training set and performed the tests as before. We want to refer to this configuration as mixed training. If we compare in Fig. 3 the performance of MRMI$_{39}$

	white	factory	babble	car
RASTA-PLP	28.0	27.3	23.3	6.9
HIST-PCA$_{39}$	25.7	29.7	43.4	8.9
HIST-MRMI$_{39}$	23.0	25.9	40.0	7.8
RASTA-PLP+HIST-PCA$_{39}$	24.6	25.1	23.3	5.3
RASTA-PLP+HIST-MRMI$_{39}$	23.3	23.6	22.0	5.0

Table 2: Word error rates averaged for an individual type of additional noise over SNR values ranging from $-5\,\mathrm{dB}\ldots$ inf. Training was performed on the mixed set, i. e. with all four types of noise added to the training set.

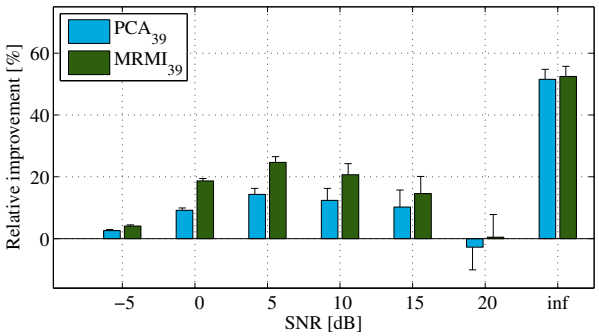

Figure 3: Relative improvements of the combination of RASTA-PLP and HIST features with different projections compared to RASTA-PLP features when training was performed on the mixed set (with all four types of noise added) and factory noise was added to the test set. The bars indicate the 95% confidence interval [23]

and PCA_{39} we can see that $MRMI_{39}$ now outperforms PCA_{39} also for low noise conditions. From Table 2 one can see that MRMI is in this mixed training case superior to PCA on average for all noise types. Please note that the combination of RASTA-PLP and HIST features now also outperforms RASTA-PLP alone for Babble noise. The HMMs were able to adapt to the high speech sensitivity of the HIST features as Babble noise was also present in the training data [8].

5. Conclusion

We compared the discriminant feature projections LDA and MRMI to PCA in the framework of our HIST features. Additionally, we investigated their generalization capabilities to mismatches between training and test conditions. We saw that $MRMI_{39}$ clearly outperforms PCA_{39} in matching conditions, i. e. at low noise levels when training was performed on clean, but yields inferior performance in non-matching, i. e. noisy, conditions. From the results we see that the discriminant projections, i. e. MRMI, select features which increase recognition performance and hence discriminative power in matching conditions. We hypothesize that with increasing dimensionality of the output feature space dimensions are selected which decrease discriminative power in non-matching conditions, i. e. training on clean and testing in noise. One solution seems to be to select a lower-dimensional output space (as the comparison between $MRMI_{39}$ and $MRMI_{20}$ shows). However, this yields reduced performance in matching conditions. An alternative is to perform the learning of the feature projection also on noisy data and thereby reestablish matching conditions when testing is also performed in noisy conditions. In this case MRMI is able to select the features which increase discriminative power in the noisy condition. The performance in clean is then comparable to that of the PCA but not superior. LDA, which we restricted to 20 dimensions as we were using 21 phoneme classes, performed worse than MRMI. Especially, LDA was much more susceptible to a mismatch between training and test condition.

6. References

[1] S. Shamma, "On the role of space and time in auditory processing," *Trends in Cognitive Sciences*, vol. 5, no. 8, pp. 340–348, 2001.

[2] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *Proc. Int. Conf. on Spoken Language Proc. (ICSLP)*, Denver, CO, 2002, ISCA.

[3] B.T. Meyer and B. Kollmeier, "Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities," in *Proc. INTERSPEECH*, Brighton, UK, 2009.

[4] T. Ezzat and T. Poggio, "Discriminative word-spotting using ordered spectro-temporal patch features," in *Proc. SAPA*, Brisbane, 2008.

[5] S.Y. Zhao, S. Ravuri, and N. Morgan, "Multi-Stream to Many-Stream: Using Spectro-Temporal Features for ASR," in *Proc. INTERSPEECH*, Brighton, UK, 2009.

[6] F. Valente and H. Hermansky, "Discriminant linear processing of time-frequency plane," in *Proc. INTERSPEECH*, Pittsburgh, PA, 2006.

[7] X. Domont, M. Heckmann, F. Joublin, and C. Goerick, "Hierarchical sectro-temporal features for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Proc. (ICASSP)*, Las Vegas, NV, 2008, pp. 4417–4420.

[8] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Communication*, vol. 53, no. 5, 2011.

[9] H. Wersing and E. Körner, "Learning Optimized Features for Hierarchical Models of Invariant Object Recognition," *Neural Computation*, vol. 15, no. 7, pp. 1559–1588, 2003.

[10] K.E. Hild II, D. Erdogmus, K. Torkkola, and J.C. Principe, "Feature Extraction Using Information-Theoretic Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1385–1392, 2006.

[11] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Proc. (ICASSP)*, San Fransisco, CA, 1992.

[12] K. Demuynck, J. Duchateau, and D.V. Compernolle, "Optimal feature sub-space selection based on discriminant analysis," in *Proc. 6th Europ. Conf. Speech Communication Techn.*, 1999, pp. 1311–1314.

[13] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Tech. Rep., Apple Computer Co., 1993, Technical report #35.

[14] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and bayesian estimation for robust formant tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 224–236, 2010.

[15] P.O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[16] D. Pearce and H.G. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in *Proc. Int. Conf. on Spoken Lang. Proc.(ICSLP)*, Bejing, China, 2000, ISCA.

[17] R. Leonard, T.I. Incorporated, and T. Dallas, "A database for speaker-independent digit recognition," in *Int. Conf. Acoustics, Speech, and Signal Proc. (ICASSP)*, San Diego, CA, 1984, vol. 9, IEEE.

[18] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[19] G. Hirsch, "FaNT - Filtering and Noise Adding Tool," Tech. Rep., Niederrhein University of Applied Sciences, Krefeld, Germany, 2005.

[20] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University, Cambridge, United Kingdom, 1995.

[21] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM*, Philadelphia, 1993.

[22] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans Speech and Audio Proc.*, vol. 2, no. 4, pp. 578–589, 1994.

[23] J.M. Vilar, "Efficient computation of confidence intervals for word error rates," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, Las Vegas, NV, 2008, pp. 5101–5104, IEEE.