# Speech - Nonspeech Discrimination using the Information Bottleneck Method and Spectro-Temporal Modulation Index

*Maria Markaki, Michael Wohlmayr, and Yannis Stylianou*

Computer Science Department, University of Crete, Greece
mmarkaki@csd.uoc.gr, micki_w@csd.uoc.gr, yannis@csd.uoc.gr

## Abstract

In this work, we adopt an information theoretic approach - the Information Bottleneck method - to extract the relevant spectro-temporal modulations for the task of speech / non-speech discrimination - non-speech events include music, noise and animal vocalizations. A compact representation (a "cluster prototype") is built for each class consisting of the maximally informative features with respect to the classification task. We assess the similarity of a sound to each representative cluster using the spectro-temporal modulation index (STMI) adapted to handle the contribution of different frequency bands. A simple threshold check is then used for discriminating speech from non-speech events. Conducted experiments have shown that the proposed method has low complexity and high accuracy of discrimination in low SNR conditions compared to recently proposed methods for the same task.

**Index Terms**: audio classification, speech discrimination, auditory model

## 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.

Speech is characterized by joint spectro-temporal energy modulations; oscillations in power across spectral and temporal axes in the spectrogram. Of particular relevance to speech intelligibility are the slow temporal modulations (few Hz) that correspond to the phonetic and syllabic rates of speech [1]. Spectrogram modulations at multiple resolutions give a highly redundant representation of sound. This might be an advantage in the presence of noise and uncertainty, provided that we select a reduced set of these features which still captures enough information about the recognition task.

Tishby et al. have proposed the *Information Bottleneck Method* (IB) for the automatic detection and selection of the task - relevant features of speech signals [2, 3]. The IB method enables to construct a compact representation for each class that maintains information about the target through clusters obtained with the IB procedure. In [3] a general speech-oriented implementation of IB has been presented, using Mel frequency cepstral coefficients (MFCC). According to the recognition task, phoneme or speaker recognition, a small subset of MFCCs was selected which preserved high mutual information about the target [3].

In this work we estimate the power distribution in the modulation spectrum of speech signals and other sounds at different frequency ranges. The auditory model of Shamma et al [4] is the basis for these estimations. The model has been successfully applied in the assessment of speech intelligibility [5] and the discrimination of speech from non-speech [6], among others. Through the sequential information bottleneck procedure (sIB), we obtain a "cluster prototype" for each class consisting of the modulation frequencies that differ most between speech and other sounds. We assess the similarity of sounds to speech cluster representative using the spectro-temporal modulation index (STMI) [5] extended to handle the contribution of different frequencies. A simple threshold check is used for discriminating between speech and non-speech events. The system is compared to the system in [6] which is based on the same auditory features but uses a multilinear dimensionality reduction technique - Higher Order Singular Value Decomposition (HOSVD) [7] - and Support Vector Machines (SVMs) for classification. We evaluate systems performance in voice activity detection under varying noise conditions, using F-measure. In low levels of additive noise, our system is almost equivalent to the system of [6], whereas in low SNR conditions the proposed method exhibits superior performance. Moreover, when we take into account the similarity of audio signals to both cluster "prototypes" (speech and non-speech), we achieve a remarkable improvement in accuracy under severe noisy conditions (SNR$\leq$ 0dB). For evaluation purposes, we have also implemented another segment-based system based on MFCCs and Zero Crossing Rates (ZCRs); these features are also extracted on a frame basis and their mean values in each segment are given as input to an SVM classifier. This system is used as a reference system to show the robustness of the auditory features to various noise conditions.

The paper is organized as follows. 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 modified STMI. In Section 4 we compare the performance of the proposed system, the system in [6] and the reference system (MFCCs and ZCRs) on voice activity detection using F-measure at various SNR conditions. Conclusions are provided in Section 5.

## 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 filters have the form of a spectro-temporal Gabor function, selective for specific frequency sweeps, bandwidth, etc., performing actually a multi-resolution 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 a fixed time window of 500 ms. The dimensionality of the 3-dimensional arrays, or third-order tensors that incur, covers 128 logarithmic frequency bands × 5 scales × 12 rates.

# 3. Information Bottleneck, IB, 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 [3]. IB method provides an information theoretic formulation and solution to the tradeoff between compactness and quality of a signal's representation [2, 8, 3]. In the supervised learning framework, features are regarded as relevant if they provide information about a target. In the case of speech processing systems, the tagging $Y$ of the audio signal (as speech / non speech, identity of speakers or phonemes) guides 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 $Y$. Given an estimate of their joint distribution $p(x, y)$, the amount of relevant information in $X$ about $Y$ can be measured through Shannon's mutual information :

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

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 the compressed representation of $x$; $x$ is transformed to $\tilde{x}$ by a (stochastic) mapping $p(\tilde{x}|x)$. We seek an $\tilde{X}$ that compresses $X$ through minimization of $I(\tilde{X}; X)$ *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 as minimization of the *IB variational functional*:

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

where $\beta$ is the (positive) Lagrange multiplier controlling the tradeoff between compression and relevance. Various iterative algorithms have been proposed that converge to a reduced representation $\tilde{X}$ given $p(x, y)$ and $\beta$. We have chosen the *sequential optimization algorithm* (sIB) [8] since we want a fixed number of hard clusters as output.

The input consists of the joint distribution $p(x, y)$, the tradeoff parameter $\beta$ and the number of clusters $M = |\tilde{X}|$. During initialization, each element $x \in X$ is randomly assigned to one of the $M$ clusters $\tilde{X}$. Afterwards, the algorithm applies sequential update steps where it cycles through all $x \in X$ and tries to assign them to a different cluster $\tilde{x} \in \tilde{X}$ in order to minimize the loss of mutual information on the relevant variable $Y$. Optimization is controlled through the Jensen-Shannon divergence, i.e., the likelihood that $\{x\}$ and the cluster $\tilde{x}$ currently being merged have a common source [8]. Each update step increases the value of the functional

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

which is equivalent to minimization of $\mathscr{L}\{p(\tilde{x}|x)\}$ in (2). Following every single assignment update, both distributions $p(y|\tilde{x})$ and $p(\tilde{x})$ are updated. This detail makes sIB to be related somehow to the incremental variant of the Expectation Maximization (EM) algorithm for maximum likelihood (see [8] for details).

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 reduce sensitivity to locally optimal solutions, we repeat the whole process with different initial random partitions [8].

### 3.1. Application to Cortical Features

The feature tensor $\mathcal{Z}$ represents a discrete set of *continuous* features $z_{i_1,i_2,i_3} = \mathcal{Z}_{i_1,i_2,i_3} \in \mathbb{R}^{+S \times R \times F}$. Let the location of a response be denoted by $x_i$, where $i = 1, \ldots, S \times R \times F$, such that $z_{i_1,i_2,i_3} = z_{x_i}$. The 3−dimensional modulation spectrum ( scale - rate - frequency) is divided then into $S \times R \times F$ bins centered at $(\Omega_{i_1}, \omega_{i_2}, f_{i_3})$. Given a training list of $N$ feature tensors $Z^{(k)}$ $k = 1, \ldots, N$, we can sort the $S \times R \times F = 7680$ bins by unsupervised "information gain" (i.e., with respect to their frequency of occurrence in all N training samples) and keep the top 2000 bins. Since we also have their corresponding targets $y^{(k)}$, $k = 1, \ldots, N$, $y = 1, 2$ (nonspeech and speech tags, respectively), we can build a count matrix $K(x, y)$ where $K(i, j)$ denotes the frequency of occurrence of the $x_i$ discrete subdivision of the modulation spectrum in the presence of the $y_j$ target value. Normalization of this count matrix yields an estimate of the joint distribution $p(x, y)$, which is all the IB framework requires.

We have clustered the features $X$ into 5 groups, one composed of features most relevant to $y_1$, the second of features relevant to $y_2$, whereas the other clusters consist of features less relevant to either class. Since this setting implies a significant degree of compression, we can ignore the trade-off parameter $\beta$ setting $\beta^{-1} = 0$ in (3) and concentrate on solutions that maximize the relevant information term only. Let $\tilde{X}$ denote a compressed representation (a reduced feature set) and $p(\tilde{x}|x)$ the (deterministic) mapping obtained by sIB algorithm. We "discard" the clusters $\tilde{X}_j$ whose contribution :

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

to $I(\tilde{X}; Y)$ is minimal. To find out the identity of the remaining clusters, we compute:

$$p(\tilde{x}, y) = \sum_{x} p(x, y) p(\tilde{x}|x) \qquad (5)$$

$$p(\tilde{x}) = \sum_{y} p(\tilde{x}, y) \qquad (6)$$

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

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:

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

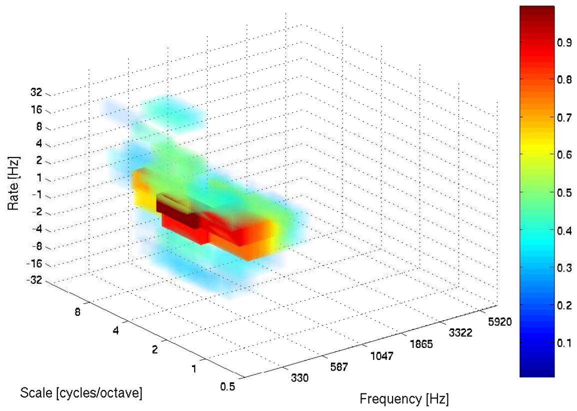


Figure 1: $p(x|\tilde{x} = \tilde{x}_2)$ for speech.

Cluster $\tilde{X}_1$ holds 3.14% and $\tilde{X}_2$ holds 6.47% of the 7680 bins. The remaining 90.39% of elements are considered irrelevant. Therefore during testing we don't need 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. Figure 1 presents an example of the relevant modulation spectrum of speech sounds. Strongest speech-relevant modulations appear between $2 - 4$ cyc/octave (scale), $-1$ and 2 Hz (rate), and inside the $300 - 600$ Hz frequency range.

The speech-relevant "cluster prototype" $p(x|\tilde{x} = \tilde{x}_2)$ permits the classification of audio signals based on their similarity to that cluster and a threshold check. We can assess the similarity between the cortical-like representations of sounds $\mathcal{Z}(\Omega, \omega, f)$ and $p(x|\tilde{x} = \tilde{x}_2)$ using the spectro-temporal modulation index (STMI) [5], defined below between corresponding $\Omega$, $\omega$ *and* $f$ channels:

$$\rho_s(\Omega, \omega, f) = \sqrt{\frac{1}{1 + \left(\frac{\mathcal{Z}(\Omega,\omega,f) - p(x|\tilde{x}=\tilde{x}_2)}{\sigma_{\mathcal{Z}}(\Omega,\omega,f)}\right)^2}} \qquad (9)$$

where $\sigma_{\mathcal{Z}}(\Omega, \omega, f)$ is the standard deviation of the auditory representation extracted over a fixed time-frame (500 ms) at each channel, whereas $\mathcal{Z}(\Omega, \omega, f)$ is the corresponding mean value.

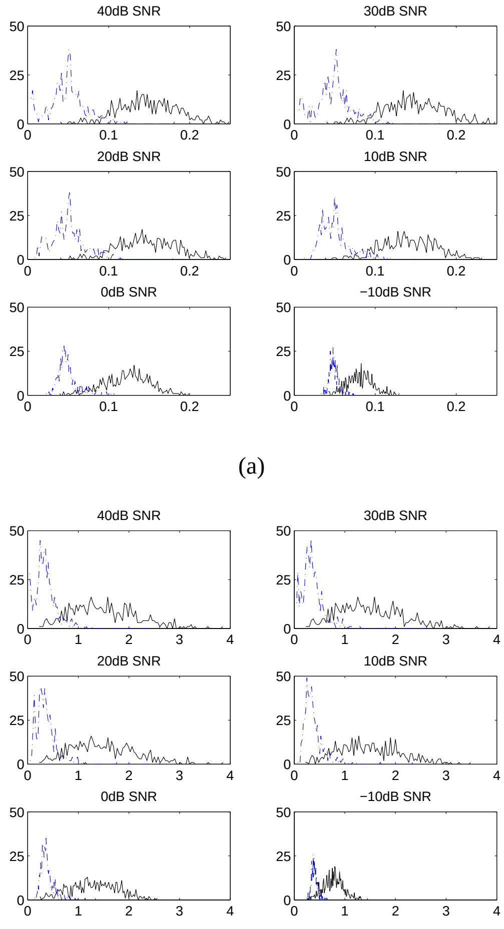
Finally we derive the average of $\rho_s(\Omega, \omega, f)$ over the channels $(\Omega, \omega, f) \in \tilde{X}_2$, i.e., over which $p(x|\tilde{x} = \tilde{x}_2) \neq 0$ :

$$\rho(\mathcal{Z}) = \frac{1}{|\tilde{X}_2|} \sum_{(\Omega,\omega,f)\in\tilde{X}_2} \rho_s(\Omega, \omega, f). \qquad (10)$$

We also compare the similarity of a sound to both typical patterns $p(x|\tilde{x} = \tilde{x}_1)$ and $p(x|\tilde{x} = \tilde{x}_2)$, $\rho_1$ and $\rho_2$ respectively, by taking their ratio:

$$R(\mathcal{Z}) = \frac{\rho_2}{\rho_1} \qquad (11)$$

We calculate the STMI ($\rho$) and corresponding ratio ($R$) for all training examples and noise conditions. Figure 2 shows the histograms of $\rho$ and $R$ computed on speech (continuous curve) and non-speech examples (dashed curve). The histograms form two distinct clusters with a small degree of overlap in the case of $\rho$, whereas decision threshold depends on the SNR condition especially for low SNR (0dB, -10dB). In the case of $R$ distribution the overlap is increased, however the decision threshold is less sensitive to the variation of SNR. This trend is reflected in the results in the benchmark test presented below.



(a)

(b)

Figure 2: *Histogram of STMIs $\rho$ (a) and ratios of STMIs $R$ (b) computed on nonspeech (dashed) and speech examples (continuous curve).*

Threshold setting implies a trade-off between the false acceptance rate ($FAR$) and false rejection rate ($FRR$) for each class. In this case, we have set threshold at a fixed value $\theta$ that minimizes the total number of segments incorrectly assigned to each class at the highest SNR level (40dB).

### 3.2. 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 [9]. The noise examples 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 segments of 500 ms each.

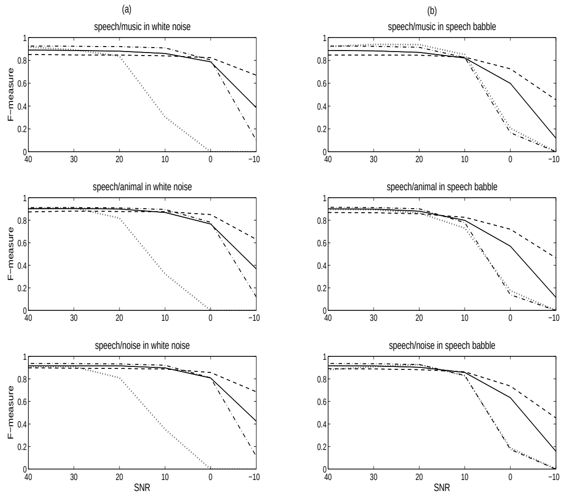


Figure 3: F measure for System 1 (dotted-dashed line), System 2 (continuous line), System 2(a): (dashed line), and System 3: (dotted line) applied to the benchmark test (a) with additive white noise (b) with speech babble.

From each of these segments, a feature tensor $\mathcal{Z}$ 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. Classification is based on STMI and a decision threshold, whereas **System 2(a)** uses the ratio of STMIs. In **System 3** mean values of MFCC and ZCR features are extracted and classification is based on SVM.

Test set consists of 260 speech and 300 non-speech examples of varying length. 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*: each signal is 30 seconds long, consisting of alternating speech - nonspeech test examples with random lengths (between 2 and 8 seconds). We create 10 such signals for each non-speech class: music, noise, or animal vocalization events, 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. Features are collected in a fixed time-window of 500 ms and at a rate of 50 ms. Every 50 ms frame is classified as speech (or non-speech) if it lies in the middle of a segment that was classified as speech (or non speech, respectively).

## 4. Experimental Results

We evaluate systems performance in terms of the F-measure for each non-speech class (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 and recall. Both measures are evaluated here with respect to the speech class. The results are presented in Figure 3.

Both systems 1 and 2 - which are based on the same auditory features - exhibit equally good performance in high SNR conditions. For white noise the proposed method presents a better generalization ability below 0 dB, whereas for speech babble the results are better even below 10 dB. The system based on ratio of STMIs is somewhat inferior in high SNRs but is clearly better in extremely noisy conditions (see also Fig. 2). 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 similar generalization ability to system 1 in the case of additive speech babble.

## 5. Conclusions

We presented an information theoretic approach to select a reduced set of auditory features which are maximally informative with respect to the target - speech / nonspeech. A simple threshold check built upon these reduced representations yields a performance close to, or better than state-of-the-art classifiers, with a significantly reduced computational load. It would be interesting to apply the proposed method to the task of recognition of other speech attributes, such as speech or speaker recognition, and language recognition.

## 6. Acknowledgements

This work was partially funded by GSRT grant 05AKMON106 and SIMILAR Network of Excellence.

## 7. References

[1] Quatieri, T.F., Discrete-Time Speech Signal Processing, Prentice-Hall Signal Processing series, 2002.

[2] 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.

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

[4] K. Wang and S.A. Shamma, ”Spectral shape analysis in the central auditory system”, IEEE Trans. Speech and Audio Proc., 3: 382–396, 1995.

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

[6] Mesgarani, N., Slaney, M., and Shamma S.A., ”Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations”, IEEE Trans. Audio, Speech and Language Proc., 14:920–930, 2006.

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

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

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