FEATURE SELECTION FOR IMPROVED BANDWIDTH EXTENSION OF SPEECH SIGNALS

Peter Jax and Peter Vary

Institute of Communication Systems and Data Processing (IND)
Aachen University (RWTH), Templergraben 55, 52056 Aachen, Germany
{jax,vary}@ind.rwth-aachen.de

ABSTRACT

The aim of artificial bandwidth extension (BWE) is to convert speech signals with "standard telephone" quality (frequencies up to 3.4 kHz) into 7 kHz wideband speech. The principal key to high quality BWE is the estimation of the spectral envelope of the wideband speech. In general, this estimation of the wideband spectral envelope is based on a number of features that are extracted from the narrowband input speech signal.

In this paper we investigate potential features and evaluate their suitability for the BWE application. The quality of each feature is quantified in terms of the statistical measures of mutual information and separability. It turns out that the best BWE results are obtained by using a large feature "super-vector" (→ high mutual information) which is subsequently reduced in dimension by a linear discriminant analysis (→ large separability). This solution also helps to reduce the computational complexity of the estimation of the wideband spectral envelope.

1. INTRODUCTION

In current public telephone systems the bandwidth of the transmitted speech is limited due to constraints of the old analogue telephone system to a frequency range of up to about 3.4 kHz. This bandwidth limitation causes the characteristic sound of "telephone speech". Establishing true wideband speech communication requires a modification of the transmission link by enhanced speech codecs. An alternative approach towards a higher (audio) bandwidth (typically up to 7 kHz, sampling rate 16 kHz) is the artificial bandwidth extension: Missing low and high frequency components of the speech signal are recovered at the receiving end of the transmission link utilizing only the bandlimited speech [1, 2, 3].

The vast majority of the adaptive BWE algorithms published to date are based on the well-known linear source-filter model of the speech production process: It is assumed that the human vocal tract can be modeled by an auto-regressive filter $1/A(z)$, which is excited by a spectrally flat excitation signal $u(k)$. Accordingly, the bandwidth extension of the speech signal is commonly performed separately for the spectral envelope and the excitation of the speech [1, 2] (cf. Fig. 1). The spectral envelope is particularly important for the subjective quality of the extended speech, e.g. [2].

In this paper the focus shall be on the feature extraction block which is preceding the step of estimating the wideband spectral envelope. In general, feature extraction and estimation of the wideband spectral envelope are performed on a frame-by-frame basis with frame lengths of about 10–30 ms. The feature extraction reduces the dimensionality of each frame of the narrowband signal $s_{\mathrm{nb}}(k)$ such that the subsequent estimation of the spectral envelope is feasible and computationally efficient. The result is the feature vector $\mathbf{x} = [x_1, \dots x_b]^{\mathrm{T}}$ with the dimension $b = \dim \mathbf{x}$. Usually, representations of the spectral envelope of the narrowband signal $s_{\mathrm{nb}}(k)$ are used as features, e.g. LPC or LSF vectors or cepstral coefficients. In some contributions additional features such as voicing criteria are taken into account.

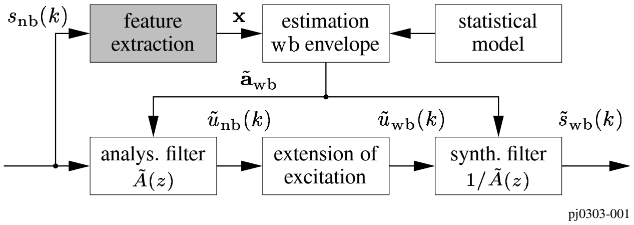
$s_{\mathrm{nb}}(k)$
feature extraction
$\mathbf{x}$
estimation wb envelope
statistical model
$\tilde{\mathbf{a}}_{\mathrm{wb}}$
$\tilde{u}_{\mathrm{nb}}(k)$
$\tilde{u}_{\mathrm{wb}}(k)$
$\tilde{s}_{\mathrm{wb}}(k)$
analys. filter $\tilde{A}(z)$
extension of excitation
synth. filter $1/\tilde{A}(z)$
pj0303-001

Fig. 1. Signal-flow of an exemplary BWE algorithm (from [4]). This contribution focuses on the feature extraction block that is shaded in gray. The subscripts nb and wb denote narrowband and wideband signals, respectively. Note that the sampling rate is 16 kHz for all signals in the diagram, i.e. it is assumed that the input signal $s_{\mathrm{nb}}(k)$ has already been interpolated if required.

For an efficient BWE algorithm a very *compact* feature vector is needed: The dimension of $\mathbf{x}$ shall be as low as possible to keep a low computational complexity but on the other hand the features shall provide as much usable information on the estimated wideband spectral envelope as possible.

A multitude of potential features $\mathbf{x}$ can be defined that are extracted by any linear or non-linear mapping from each frame of the narrowband speech signal $s_{\mathrm{nb}}(k)$. In this paper the usability of different features, well-tried and new ones, for the bandwidth extension problem shall be investigated. The "quality" of the features is quantified by two instrumental measures:

- Shannon's *mutual information* between the feature set $\mathbf{x}$ and the estimated quantity can be regarded as an indication of the feasibility of the estimation task [5, 6].
- The *separability* measure is well-known from statistics (e.g. [7]). It quantifies the discriminative power of a feature set for a classification task.

Note that the insights and results of this paper are mostly independent from the particular approach used for estimating the wideband spectral envelope.

2. MUTUAL INFORMATION

In an information theoretic perspective, the dependencies between different signals are described by their mutual information (MI).

The MI covers all kinds of linear and non-linear dependencies. Here, we want to investigate the mutual information $I(\mathbf{x};\mathbf{y})$ between the feature set $\mathbf{x}$ and the estimated quantity $\mathbf{y}$, i.e. the parameters representing the (true) wideband spectral envelope or the "missing" sub-band thereof[1]. This is motivated by [6] where it has been shown that for a specific MI the minimum achievable mean square estimation error is lower bounded. The larger the MI the lower is the bound. Hence, high mutual information $I(\mathbf{x};\mathbf{y})$ is a necessary condition for a high quality estimation of $\mathbf{y}$ from the observations $\mathbf{x}$. Note, however, that the bound is not necessarily tight such that a large mutual information alone is not sufficient to guarantee good estimation performance.

2.1. Numerical Approximation

For estimating the mutual information $I(\mathbf{x};\mathbf{y})$ we use a parametric approach because of the high dimension of the continuous vectors $\mathbf{x}$ and $\mathbf{y}$. The joint probability density function (PDF) $p(\mathbf{x},\mathbf{y})$ is approximated by a Gaussian mixture model (GMM), i.e. a sum of L weighted multivariate Gaussian densities $\mathcal{N}(\cdot)$ with mean vectors μ_l and covariance matrices $\mathbf{V}_l$

$$\tilde{p}(\mathbf{x},\mathbf{y}) = \sum_{l=1}^{L} \rho_l \, \mathcal{N}(\mathbf{x},\mathbf{y};\mu_l,\Sigma_l) \approx p(\mathbf{x},\mathbf{y}) \; . \tag{1}$$

The scalar weights ρ_l and the parameters μ_l and $\mathbf{V}_l$ of the individual Gaussians are trained by the expectation-maximization (EM) algorithm. Then, the mutual information is calculated numerically from the parameters of the GMM [8]

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &\approx E_{\tilde{p}(\mathbf{x},\mathbf{y})}\left\{\log \frac{\tilde{p}(\dot{\mathbf{x}},\dot{\mathbf{y}})}{\tilde{p}(\dot{\mathbf{x}})\,\tilde{p}(\dot{\mathbf{y}})}\right\} \\ &\approx \frac{1}{M}\sum_{\nu=1}^{M} \log \frac{\tilde{p}(\dot{\mathbf{x}}(\nu),\dot{\mathbf{y}}(\nu))}{\tilde{p}(\dot{\mathbf{x}}(\nu))\,\tilde{p}(\dot{\mathbf{y}}(\nu))} \; , \end{aligned} \tag{2}$$

where the vector pairs $\dot{\mathbf{x}}(\nu), \dot{\mathbf{y}}(\nu)$ are generated synthetically according to the model PDF $\tilde{p}(\mathbf{x},\mathbf{y})$. In our investigations we have used $L = 256$ Gaussians with full covariance matrices. The numerical evaluation of Eq. (2) was performed with $M = 10^6$ synthetic vector pairs.

2.2. Properties

From the definition of mutual information, e.g. [9], a number of properties of this measure for feature selection can be found:

- If the relation between two different feature vectors is defined by a bijective mapping, the MI is identical for both feature vectors. In this case, the MI measure does not provide any hint which feature set shall be preferred.
- If several parameters of the narrowband speech (say x_A, x_B and x_C) form a Markov chain $x_A \rightarrow x_B \rightarrow x_C$, i.e., if x_C can be calculated from x_B, and x_B is calculated from x_A, it appears favorable to select the very first element x_A of the chain as a feature. Due to the data processing inequality [9], MI is maximized by this choice.
- For combined feature vectors the MI cannot be simply added. In general, the inequality $I(\mathbf{x}_A,\mathbf{x}_B;\mathbf{y}) \leq I(\mathbf{x}_A;\mathbf{y}) + I(\mathbf{x}_B;\mathbf{y})$ applies.

3. SEPARABILITY

From the field of pattern recognition the separability is known as a measure for the quality of a particular feature set for a classification problem [7]. In the BWE application the class definitions should best be adopted to the method used to estimate the wideband spectral envelope: for example, if codebook mapping is used [2] the classes should correspond to the correct codebook indices as computed from true wideband speech. For an HMM-based approach [4] the classes should be the true HMM state information.

The separability measure can be calculated from a labeled set of training data, i.e. for each feature vector in the set the corresponding class must be known. Let Ξ_i denote the set of feature vectors $\mathbf{x}$ assigned to the i-th class. The number of feature vectors in the i-th set is $N_{\Xi_i} = |\Xi_i|$. The total number of frames in the training data is denoted by N_m. From the training data the *within-class* covariance matrix

$$\mathbf{V}_\mathbf{x} = \frac{1}{N_m}\sum_{i=1}^{N_S}\sum_{\mathbf{x}\in\Xi_i}(\mathbf{x}-\mu_i)(\mathbf{x}-\mu_i)^T \tag{3}$$

and the *between-class* covariance matrix

$$\mathbf{B}_\mathbf{x} = \sum_{i=1}^{N_S}\frac{N_{\Xi_i}}{N_m}(\mu_i-\mu)(\mu_i-\mu)^T \tag{4}$$

are calculated, where

$$\mu_i = \frac{1}{N_{\Xi_i}}\sum_{\mathbf{x}\in\Xi_i}\mathbf{x} \quad \text{and} \quad \mu = \sum_{i=1}^{N_S}\frac{N_{\Xi_i}}{N_m}\,\mu_i \; . \tag{5}$$

The separability measure shall be larger if the between-class covariance gets smaller or if the within-class covariance gets larger. Accordingly, the separability measure is empirically defined by the term $\mathbf{J}_\mathbf{x} = \mathbf{V}_\mathbf{x}^{-1}\,\mathbf{B}_\mathbf{x}$. To obtain a scalar measure for the separability of the classes a trace criterion is used [7]

$$\zeta(\mathbf{x}) = \text{tr}\;\mathbf{J}_\mathbf{x} = \text{tr}\;\left(\mathbf{V}_\mathbf{x}^{-1}\,\mathbf{B}_\mathbf{x}\right) \; . \tag{6}$$

The separability depends on the definition of the classes. Comparing $\zeta(\mathbf{x})$ for different feature vectors $\mathbf{x}$ with the same class definitions, a larger value indicates a better suitability of the corresponding feature vector for classification and estimation.

The separability measure has the following properties:

- The definition of the separability measure is based on the implicit assumption of a normal distribution of the feature vectors that are assigned to each class. If this assumption is not valid, the significance of the separability measure is reduced.
- By the separability measure all classes are treated alike. Therefore, the separability of two very similar classes (w.r.t. the represented speech sound) is rated like the separability of two very different classes. Hence, a maximization of the separability not necessarily leads to the optimum achievable estimation performance (in the MMSE sense) of the subsequent estimation rule.
- In general the values of the separabilities can not be added up if several features are assembled to a composite feature vector. In this case the separability of the composite feature vector must be measured anew.

[1] The true values of the representation $\mathbf{y}$ can be calculated if wideband speech is available (cf. [6]).

3.1. Linear Discriminant Analysis

The purpose of the *linear discriminant analysis* (LDA) is to obtain a feature vector with maximal compactness [7]: starting from the high-dimensional "super-vector" $\mathbf{x}_0$ the dimension of the feature vector $\mathbf{x}$ shall be reduced while the discriminating power shall be retained or decreased as little as possible. The reduction of dimension is performed (during BWE) by means of a linear transformation

$$\mathbf{x} = \mathbf{H}^{-1}\,\mathbf{x}_0 \tag{7}$$

where the matrix $\mathbf{H}$ is a $\beta \times b$ matrix with $b = \dim \mathbf{x} < \beta = \dim \mathbf{x}_0$. The transformation matrix $\mathbf{H}$ shall further be an orthonormal matrix.

The matrix $\mathbf{H}$ is optimized such that the separability of $\mathbf{x}$ is maximized [7]

$$\begin{aligned} \mathbf{H} &= \arg\max_{\mathbf{H}} \zeta(\mathbf{x})\ , \quad \text{where} & (8)\\ \zeta(\mathbf{x}) &= \mathrm{tr}\left(\mathbf{V}_{\mathbf{x}}^{-1}\mathbf{B}_{\mathbf{x}}\right) = \mathrm{tr}\left(\mathbf{H}^T\mathbf{V}_{\mathbf{x}_0}^{-1}\mathbf{B}_{\mathbf{x}_0}\mathbf{H}\right)\ . \end{aligned}$$

The solution to Eq. (8) is achieved by composing the matrix $\mathbf{H}$ from the eigenvectors $\Phi_1, \Phi_2 \ldots \Phi_b$ that are assigned to the b largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_b$ of $\mathbf{V}_{\mathbf{x}_0}^{-1}\mathbf{B}_{\mathbf{x}_0}$. The computationally complex preparation of the transformation matrix $\mathbf{H}$ is performed off-line during the training phase of the BWE algorithm.

The LDA makes it possible to take many primary features of the bandlimited speech signal into account, using a high-dimensional super-vector $\mathbf{x}_0$. Nevertheless, the dimension of $\mathbf{x}$ can be small — without loosing too much discriminating power — such that the complexity and memory consumption of the subsequent estimation algorithm are low.

4. EVALUATION

In this section the typical application of *high* frequency bandwidth extension will be considered. That is, the narrowband speech signal $s_{\mathrm{nb}}(k)$ has frequency components in the range of 0.3–3.4 kHz. By the BWE algorithm a wideband signal $s_{\mathrm{wb}}(k)$ with frequency components up to 7 kHz shall be produced.

For measuring mutual information and separability, the speech signals are subdivided into frames with a length of 20 ms. For each signal frame the primary features and (from corresponding wideband speech) the vector $\mathbf{y}$ is determined. The vector $\mathbf{y}$ consists of weighted cepstral coefficients representing the gain and shape of the spectral envelope within the missing frequency band (3.4–7 kHz) [6, 4]. All of the measurements were performed using the BAS[2] SI100 speech corpus consisting of about 36 hours of clean German speech spoken by about 100 male and female speakers.

4.1. Primary Features

The following elementary feature vectors and scalar features were investigated:

- The normalized auto-correlation function (ACF). Both the ACF vector consisting of the first ten coefficients as well as the (scalar) auto-correlation coefficients for the lags of 1 and the pitch period are considered.
- The LPC and LSF coefficients as calculated from the ACF as well as the LPC-derived cepstrum.
- The cepstrum and the *mel frequency cepstral coefficients* (MFCC), both calculated directly from the narrowband speech frame.
- The normed frame energy (frame index m)

$$x_{\mathrm{nrp}}(m) = \frac{\log E(m) - \log E_{\min}(m)}{\log \bar{E}(m) - \log E_{\min}(m)}\ , \tag{9}$$

  with

$$\begin{aligned} E(m) &= \sum_{\kappa=0}^{N-1} s_{\mathrm{nb}}^2(\kappa)\\ E_{\min}(m) &= \min_{\mu=0}^{N_{\min}} E(m-\mu)\\ \bar{E}(m) &= \alpha\,\bar{E}(m-1) + (1-\alpha)\,E(m)\ , \end{aligned}$$

  where the forgetting factor is adjusted to $\alpha = 0.96$, and the minimum search window has a size of $N_{\min} = 200$. The number of samples per frame is N.

- The gradient index [10]

$$x_{\mathrm{gi}} = \sum_{\kappa=2}^{N} \frac{\Psi(\kappa)\,|s_{\mathrm{nb}}(\kappa) - s_{\mathrm{nb}}(\kappa-1)|}{\sqrt{\frac{1}{N}E(m)}}\ . \tag{10}$$

  The variable $\psi(\kappa)$ denotes the sign of the gradient $s_{\mathrm{nb}}(\kappa) - s_{\mathrm{nb}}(\kappa-1)$, i.e., $\psi(\kappa) \in \{-1, 1\}$, and $\Psi(\kappa) = 1/2\,|\psi(\kappa) - \psi(\kappa-1)|$.

- The zero crossing rate.
- The pitch period, estimated by the lag in the range of 20 to 130 (sampling rate 8 kHz) for which the ACF is maximal.
- An estimate of the local kurtosis

$$x_{\mathrm{k}} = \log \frac{1}{N}\sum_{\kappa=0}^{N-1} s_{\mathrm{nb}}^4(\kappa) - 2\log\frac{1}{N}E(m)\ . \tag{11}$$

- The spectral centroid

$$x_{\mathrm{sc}} = \frac{\sum_{i=0}^{M/2} i \cdot \left|S_{\mathrm{nb}}(e^{j\Omega_i})\right|}{\left(\frac{M}{2}+1\right)\sum_{i=0}^{M/2}\left|S_{\mathrm{nb}}(e^{j\Omega_i})\right|}\ . \tag{12}$$

  The quantity $S_{\mathrm{nb}}(e^{j\Omega_i})$ labels the i-th coefficient of a discrete Fourier transformation of the length M of the input signal frame.

- The spectral flatness

$$x_{\mathrm{sfm}} = \frac{\sqrt[M]{\prod_{i=0}^{M-1}\left|S_{\mathrm{nb}}(e^{j\Omega_i})\right|^2}}{\frac{1}{M}\sum_{i=0}^{M-1}\left|S_{\mathrm{nb}}(e^{j\Omega_i})\right|^2}\ . \tag{13}$$

The reader is refered to [11] for a more detailed description of these primary features.

4.2. Mutual Information and Separability

The estimated mutual information between $\mathbf{y}$ and the investigated primary features are listed in Tab. 1. It can be subsumed that the features describing the spectral envelope of the bandlimited speech in fact play a major role for the bandwidth extension. Both the mutual information and the separability measures are maximal for

[2] Bavarian Archive for Speech Signals
(`http://www.phonetik.uni-muenchen.de/Bas`)

these features. It must be taken into account, however, that the dimension of the primary features from this group is ten times higher than those of the scalar features.

feature vector $\mathbf{x}$	dim $\mathbf{x}$	$I(\mathbf{x};\mathbf{y})$	$\zeta(\mathbf{x})$
		[bit/frame]	(16 classes)
ACF	10	2.6089	1.6349
LPC	10	2.3054	1.5295
LSF	10	2.3597	1.5596
LPC-cepstrum	10	2.2401	1.4282
cepstrum	10	2.3075	1.5483
MFCC	10	2.3325	2.2659
ACF (1)	1	0.7514	1.1237
ACF (pitch period)	1	0.4450	0.4058
frame energy	1	0.9285	1.0756
gradient index	1	0.8011	1.2520
zero crossing rate	1	0.7453	1.0795
pitch period	1	0.2451	0.0530
local kurtosis	1	0.2037	0.0225
spectral centroid	1	0.7913	1.0179
spectral flatness	1	0.4387	0.3538

Table 1. Estimates of mutual information $I(\mathbf{x};\mathbf{y})$ and separability $\zeta(\mathbf{x})$ for BWE of the high frequency band (3.4–8 kHz) from telephone speech (0.3–3.4 kHz). For calculating the separability the 16 classes were defined by vector quantizing $\mathbf{y}$ [4].

To achieve the best results with the BWE algorithm it can further be motivated from Tab. 1 to additionally include certain scalar features in the feature vector. Particularly, the consideration of the frame energy as well as the gradient index, zero crossing rate and/or spectral centroid seems to be very promising.

4.3. Linear Discriminant Analysis

To evaluate the impact of the linear discriminant analysis, the estimation quality obtained with the transformed feature vectors was determined. The HMM-based MMSE estimation rule from [4] was used with 64 HMM states and 16 mixture components in the state-specific GMMs. Both speaker-dependent and speaker-independent models were investigated. The results are expressed in terms of the *root mean square log spectral distortion* (RMS LSD) of the estimated spectral envelope within the missing frequency band (3.4–8 kHz) [6, 4]. The 15-dimensional feature super-vector $\mathbf{x}_0$ consisted of the first ten normalized auto-correlation coefficients, the zero crossing rate, the normed frame energy, the gradient index, the local kurtosis, and the spectral centroid.

In Fig. 2 the mean performances are depicted that were obtained both without and with the application of LDAs for the dimensions $b = 1 \ldots 5$. As expected, the distortions of the estimates are decreased by increasing the dimension of the LDA transform. Remarkably, the achieved performances with a dimension of the LDA transform of $b = 5$ are even superior to those of the estimator that uses the original non-transformed feature vectors with a dimension of $\beta = 15$. This effect is the result of the improved compactness of the feature vectors: if the dimension of the feature vectors $\mathbf{x}$ is reduced significantly, the quality of the statistical modeling is enhanced.

5. CONCLUSIONS

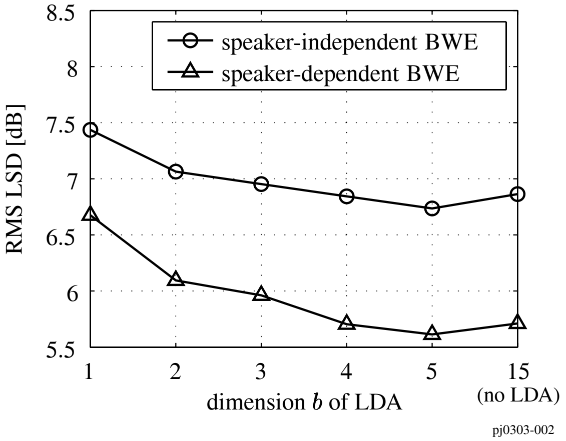

Fig. 2. Impact of a linear discriminant analysis on the estimation performance.

It has been shown that, in addition to the well-tried spectral envelope parameters of the bandlimited speech, characteristics of the excitation of the input speech, such as gain or voicing, should be included in the feature vector $\mathbf{x}$. Furthermore, it is found that by utilizing a linear discriminant analysis the performance and robustness of the bandwidth extension system can be improved, yet simultaneously reducing the computational complexity of the estimation algorithm substantially.

6. REFERENCES

[1] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 4, pp. 544–548, Oct. 1994.

[2] H. Carl, *Untersuchung verschiedener Methoden der Sprachkodierung und eine Anwendung zur Bandbreitenvergrößerung von Schmalband-Sprachsignalen*, Ph.D. thesis, Ruhr-Universität Bochum, Bochum, Germany, 1994, (in German).

[3] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. of EUSIPCO*, Edinburgh, Scotland, Sept. 1994, vol. 2, pp. 1178–1181.

[4] P. Jax and P. Vary, "On artificial bandwidth extension of speech signals," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.

[5] M. Nilsson, H. Gustafsson, S. V. Andersen, and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. of ICASSP*, Orlando, FL, USA, May 2002, vol. 1, pp. 525–528.

[6] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals," in *Proc. of ICASSP*, Orlando, FL, USA, May 2002, vol. 1, pp. 237–240.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Morgan Kaufmann, Academic Press, San Francisco, San Diego, 2nd edition, 1990.

[8] P. Hedelin and J. Skoglund, "Vector quantization based on Gaussian mixture models," *IEEE Trans. Speech and Audio Proc.*, vol. 8, no. 4, pp. 385–401, July 2000.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, 1991.

[10] J. W. Paulus, "Variable rate wideband speech coding using perceptually motivated thresholds," in *IEEE Speech Coding Worksh.*, Annapolis, MD, USA, Sept. 1995, pp. 35–36.

[11] P. Jax, *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*, Ph.D. thesis, Aachen University (RWTH), Aachen, Germany, 2002.