SPEECH - NONSPEECH DISCRIMINATION BASED ON SPEECH-RELEVANT SPECTROGRAM MODULATIONS

Michael Wohlmayr, Maria Markaki, and Yannis Stylianou

Computer Science Department, University of Crete
Knossou Ave., 71409, Heraklion, Greece
phone: + (30) 2810 393559, fax: + (30) 2810 393592, {micki_w, mmarkaki, yannis}@csd.uoc.gr

ABSTRACT

In this work, we adopt an information theoretic approach - the Information Bottleneck method - to extract the relevant modulation frequencies across both dimensions of a spectrogram, for speech / non-speech discrimination (music, animal vocalizations, environmental noises). A compact representation is built for each sound ensemble, consisting of the maximally informative features. We demonstrate the effectiveness of a simple thresholding classifier which is based on the similarity of a sound to each characteristic modulation spectrum. When we assess the performance of the classification system at various SNR conditions using F-measure, results are equally good to a recently proposed method based on the same features but having significantly greater complexity.

1. INTRODUCTION

Robust automatic audio classification and segmentation in real world conditions is a research area of great interest with applications in many areas of speech technology like speech and speaker recognition, and in multimedia processing for automatic labeling and extraction of semantic information. It has been argued [1] that the statistical analysis of natural sounds - including animal vocalizations and speech - could reveal the neural basis of acoustical perception. Insights in the auditory processing could be exploited in the speech and audio engineering applications listed above.

It is worth to note that all natural sounds are characterized by slow spectral and temporal modulations [1]. However, auditory neurons seem to be able to discriminate relevant from irrelevant sound ensembles, by tuning to the auditory features that differ most across them [2]. Speech is characterized by joint spectro-temporal energy modulations; oscillations in power across spectral and temporal axes in spectrogram reflect formant peaks and their transitions, spectral edges, and fast amplitude modulations at onsets-offsets. Of particular relevance to speech intelligibility are the slow temporal modulations (few Hz) which correspond to the phonetic and syllabic rates of speech [3].

Spectrogram modulations at multiple resolutions can be estimated using the auditory model of Shamma et al [4]. The model has been successfully applied in the assessment of speech intelligibility [5], the discrimination of speech from non-speech [6], and other simulations of psychoacoustical phenomena [7]. These auditory representations of sounds are highly redundant, which might yield an advantage in the presence of noise and uncertainty since this adds robustness. However, the *curse of dimensionality* states that the number of training examples required to achieve a fixed upper bound on a classifier generalization error, grows exponentially with the feature dimensions. It is crucial, then, to reduce dimensionality in such a way that the remaining set of features still captures enough information about a class.

A generalization of the Singular Value Decomposition (SVD) to higher - order tensors, *Higher Order SVD* (HOSVD) [8], has been applied to the auditory features in [6]. HOSVD allows to remove redundancies from each subspace separately, permitting to choose the number of dimensions to keep per subspace. Application of HOSVD to tensors is quite similar to principal component analysis (PCA) of vectors. These techniques yield the dimensions which best represent the data, but might be suboptimal for data classification [9]. An alternative method of dimensionality reduction is the *Information Bottleneck Method* (IB) proposed by Tishby et al [11]. The IB method enables to construct a compact representation for each class, maintaining its most relevant features. In [12], a general speech-oriented implementation of IB has been presented, using Mel frequency cepstral coefficients (MFCC). According to the recognition task, a small subset of MFCCs was selected which preserved high mutual information about the target variable [12].

In this paper, we estimate the power distribution in the modulation spectrum of speech signals, and compare it to the modulation statistics of other sounds. The auditory model of Shamma et al [4] is the basis for these estimations. Using IB method, we show that an efficient dimensionality reduction is achieved while modulation frequencies which distinguish speech from other sounds are preserved (and estimated). A simple thresholding classifier is proposed, which is based on the similarity of sounds to the compact modulation spectra. Its performance is compared to the system of [6] which uses HOSVD [8] before classification with Support Vector Machines (SVMs). According to F-measure, our system is almost equivalent to the system of [6], in spite of its significantly lower complexity. For evaluation purposes, we have also implemented another system based on Mel Frequency Cepstral Coefficients (MFCCs), Zero Crossing Rates (ZCRs) and SVM classifiers. This served as a reference system to show the robustness of auditory features to various noise conditions.

The auditory model of Shamma et al [4] is presented in brief in Section 2. In Section 3 we describe the information theoretic principle, the sequential information bottleneck procedure applied to auditory features and the thresholding classifier. In Section 4 we compare the performance of the proposed system, the system in [6] and the reference system (MFCCs and ZCRs) on a benchmark set using F-measure at various SNR conditions.

2. COMPUTATIONAL AUDITORY MODEL

Early stages of the model estimate an enhanced spectrum of sounds, while at later stages spectrum analysis occurs: fast and slow modulation patterns are detected by arrays of filters centered at different frequencies, with Spectro-Temporal Response Functions (STRFs) resembling the receptive fields of auditory midbrain neurons [5]. These have the form of a spectro-temporal Gabor function, selective for specific frequency sweeps, bandwidth, etc., performing actually a multiresolution wavelet analysis of the spectrogram [4]. The auditory based features are collected from an audio signal in a frame-per-frame scheme. For each time frame, the auditory representation is calculated on a range of frequencies, scales (of spectral resolution) and rates (temporal resolution). In this study, the scales are set to $s = [0.5, 1, 2, 4, 8]$ cyc/oct, the rates (positive and negative) to $r = [1, 2, 4, 8, 16, 32]$ Hz. The extracted information is averaged over time, therefore resulting in a 3-dimensional array, or third-order tensor. The dimensionality of this set covers 128 logarithmic frequency bands $\times$ 5 scales $\times$ 12 rates. We have used the "NSL Tools" MATLAB package (courtesy of the Neural Systems Laboratory, University of Maryland, College Park, downloadable from http://www.isr.umd.edu/CAAR/pubs.html).

3. INFORMATION BOTTLENECK METHOD

In Rate Distortion theory a quantitative measure for the quality of a compact representation is provided by a *distortion function*. In general, definition of this function depends on the application: in speech processing, the relevant acoustic distortion measure is rather unknown, since it is a complex function of perceptual and linguistic variables [12]. IB method provides an information theoretic formulation and solution to the tradeoff between compactness and quality of a signal's representation [11, 13, 12]. In the supervised learning framework, features are regarded as relevant if they provide information about a target. IB method assumes that this additional variable Y (the target) is available. In the case of speech processing systems, the available tagging Y of the audio signal (as speech / non speech, speakers or phonemes) guides the selection of features during training. The relevance of information in the representation of an audio signal, denoted by X, is defined as the amount of information it holds about the other variable Y. If we have an estimate of their joint distribution $p(x,y)$, a natural measure for the amount of relevant information in X about Y is given by Shannon's mutual information between these two variables:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where the discrete random variables $x \in X$ and $y \in Y$ are distributed according to $p(x)$, and $p(y)$, respectively. Further, let $\tilde{x} \in \tilde{X}$ be another random variable which denotes the compressed representation of x; x is transformed to $\tilde{x}$ by a (stochastic) mapping $p(\tilde{x}|x)$. Our aim is to find an $\tilde{X}$ that compresses X through minimization of $I(\tilde{X};X)$, i.e. the mutual information between the compressed and the original variable. At the same time, the compression of the resulting representation $\tilde{X}$ should be minimal *under the constraint* that the relevant information in $\tilde{X}$ about Y, $I(\tilde{X};Y)$ stays above a certain level. This constrained optimization problem can be expressed via Lagrange multipliers, with the minimization of the *IB variational functional*:

$$\mathscr{L}\{p(\tilde{x}|x)\} = I(\tilde{X};X) - \beta I(\tilde{X};Y) \tag{2}$$

where β, the positive Lagrange multiplier, controls the tradeoff between compression and relevance. The solution to this constrained optimization problem has yielded various iterative algorithms that converge to a reduced representation $\tilde{X}$, given $p(x,y)$ and β [13]. We choose the *sequential optimization algorithm* (sIB), as we want a fixed number of hard clusters as output. We use the "IBA-1.0: Matlab Code for Information Bottleneck Clustering Algorithms" (N. Slonim, http://www.cs.huji.ac.il/~noamm, 2003).

The input consists of the joint distribution $p(x,y)$, the tradeoff parameter β and the number of clusters $M = |\tilde{X}|$. During initialization, the algorithm creates a random partition $\tilde{X}$, i.e. each element $x \in X$ is randomly assigned to one of the M clusters $\tilde{X}$. Afterwards, the algorithm enters an iteration loop. At each iteration step, it cycles through all $x \in X$ and tries to assign them to a different cluster $\tilde{X}$ in order to *increase* the IB functional:

$$\mathscr{L}_{max} = I(\tilde{X};Y) - \beta^{-1} I(\tilde{X};X). \tag{3}$$

This is equivalent to minimization of the functional defined in equation 2, and it is used for consistency with [13]. The algorithm terminates when the partition does not change during one iteration. This is guaranteed because $\mathscr{L}_{max}$ is always upper bounded by some finite value. To prevent the convergence of the algorithm to a local maximum (i.e., a suboptimal solution), we perform several runs with different initial random partitions [13].

3.1 Application to Cortical Features

The feature tensor $\mathscr{Z} \in \mathbb{R}^{+F \times R \times S}$ represents a discrete set of *continuous* features $z_{i_1,i_2,i_3} = (\mathscr{Z})_{i_1,i_2,i_3}$, where F, R and S denote the number of frequencies, rates and scales, respectively. Since each response z_{i_1,i_2,i_3} is collected over a time frame, it can be interpreted as the average count of an inherent binary event (in the case of a neuron, this would be a spike). We therefore consider each response at a location indexed by (i_1,i_2,i_3), as a binary feature whose number of occurences in a time interval is represented by z_{i_1,i_2,i_3}.

Let the location of a response be denoted by x_i, where $i = 1,\ldots,F \times R \times S$, such that $z_{i_1,i_2,i_3} = z_{x_i}$. The 3−dimensional modulation spectrum (frequency - rate - scale) is divided then into $F \times R \times S$ bins centered at $(f_{i_1}, r_{i_2}, s_{i_3})$. Given a training list of N feature tensors $\mathscr{Z}^{(k)}$ and their corresponding targets $y^{(k)}$, $k = 1,\ldots,N$, $y = 1,2$ (the nonspeech and speech tags, respectively), we can now build a count matrix $K(x,y)$ which indicates the frequency of occupancy of the i^{th} discrete subdivision of the modulation spectrum in the presence of a certain target value $y^{(k)}$. Normalizing this count matrix such that its elements sum to 1, provides an estimate of the joint distribution $p(x,y)$, which is all the IB framework requires. We assume that N is large enough such that the estimate of $p(x,y)$ is reliable, although it has been reported that satisfactory results were achieved even in cases of extreme undersampling [13].

For the purpose of discrimination, the target variable Y has only two possible values, $y_1 = 1$ (nonspeech) and $y_2 = 2$ (speech). We choose to cluster the features X into 3 groups,

one composed of features relevant to y_1, the second of features relevant to y_2, whereas the third cluster includes features less relevant to a specific class. Since this setting already implies a degree of compression, we decided to set $\beta^{-1} = 0$ and concentrate on solutions that maximize the relevant information term only. Let us denote a compressed representation (a reduced feature set) by $\tilde{X}$ and the deterministic mapping obtained by sIB algorithm as $p(\tilde{x}|x)$. We discard the cluster $\tilde{X}_j$ whose contribution :

$$C_{I(\tilde{X};Y)}(\tilde{X}_j) = \sum_y p(\tilde{x}_j, y) \log \frac{p(\tilde{x}_j, y)}{p(\tilde{x}_j)p(y)} \tag{4}$$

to $I(\tilde{X};Y)$ is minimal, because its features are mostly irrelevant in this case. Therefore, we don't even have to estimate the responses at these locations of the modulation spectrum (in contrast to the HOSVD approach [6]). This implies an important reduction in computational load, still keeping the maximally informative features with respect to the task of speech-nonspeech discrimination. To find out the identity of the remaining two clusters, we compute:

$$p(\tilde{x}, y) = \sum_x p(x, y) p(\tilde{x}|x) \tag{5}$$

$$p(\tilde{x}) = \sum_y p(\tilde{x}, y) \tag{6}$$

$$p(y|\tilde{x}) = \frac{p(\tilde{x}, y)}{p(\tilde{x})} \tag{7}$$

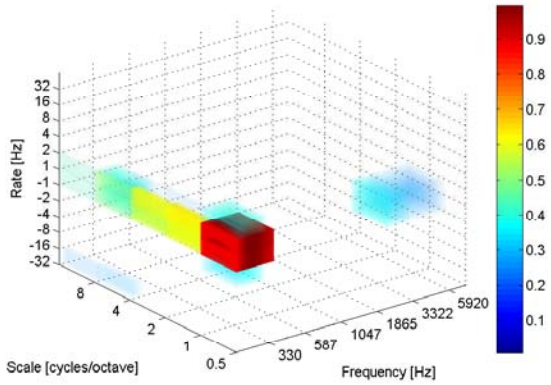
The cluster that maximizes the likelihood $p(y_1|\tilde{x})$ contains the most relevant features for y_1; the other for y_2. We denote, hence, the first cluster as $\tilde{X}_1$ and the latter as $\tilde{X}_2$. The typical pattern (3-dimensional distribution) of features relevant for y_1 is given by $p(x|\tilde{x} = \tilde{x}_1)$, while for y_2 is given by $p(x|\tilde{x} = \tilde{x}_2)$. According to Bayes rule, these are defined as:

$$p(x|\tilde{x} = \tilde{x}_j) = \frac{p(\tilde{x} = \tilde{x}_j|x)p(x)}{p(\tilde{x} = \tilde{x}_j)}, \qquad j = 1,2 \tag{8}$$

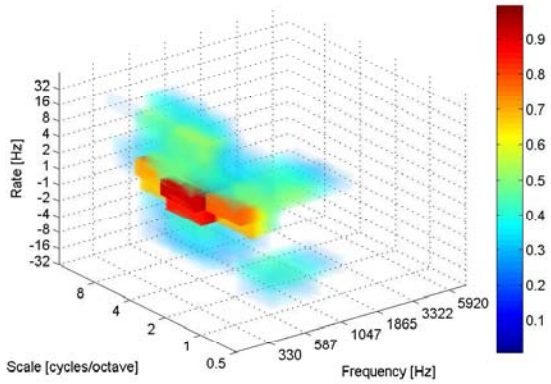
Figure 1 presents an example of the relevant modulation spectrum of each sound ensemble, speech and non-speech. On average, strongest speech-relevant modulations are between $1-8$ cyc/octave (scale), -1 and 2 Hz (rate), and in the $300-600$ Hz frequency range. Knowledge of such compact modulation patterns allows us to classify new incoming sounds based on the similarity of their cortical-like representation (the feature tensor $\mathscr{Z}$) to the typical pattern $p(x|\tilde{x} = \tilde{x}_1)$ or $p(x|\tilde{x} = \tilde{x}_2)$. We assess the similarity (or correlation) of $\mathscr{Z}$ to both patterns by their inner (tensor) product (a compact one dimensional feature). We propose the ratio of these similarity measures, denoted as *relevant response ratio*:

$$R(\mathscr{Z}) = \frac{< \hat{\mathscr{Z}}, p(x|\tilde{x} = \tilde{x}_2) >}{< \hat{\mathscr{Z}}, p(x|\tilde{x} = \tilde{x}_1) >} \gtrless \lambda \tag{9}$$

where $\hat{\mathscr{Z}}$ is the normalized feature tensor. Large values of R give strong indications toward target y_2, small values toward y_1. For the purpose of classification a threshold (λ) has to be defined such that any sound whose corresponding relevant response ratio R is above λ is classified as speech, otherwise as nonspeech. We calculate the relevant response ratio R for all training examples and noise conditions. Figure 2 shows



(a)



(b)

Figure 1: $p(x|\tilde{x} = \tilde{x}_1)$ for non-speech (a) and $p(x|\tilde{x} = \tilde{x}_2)$ for speech (b). Cluster $\tilde{X}_1$ holds 24.7% and $\tilde{X}_2$ holds 37.5% of all responses. The remaining 37.8% are discarded as irrelevant.

the histograms of R computed on speech and non-speech examples. It is worth to note that the histograms form two distinct clusters for every SNR, with a small degree of overlap. Obviously, decision threshold λ is highly dependent on the SNR condition under which the features are extracted. This is especially true for low SNR conditions (0dB, -10dB).

3.2 Computational Complexity

It is worth to compare the computational complexity of this approach (system 2) with the system based on HOSVD [6] (system 1), both for training and testing.

Training complexity of system 1 is dominated by the HOSVD [8], which for this purpose boils down to three matrix SVDs. The dimension of these matrices is $F \times RSN$, $R \times FSN$ and $S \times FSN$, respectively, where N denotes the number of training examples. As in each case the number of rows is much smaller than the number of columns, it is beneficial to compute the right singular vectors of each *transposed* matrix using the modified Golub-Reinsch algorithm, which, for an arbitrary $m \times n$ matrix with $m > n$, is reported to have complexity $2mn^2 + 11n^3$ [10]. Thus, system 1 is dominated by an overall complexity of $O\left(R^3 + S^3 + F^3\right) + N(FSR^2 + RFS^2 + RSF^2)$.

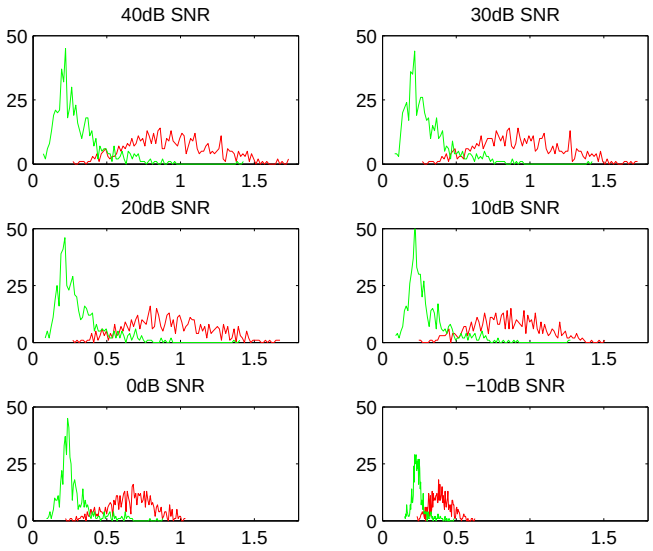


Figure 2: Histogram of relevant response ratios computed on nonspeech (gray/green) and speech examples (black/red).

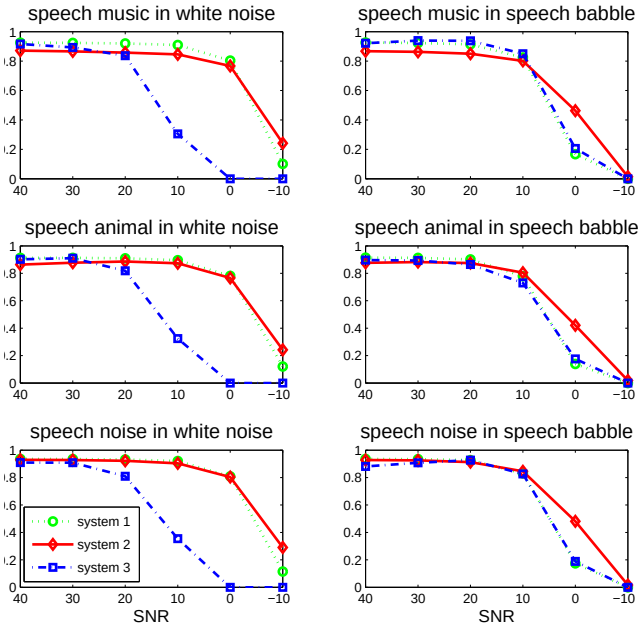


Figure 3: F measure of systems applied to all signal types of the benchmark test : (Left) with additive white noise (right) with speech babble.

In contrast, system 2 is dominated by the complexity of the sIB algorithm, which is reported to be $O(|X||\tilde{X}||Y|)$ [13], where in the current setting $|X| = FRS$, $|\tilde{X}| = 3$ and $|Y| = 2$.

For testing, the complexity of system 1 is dominated by the reduction of each tensorial feature $\mathscr{X} \in \mathbb{R}^{+F\times R\times S}$ to its shrinked version $\tilde{\mathscr{X}} \in \mathbb{R}^{+\tilde{F}\times\tilde{R}\times\tilde{S}}$. This is achieved by the n-mode multiplication of $\mathscr{X}$ with three matrices, $\mathbf{U}^{(1)} \in \mathbb{R}^{\tilde{F}\times F}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{\tilde{R}\times R}$ and $\mathbf{U}^{(3)} \in \mathbb{R}^{\tilde{S}\times S}$. Complexity of the n-mode product of an arbitrary tensor $\mathscr{A} \in \mathbb{R}^{I_1\times\ldots\times I_{n-1}\times I_n\times I_{n+1}\times\ldots\times I_N}$ and a matrix $\mathbf{U} \in \mathbb{R}^{J_n\times I_n}$ is governed by the computation of the product $\mathbf{U}\mathbf{A}_{(n)}$, where $\mathbf{A}_{(n)} \in \mathbb{R}^{I_n\times I_1\ldots I_{n-1}I_{n+1}\ldots I_N}$ denotes the n-mode matrix unfolding of $\mathscr{A}$ [8]. Assuming for simplicity a straightforward implementation of the matrix-product, the complexity of the overall reduction step os system 1 is $O\left(FRS\tilde{F} + RS\tilde{F}\tilde{R} + S\tilde{F}\tilde{R}\tilde{S}\right)$.

Again, system 2 exhibits a lower complexity for testing. First, the set of relevant response features, obtained from training, guides the decision which modulations do not even need to be computed. Then, for the remaining set of features, complexity is governed by normalizing the obtained feature tensor and computing the relevant response ratio: $O(FRS)$.

3.3 Database and feature extraction

Speech examples were taken from the TIMIT Acoustic-Phonetic Continuous Speech Corpus. Music examples were selected from the authors' music collection. Animal vocalizations consist of bird sounds [14]. The noise examples (taken from Noisex) consist of background speech babble in locations such as restaurants and railway stations, machinery noise and noisy recordings inside cars and planes. Training set consists of 500 speech and 560 non-speech samples. One single frame of 500 ms is extracted from each example, starting at a certain sample offset in order to skip initial periods of silence.

From each of these frames, a feature tensor $\mathscr{X}$ holding the cortical responses is extracted to train the systems which are based on the same auditory features: System 1 reduces their dimensionality using the HOSVD, and classifies the final set of features with SVM [6]. System 2 (the proposed one) defines relevant subsets of auditory features according to IB method, and classifies them with the Relevant Response Ratio and a fixed threshold. Likewise, one feature vector $\mathbf{z}$ holding MFCC and ZCR features is extracted from each of these frames to train the 3^{rd} system. This system subsequently uses SVM classification. We train each system in a specific SNR condition chosen such that the expected classification performance is high for a broad range of test conditions: this is 10 dB for systems 1 and 2, and 40 dB for system 3.

Test set consists of 260 speech and 300 non-speech examples. Sentences and speakers in test examples are different from the training examples. Since we want to evaluate the robustness and applicability of the systems under realistic conditions, we construct a *benchmark test* consisting of a variety of labeled sound signals. Each signal is 30 seconds long, and consists of alternating speech - nonspeech test examples with random length (between 2 and 8 seconds). We create 30 such signals, consisting of alternating speech and music, noise, or animal vocalization events. Each of them is corrupted either by additive white noise, or speech babble, at SNRs of 40, 30, 20, 10, 0, and -10 dB, resulting in 360 test signals.

4. EXPERIMENTAL RESULTS

We evaluate systems performance in terms of the F-measure for each non-speech ensemble (music, noise, or animal vocalizations), noise type and level. The F-measure is a common tool to assess the performance of an information retrieval system based on two quantitative measures, precision P and recall R. The results are presented in Figure 3. Both

systems 1 and 2 - which are based on the same auditory features - exhibit equally good performance, with generalization ability to various noise conditions for both types of noise. The performance of the 3rd system, which is based on MFCC and ZCR features, degrades remarkably when corrupted by additive white noise, whereas it exhibits a better generalization ability in the case of additive speech babble.

5. CONCLUSIONS

Classical methods of dimensionality reduction seek the optimal projections to represent the data in a low - dimensional space. Dimensions are discarded based on the relative magnitude of the corresponding singular values, even if these particular dimensions could give a clue for classification. In this paper, an information theoretic approach enables the selection of a reduced set of auditory features which are maximally informative in respect to the target - speech or non-speech in this case. A simple thresholding technique is proposed, built upon these reduced representations. It yields a performance close to state-of-the-art classifiers, such as SVMs, with a significantly reduced computational load. An obvious refinement of the system would be the inclusion of a noise energy measure in order to adapt the decision threshold to the observed SNR (according to Figure 2).

Since we wanted to evaluate the process of feature selection per se, we preferred not to use more complex classifiers in this task. In future work, we could test an unsupervised clustering method for the classification of test examples, using the same sequential optimization routine of the sIB algorithm [13]. The method could also be tailored to the recognition of other speech attributes, such as speech or speaker recognition, based upon other features [12] in addition to the spectro-temporal modulations.

6. ACKNOWLEDGEMENTS

This work has been supported by the General Secretariat of Research and Technology (GGET) and SIMILAR Network of Excellence.

REFERENCES

[1] N.C. Singh and F.E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing", *J. Acoust. Soc. Amer.*, vol. 114, pp. 3394–3411, 2003.

[2] S.M.N. Woolley, T.E. Fremouw, A. Hsu and F.E. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," *Nature Neuroscience*, vol. 8, pp. 1371–1379, 2005.

[3] T.F. Quatieri, *Discrete-Time Speech Signal Processing*. Address: Prentice-Hall Signal Processing series, 2002.

[4] K. Wang and S.A. Shamma, "Spectral shape analysis in the central auditory system", *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 382–396, 1995.

[5] M. Elhilali, T. Chi and S.A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility", *Speech communication*, vol. 41, pp. 331–348, 2003.

[6] N. Mesgarani, M. Slaney and S.A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 920–930, May 2006.

[7] R.P. Carlyon and S.A. Shamma, "An account of monaural phase sensitivity", *J Acoust Soc Am*, vol. 114, pp. 333–346, 2003.

[8] L. De Lathauwer, B. De Moor and J. Vandewalle, "A multilinear singular value decomposition", *SIAM J Matrix Anal Appl*, vol. 21, pp. 1253–1278, 2000.

[9] R. Duda and P. Hart, *Pattern Classification*. Address: Wiley-Interscience, New York, 1999.

[10] G. Golub and C. van Loan, *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996

[11] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.

[12] R.M. Hecht and N. Tishby, "Extraction of relevant speech features using the Information Bottleneck method, " in *Proceedings of Interspeech*, Lisbon, Portugal, 2005.

[13] N. Slonim, *The Information Bottleneck: Theory and Applications*. PhD thesis: School of Engineering and Computer Science, Hebrew University, 2002.

[14] R. Specht, *Animal Sound Recordings, Avisoft Bioacoustics*. www.avisoft.com, 2006.