Selection of Relevant Features for Audio Classification tasks

Maria Markaki

Computer Science Department
University of Crete

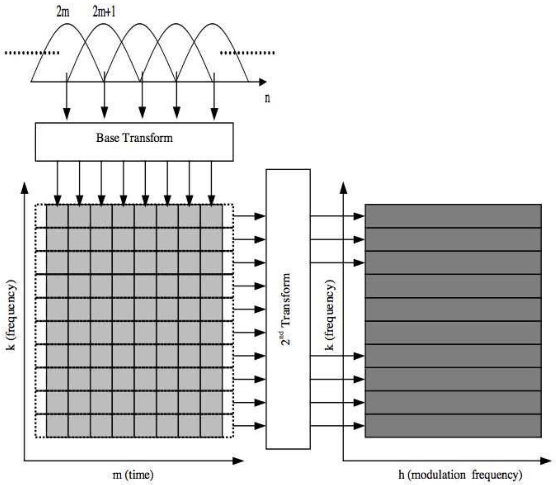
21 October 2011

Outline

Non-stationary Signal Analysis

- The analysis of human speech was the main reason for the development in the 1940s of time-frequency analysis
  - Time-frequency representations depict simultaneous measurements of the acoustic energy in both time and frequency domains
  - The main method was - and still is - the short-time Fourier transform whose the squared magnitude is the spectrogram
- Similar to a Fourier analyser, our auditory system maps the one-dimensional sound waveform to a time-frequency representation through the cochlea
- During later auditory stages, spectrum analysis occurs: fast and slow modulation patterns are detected by arrays of filters centred at different frequencies

Principle of Modulation Spectra



In Equations

- Short-time Fourier transform:

$$X_k(m) = \sum_{n=-\infty}^{\infty} h(mM - n)x(n)W_K^{kn},$$

where $k = 0, \ldots, K-1$, $W_K = e^{-j(2\pi/K)}$, $h(n)$: acoustic frequency analysis window.

- Subband envelope detection & frequency analysis:

$$X_l(k, i) = \sum_{m=-\infty}^{\infty} g(lL - m)|X_k(m)|W_l^{im},$$

where $i = 0, \ldots, I-1$, $g(m)$: modulation frequency analysis window [1]

EXAMPLE

- Joint acoustic / modulation frequency representations and their combination with cepstrum represent a simple interpretation of the computational auditory model [1].

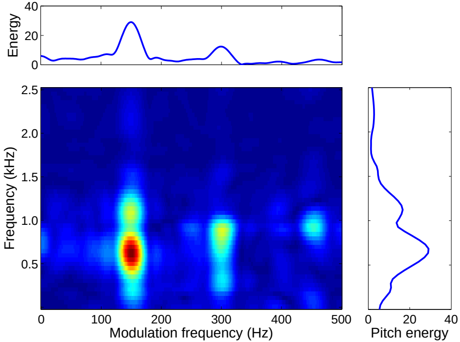


FIGURE: Modulation spectrogram of sustained vowel /AH/ by a normal speaker. The two side plots present the slices intersecting at the point of maximum energy; its coordinates coincide with the fundamental frequency and the first formant of /AH/ (~ 590 Hz).

Parameters

- Tapered windows $h(n)$ and $g(m)$:
  - reduced sidelobes of frequency estimates
- Length of the analysis window $h(n)$:
  - trade-off between resolution in the acoustic and modulation frequency axes
- Overlap between successive windows :
  - upper limit of the subband sampling rate during modulation transform
- Modulation spectral energy in the joint acoustic / modulation frequency plane:
  - a 2D-matrix $|X_l(k,i)| \in R^{K \times I}$
  - N training matrices: a 3D-tensor $\mathcal{A} \in R^{K \times I \times N}$

Outline

Curse of Dimensionality

- Classification algorithms detect and exploit complex patterns in data during training, validation and testing
- High dimensional features pose challenging problems to learning algorithms:
  - high computational cost and storage volumes for the representation of signals
  - difficult exclusion of accidental, unstable patterns which lead to over-fitting of the training system:
    - the generalization error, and
    - the number of training examples required for achieving a given error level

  both increase with data dimension
- In order to obtain a low-dimensional representation of the signals suitable for classification, we can employ:
  - feature selection techniques
  - dimensionality reduction

Maximal Statistical Dependency

- Minimal classification error $\simeq$ maximal statistical dependency of target class c on the data distribution
- Max-Dependency criterion $\rightarrow$ a set S of m features $\{x_i\}$ which jointly have the largest dependency on the target class $\max D(S, c)$
  - statistical dependency of variables is measured by mutual information (MI):

$$D(S, c) = I(\{x_i, i = 1, \ldots, m\}; c) =$$

$$\int \ldots \int p(x_1, \ldots, x_m, c) \log \frac{p(x_1, \ldots, x_m, c)}{p(x_1, \ldots, x_m) p(c)} dx_1 \ldots dx_m dc$$

  - requires multivariate densities for MI estimation - hard to implement

Feature Selection based on MI

- Shannon's MI between two variables measures the amount of relevant and redundant information :
  - in the supervised learning framework, feature x_j is regarded as relevant if it provides information about a target c
  - redundancy between features x_j and x_i is defined as the amount of information variable x_j holds about variable x_i
- Max-Relevance criterion:

$$\max D(S, c), \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c)$$

  - features selected might depend on each other

$\Rightarrow$ add a minimal redundancy condition [2]:

$$\min R(S), \quad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j)$$

Max-Relevance-Min-Redundancy Criterion (mRMR)

- Incremental algorithm: selects the m^{th} feature from the set $\{X - S_{m-1}\}$ of m features:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right]$$

  - low computational complexity of incremental search method
  - equivalent to Max-Dependency for first order incremental feature selection [2]

- Still, heuristics are necessary during training for discovering the optimal relation between relevance and redundancy

Idea : reduce features redundancy first so that multivariate probability densities almost equal the product of marginal densities

Higher Order SVD

- Higher Order Singular Value Decomposition (HOSVD) is a generalization of SVD to tensors [3]
  - SVD first proposed for the Wigner distribution
- Real signals contain noise spread out over all the terms of the decomposition, whereas signals are well represented by the first few terms
  - Truncation of the series after the first few terms, significantly reduces noise while retaining most of the signal
  - The signal representations can be approximated in a lower-dimensional space producing a compact feature set suitable for classification

3^{rd}-order Singular Value Decomposition

- A generalization of SVD to tensors :

$$\mathcal{A} = \mathcal{S} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}$$

where:

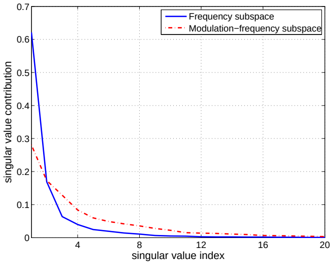
- $U^{(n)} = [U_1^{(n)}, \dots U_{I_n}^{(n)}]$, the matrix of left singular vectors of the matrix unfolding $A_{(n)}$
- $\mathcal{S} \in R^{(I_1 \times I_2 \times I_3)}$ has all-orthogonal subtensors with ordered Frobenius-norms:

$$\|\mathcal{S}_{i_n=1}\| \geq \|\mathcal{S}_{i_n=2}\| \geq \dots \geq \|\mathcal{S}_{i_n=I_n}\| \geq 0$$

- $\|\mathcal{S}_{i_n=i}\| \equiv \sigma_i^{(n)}$ are $n-$mode singular values of $\mathcal{A} \equiv$ singular values of the matrix unfolding $A_{(n)}$

"Rank" of the Matrix Unfolding

- Ordering of $n-$mode singular values $\sigma_{i_n}^{(n)}$ implies that the "energy" of tensor $\mathcal{A}$ is concentrated in the singular vectors $U_i^{(n)}$ with the lowest values of i in every subspace
- Based on the data accuracy, we define a threshold τ and retain the singular vectors with $\sigma_{i_n}^{(n)}$ exceeding it



Maximum Contribution Criterion

- Dimensionality of the embedding can be selected through traditional model selection methods such as cross-validation
- Dimensionality reduction can preserve information from all the original input variables, promoting generalization
  - still, purely unsupervised techniques might throw away low variance dimensions which are highly predictive for a classification task
- Goal: to combine both unsupervised and supervised techniques to gain the benefit of both approaches

Optimal "Independent" Features

- Project $|X_l(k,i)|$ to the basis vectors contributing more than τ to the "energy" of each subspace
  - $U_i^{(1)}$, $i = 1, \ldots, i_1$ in the acoustic frequency space
  - $U_i^{(2)}$, $i = 1, \ldots, i_2$ in the modulation frequency space
- Select the most relevant "independent" features

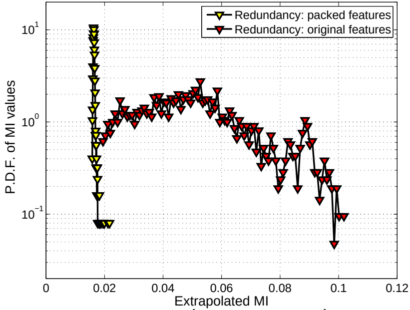


FIGURE: Redundancy of original (red triangles) and "independent" features, after applying HOSVD (yellow triangles).

Maximum Relevance Criterion applied after HOSVD

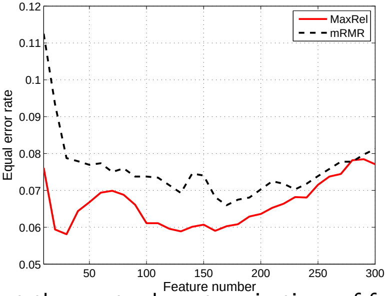


Figure: We select the most relevant projections of features among those contributing more than a threshold, through cross-validation procedure. SVM classifier equal error rate using mRMR and MaxRel features for speech/nonspeech discrimination on broadcast news.

Outline

Speech Discrimination based on Modulation Spectra

- The discrimination of speech and non-speech is the first processing step before speaker segmentation and recognition, or speech transcription
- We design a content based speech discrimination algorithm which exploits long-term information inherent in modulation spectrum
  - the system is built upon a segment based SVM classifier
- Detection experiments on Greek and U.S. English broadcast news data, suggest that the system provides complementary information to state-of-the-art mel-cepstral features

Relevance of Features

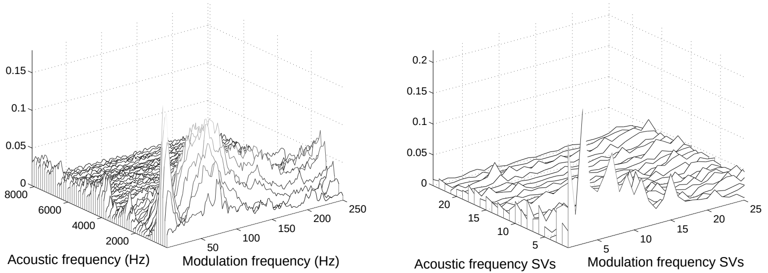


FIGURE: Relevance of the original and compressed modulation spectral features: Mutual information (MI) between the speech / non-speech class variable and (left) the acoustic and modulation frequencies (65 × 125 dimensions) and (right) the first 25 singular vectors in each subspace.

Maximum Relevance vs Maximum Contribution Criterion

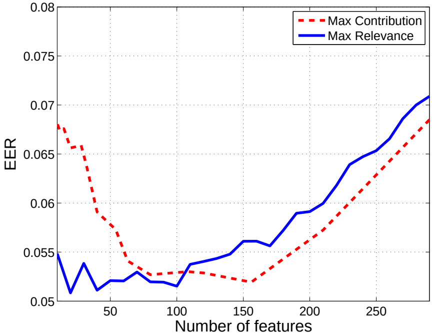


FIGURE: SVM classifier equal error rate (EER) as a function of number of features selected in terms of maximum relevance or maximum contribution.

Approximate Representations with Optimal Performance

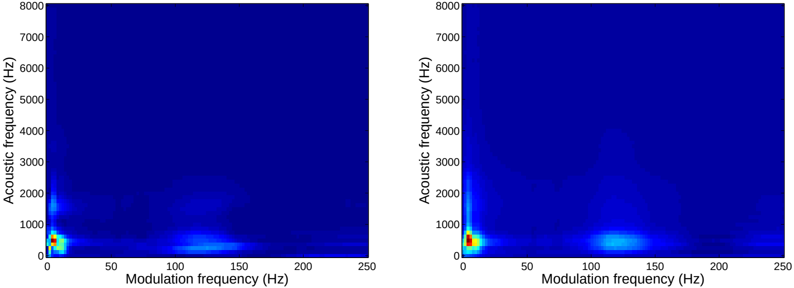


FIGURE: (Left) Rank$-(13,12)$ approximation of modulation spectrum for 500 ms of a speech signal. (Right) 21 features approximation for the same speech signal. Energy at modulations corresponding to pitch (~ 120 Hz) and syllabic and phonetic rates (< 40 Hz) remain prominent.

Outline

1. Feature Extraction from Sound Signals
   - Modulation Frequency Analysis
2. Feature Selection for Classification
   - Feature Selection based on MI
   - Redundancy Reduction using HOSVD
3. Speech Discrimination on Broadcast news
4. Pathological Voice Quality Assessment
5. Systolic Heart Murmur Classification

Pathological Voice Quality Assessment

- Objectively evaluate the degree of voice alterations in a non-invasive manner, using acoustic analysis
  - assist the perceptual evaluation of dysphonic voice quality used by the clinicians
- Identify acoustic measures that highly correlate with pathological voice qualities
- Modulation frequency analysis for voice pathology detection and classification

Relevant Features without Normalization

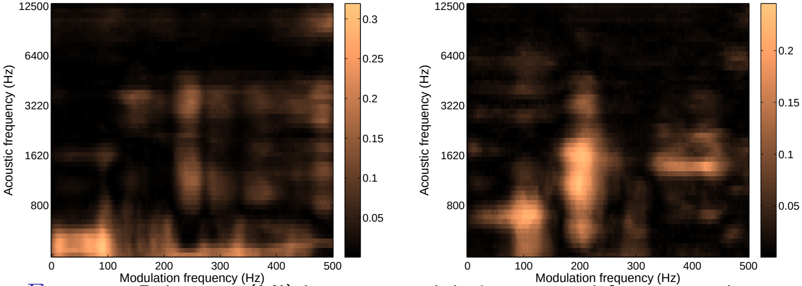


FIGURE: Relevance (MI) between modulation spectral features and pathologic voice class *without normalization* in MEEI (left), and in PdA (Right).

Normalization of modulation spectra

- The distribution of envelope amplitudes of voiced speech has a strong exponential component
  - we calculate modulation spectra using a log transformation of the amplitude values $|X_k(m)|$ and subtracting their mean log amplitude before windowing :

$$\hat{X}_k(m) = \log|X_k(m)| - \overline{\log|X_k(m)|} \qquad (1)$$

    where $\overline{\log|X_k(m)|}$ denotes the average of $\log|X_k(m)|$ over m
  - analogous to the cepstral mean subtraction approach, which compensates for convolutional noise in MFCC features
- Next, we normalize every acoustic frequency subband with the marginal of the modulation frequency representation (Sukittanon et al 2004):

$$X_{l,sub}(k,i) = \frac{X_l(k,i)}{\sum_i X_l(k,i)} \qquad (2)$$

Relevant Features after Normalization

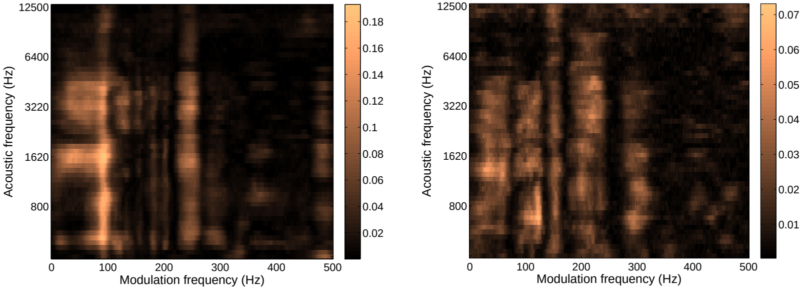


Figure: Relevance (MI) between modulation spectral features and pathologic voice class *after normalization* in MEEI (left), and in PdA (right).

Performance of MFCC and mRMS features in MEEI

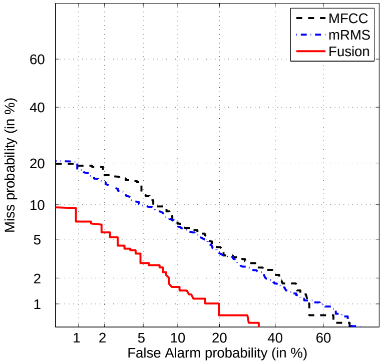


FIGURE: Detection Error Trade-off (DET) curve using mRMS features, MFCC and their fusion (concatenation of feature vectors) in MEEI.

Performance of MFCC and mRMS features in PdA

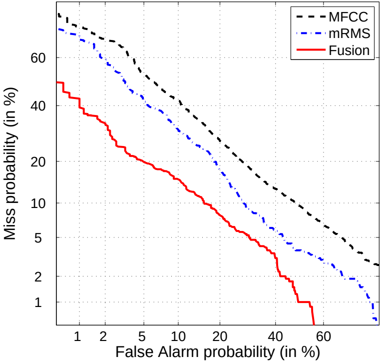


Figure: DET curve using mRMS features, MFCC and the concatenated feature vector in PdA.

Cross-database performance of MFCC and mRMS features

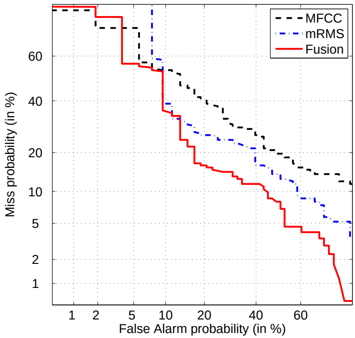


Figure: DET curve using mRMS features, MFCC and the concatenated feature vector when training is performed in PdA and testing in MEEI.

Results: Classification of Pathologies in MEEI

Classify: *vocal fold polyp, adductor spasmodic dysphonia, keratosis leukoplakia, and vocal nodules*

	mRMS			FD-GA
	DCF_{opt} (%)	AUC (%)	m	DR (%)
Pol/Add	88.33 ± 2.64	95.74	60	82.5
Pol/Ker	86.11 ± 5.52	93.61	80	81.8
Pol/Mod	91.25 ± 3.13	95.03	20	87.5

where: FD-GA stands for *Fisher distance and Genetic Algorithms* (Hosseini et al. 2008)

Outline

Systolic Heart Murmur Classification

- Classic heart auscultation using a stethoscope
  - the most common method to screen the health of cardiovascular system
  - simple, fast, with minimal cost
- Detection of pathological heart sounds - murmurs or additional sounds
  - indication of structural abnormalities of the cardiovascular system
- A significant percentage of children presents some innocent functional murmurs
  - accurate discrimination between pathological and innocent murmurs is a skill that can take years to acquire and refine

Automatic Preprocessing of PCG recordings

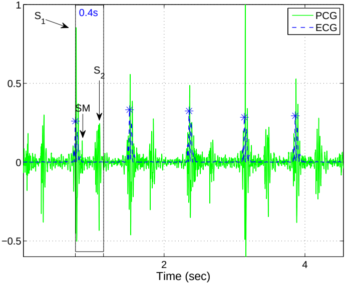


FIGURE: Phonocardiographic signal (solid) and envelope of electrocardiographic signal (dash) of an 10-years old with innocent early to midsystolic murmur. A 400ms segment at the beginning of a heart cycle is highlighted, including S_1, the systolic murmur (SM) and S_2.

Reassigned spectrogram

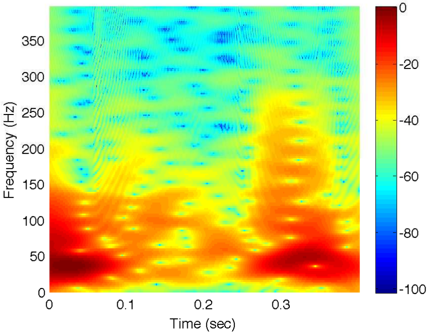


FIGURE: Energy (relative sound intensity in dB) of the reassigned spectrogram [4] of the PCG (shown in previous slide) with innocent early to midsystolic murmur - first 400 ms of one heart cycle.

Children PCG Database

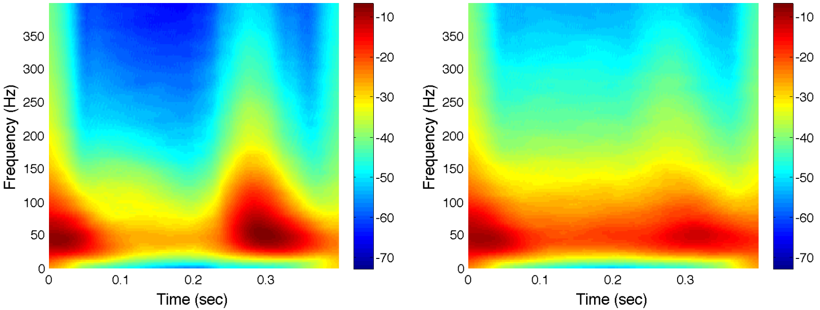


Figure: Mean values for the energy (relative sound intensity in dB) of the reassigned spectra of the PCG from 25 subjects with (left) innocent systolic murmurs, (right) pathological systolic murmurs - 3 recordings with 5 consequent heart cycles per recording.

Visualization of Useful Information

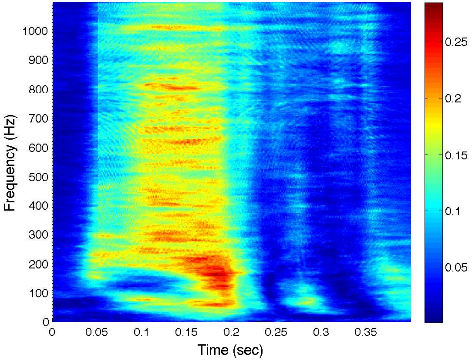


FIGURE: Relevance - estimated as mutual information - of the reassigned spectral features of the PCG (the first 400ms of the heart cycle) for discrimination of abnormal murmurs.

System Performance

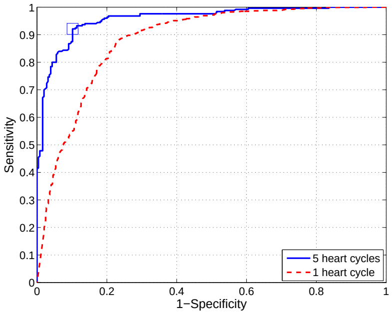


Figure: Average ROC curves of 25 cross-validation runs using SVM based on one heart cycle (red dashed) or five heart cycles segments (blue solid line). The best classification score for one recording corresponds to a sensitivity of 92.11% and a specificity of 89.82% (blue square).

Comparison of System Performance to General Doctors

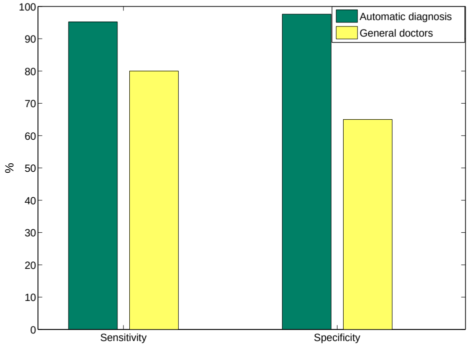


Figure: Sensitivity and specificity of the system (green bars) compared to general doctors (yellow bars).

Summary of Contributions

- Adaptation of the maximum dependency criterion for feature selection in two steps:
  1. redundancy reduction through HOSVD
  2. selection of the most relevant independent features through cross-validation
- Application of Max-Dep criterion to speech discrimination and pathological voice quality assessment based on modulation spectra
- Application of Max-Dep criterion to heart murmur classification based on reassigned spectra
- Normalization of modulation frequency features for cross-database experiments on voice quality assessment

Future Work

- Apply the algorithm using more elaborate representations for various signal classification tasks
- Experiment with recent feature selection techniques, e.g., based on Markov Blanket theory
- Comparison of the heart murmur classification system to a state-of-the-art method on the same data
- Classification of a sequence of spectra, as in video, adding an extra dimension of time before HOSVD

L. Atlas and S.A. Shamma.

Joint acoustic and modulation frequency.

EURASIP Journal on Applied Signal Processing, 7:668–675, 2003.

H. Peng, F. Long, and C. Ding.

Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.

IEEE Trans. Pattern Anal. Mach. Intell., 27:1226–1238, 2005.

L. De Lathauwer, B. De Moor, and J. Vandewalle.

A multilinear singular value decomposition.

SIAM J. Matrix Anal. Appl., 21:1253–1278, 2000.

F. Auger and P. Flandrin.

Improving the readability of time-frequency and time-scale representations by the reassignment method.

IEEE Trans. Signal Process., 43(5):1068–1089, 1995.

Conference Publications

1. “Speech - Nonspeech Discrimination using the Information Bottleneck Method and Spectro-Temporal Modulation Index”, Markaki M., Wohlmayr M. and Stylianou Y., InterSpeech ICSLP, 2007
2. “Discrimination of Speech from nonspeech in broadcast news based on modulation frequency features”, Markaki M. and Stylianou Y., ISCA, 2008
3. “Dimensionality Reduction of Modulation Frequency Features for Speech Discrimination”, Markaki M. and Stylianou Y., InterSpeech, 2008
4. “Singing Voice Detection using Modulation Frequency Features”, Markaki M., Holzapfel A. and Stylianou Y., ISCA, 2008
5. “Evaluation of Modulation Frequency Features for Speaker Verification and Identification”, Markaki M. and Stylianou Y., EUSIPCO, 2009

Conference Publications

6. "Using Modulation Spectra for Voice Pathology Detection and Classification", Markaki M. and Stylianou Y., IEEE EMBC, 2009
7. "Normalized Modulation Spectral Features for Cross-Database Voice Pathology Detection", Markaki M. and Stylianou Y., InterSpeech, 2009
8. "Modulation Spectral Features for Objective Voice Quality Assessment: the Breathiness case", Markaki M. and Stylianou Y., MAVEBA, 2009
9. "Modulation Spectral Features for Objective Voice Quality Assessment", Markaki M. and Stylianou Y., IEEE ISCCSP, 2010
10. "Dysphonia Detection based on Modulation Spectral Features and Cepstral Coefficients", Markaki M., Stylianou Y., Arias-Londono J.D. and Godino-Llorente J.I., ICASSP, 2010

Journal & Book Publications

1. "Extraction of Speech-Relevant Information from Modulation Spectrograms", Markaki M., Wohlmayr M. and Stylianou Y., Progress in Nonlinear Speech Processing, Springer, pp. 78 - 88, 2007
2. "Discrimination of Speech from Nonspeeech in Broadcast News Based on Modulation Frequency Features", Markaki M. and Stylianou Y., Speech Communication, 2010
3. "On combining information from Modulation Spectra and Mel-Frequency Cepstral Coefficients for automatic detection of pathological voices", Arias-Londono J.D., Godino-Llorente J.I., Markaki M. and Stylianou Y., Logopedics Phoniatrics Vocology, 2010
4. "Voice Pathology Detection and Discrimination Based on Modulation Spectral Features", Markaki M. and Stylianou Y., IEEE Transactions on Speech and Audio Processing, 2011

THANK YOU

for your attention