Advances in Speech Signal Processing for Voice Quality Assessment

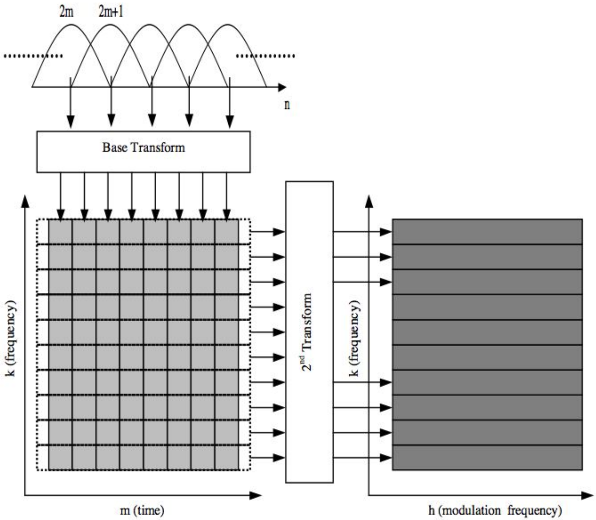
Part II

Yannis Stylianou

University of Crete, Computer Science Dept., Multimedia Informatics Lab
yannis@csd.uoc.gr

Bilbao, 2011 September

Principle


2m
2m+1
n
Base Transform
k (frequency)
m (time)
2nd Transform
k (frequency)
h (modulation frequency)

In equations

First Step: STFT

$$
\begin{aligned}
X_m(k) &= \sum_{n=-\infty}^{\infty} h(mM-n)x(n)W_{I_1}^{kn}, \\
k &= 0, \ldots, I_1 - 1,
\end{aligned}
$$

where:

- I_1 denotes the number of frequency bins in the acoustic frequency axis,
- $W_{I_1} = \exp(-j\pi/I_1)$,
- M is the shift parameter (or, hop size) in the computation of the STFT,
- $h(n)$ is the acoustic frequency analysis window.

In equations

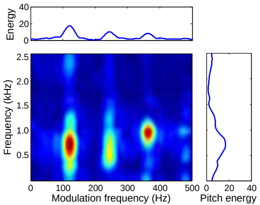
Second Step: Modulation frequencies estimation of the Subband Envelopes

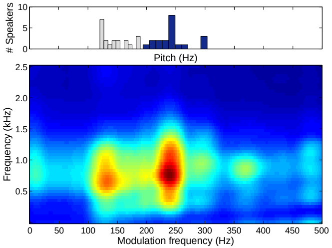
$$
\begin{aligned}
X_l(k,i) &= \sum_{m=-\infty}^{\infty} g(lL-m)|X_m(k)|W_{I_2}^{im}, \\
i &= 0,\ldots,I_2-1,
\end{aligned}
$$

where:

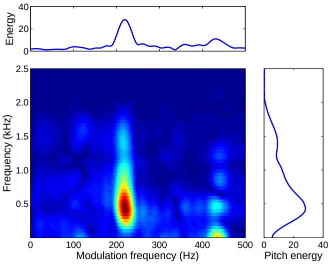
- I_2 is the number of frequency bins along the modulation frequency axis,
- $W_{I_2} = \exp\left(-j(f_M/F_s)\,\pi/I_2\right)$,
- f_M and F_s denoting the maximum modulation frequency we search for, and the sampling frequency, respectively,
- L is the shift parameter of the second STFT, and
- $g(m)$ is the modulation frequency analysis window.

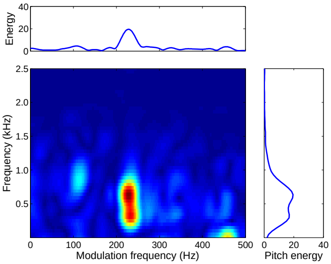
Example I: one speaker (left), mean of speakers (right)



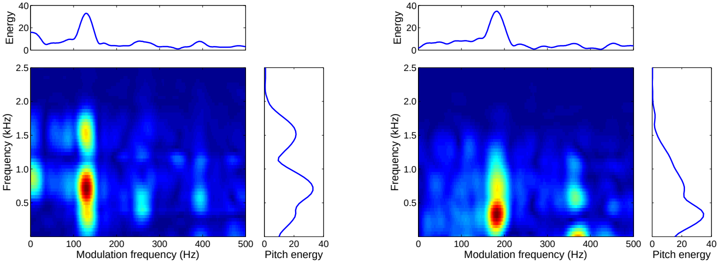


Example II: polyps (left), spasmodic dysphonia (right)





Example III: keratosis (left), nodules (right)



Dimensionality Reduction, HO-SVD

1. Create tensors: $\mathcal{D} \in R^{I_1 \times I_2 \times I_3}$
2. Decompose of tensor $\mathcal{D}$ to its $n-$mode singular vectors:

$$\mathcal{D} = \mathcal{S} \times_1 U_{af} \times_2 U_{mf} \times_3 U_{samples}$$

   where S and U are referred to as *core tensor* and *unitary matrix*, respectively and $\times_n$ denotes the $n-$mode product.
3. Rank the $n-$mode singular values
4. Near-optimal projections (PCs): truncate Singular Matrices so that we keep $a\%$ "energy" of $\mathcal{D}$

Dimensionality Reduction, HO-SVD

$n-$mode singular vectors:

- Let consider tensor $\mathcal{D} \in R^{I_1 \times I_2 \times I_3}$
- Unfold $\mathcal{D}$ to $\mathbf{D}_{(n)}$:
  1. $I_1 \times I_2 I_3$ matrix $\mathbf{D}_{(1)}$
  2. $I_2 \times I_3 I_1$ matrix $\mathbf{D}_{(2)}$
  3. $I_3 \times I_1 I_2$ matrix $\mathbf{D}_{(3)}$

  The n-mode singular values and vectors: SVD of $\mathbf{D}_{(n)}$.

Dimensionality Reduction, HO-SVD

Definition (Unitary matrix)

An $(I_n \times I_n)$ unitary matrix $\mathbf{U}^{(n)}$, $n = 1, 2, 3$, contains the n-mode singular vectors (SVs):

$$\mathbf{U}^{(n)} = \left[U_1^{(n)} \; U_2^{(n)} \; \ldots \; U_{I_n}^{(n)} \right]. \qquad (1)$$

Each matrix $\mathbf{U}^{(n)}$ can directly be obtained as the matrix of left singular vectors of the "matrix unfolding" $\mathbf{D}_{(n)}$ of $\mathcal{D}$ along the corresponding mode.

Dimensionality Reduction, HO-SVD

$$\mathcal{D} = \mathcal{S} \times_1 U_{af} \times_2 U_{mf} \times_3 U_{samples}$$

- S is referred to as core tensor (same dimensions as $\mathcal{D}$)
- $\mathbf{U}_{af} \in \mathbb{R}^{I_1 \times I_1}$, is the unitary matrix of the acoustic frequency subspace.
- $\mathbf{U}_{mf} \in \mathbb{R}^{I_2 \times I_2}$, is the unitary matrix of the modulation frequency subspace.
- $\mathbf{U}_{s} \in \mathbb{R}^{I_3 \times I_3}$ is the samples subspace matrix.
- $\times_n$ denotes $n-$mode product.

Dimensionality Reduction, HO-SVD

Defining n-product $\mathcal{S} \times_n \mathbf{U}^{(n)}$:

- $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$
- $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times I_n}$
- Example; for $n = 2$ this is an $(I_1 \times I_2 \times I_3)$ tensor given by

$$\left(\mathcal{S} \times_2 \mathbf{U}^{(2)}\right)_{i_1 i_2 i_3} \stackrel{\text{def}}{=} \sum_{i_2=1}^{I_2} s_{i_1 i_2 i_3} u_{i_2 i_2}.$$

Dimensionality Reduction, HO-SVD

1. Create tensors: $\mathcal{D} \in R^{I_1 \times I_2 \times I_3}$ and decompose it to its $n-$mode singular vectors:

$$\mathcal{D} = \mathcal{S} \times_1 U_{af} \times_2 U_{mf} \times_3 U_{samples}$$

2. Rank the $n-$mode singular values
3. Near-optimal projections (PCs): truncate Singular Matrices so that we keep $a\%$ "energy" of $\mathcal{D}$

Dimensionality Reduction, HO-SVD

- Contribution of the j^{th} n-mode singular vector $U_j^{(n)}$:

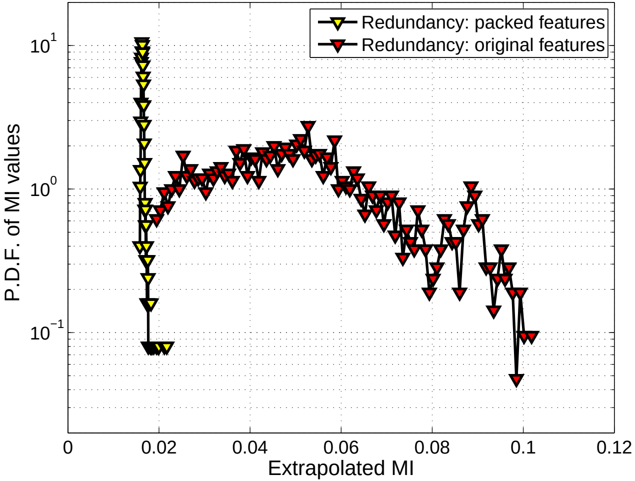
$$\alpha_{n,j} = \lambda_{n,j} \, / \sum_{j=1}^{I_n} \lambda_{n,j}$$

  where $\lambda_{n,j}$ is the corresponding singular value
- Put a threshold on $\alpha_{n,j}$ and retain the $R_n \, (n = 1, 2)$ singular vectors
- Truncate matrices: $\hat{\mathbf{U}}^{(1)} \equiv \hat{\mathbf{U}}_{af} \in \mathbb{R}^{I_1 \times R_1}$ and $\hat{\mathbf{U}}^{(2)} \equiv \hat{\mathbf{U}}_{mf} \in \mathbb{R}^{I_2 \times R_2}$
- Project new MS data on to the truncated matrices:

$$\mathbf{Z} = \hat{\mathbf{U}}_{af}^T \, \mathbf{B} \, \hat{\mathbf{U}}_{mf}$$

  where $\mathbf{B} \equiv |X_l(k,i)| \in \mathbb{R}^{I_1 \times I_2}$ and $\mathbf{Z} \in \mathbb{R}^{R_1 \times R_2}$

Redundancy Reduction with HOSVD



Mutual Information

Mutual Information between two random variables x_i and x_j is defined as:

$$I(x_i; x_j) = \int dx_i \int dx_j P_{ij}(x_i, x_j) \log_2 \left[\frac{P_{ij}(x_i, x_j)}{P_i(x_i) P_j(x_j)} \right]$$

where

- $P_{ij}(x_i, x_j)$ denotes the joint probability density function (pdf)
- $P_i(x_i)$ and $P_j(x_j)$ denote the marginal pdfs

Maximal Relevance Criterion

Select the most relevant features to the target class c:

1. Compute the mutual information $I(x_j; c)$ between feature x_j and class c
2. Rank all the computed $I(x_j; c)$
3. Select the top m features

Database & Conditions

- Sustained vowel /AH/ from MEEI
- Subset of the database (53 normophonic, 173 dysphonic speakers)
- Signals sampled at 25 kHz
- Classifier: SVM with a radial basis function (RBF) kernel
- 4-fold stratified cross-validation, repeated 400 times
- Training/Testing: 75%25%
- Decision per segment
- Evaluation: Detection Error Trade-off (DET) curves

Feature extraction

- Data tensor $\mathcal{D} \in \mathbb{R}^{257 \times 257 \times 600}$
- $\hat{U}_{af} \in \mathbb{R}^{257 \times 34}$
- $\hat{U}_{mf} \in \mathbb{R}^{257 \times 34}$
- $\mathbf{Z} \in \mathbb{R}^{34 \times 34}$

Results: Detection

Normophonic/Dysphonic:
Optimal detection accuracy (DCF_{opt}): 94.08% ($\pm$0.86) using the top $m = 25$ features ($AUC = 97.75\%$ in terms of ROC)

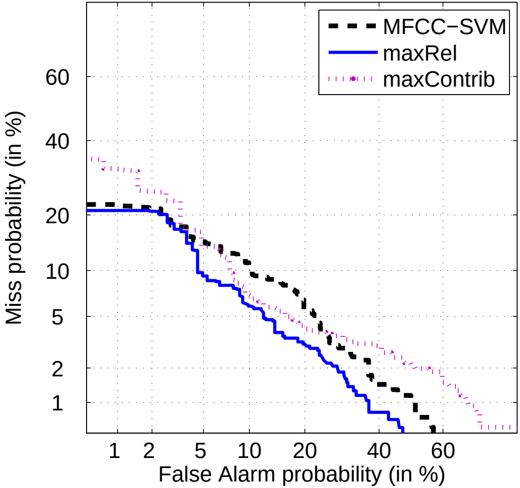
Results: Classification

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

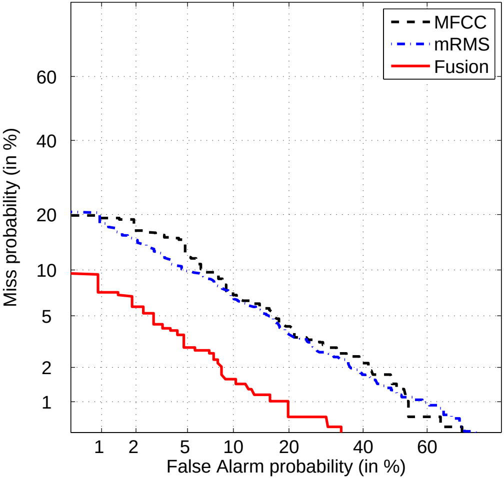
	MSMR			FD-GA
	DCF_{opt} (%)	AUC (%)	m	DR (%)
Pol/Add	88.33 ± 2.64	95.74	60	82.5
Pol/Ker	86.11 ± 5.52	93.61	80	81.8
Pol/Mod	91.25 ± 3.13	95.03	20	87.5

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

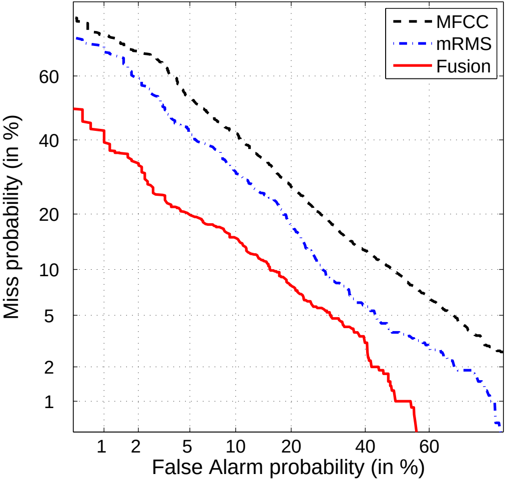
MEEI: COMPARISON



MEEI: FUSION


MFCC
mRMS
Fusion
Miss probability (in %)
60
40
20
10
5
2
1
1
2
5
10
20
40
60
False Alarm probability (in %)

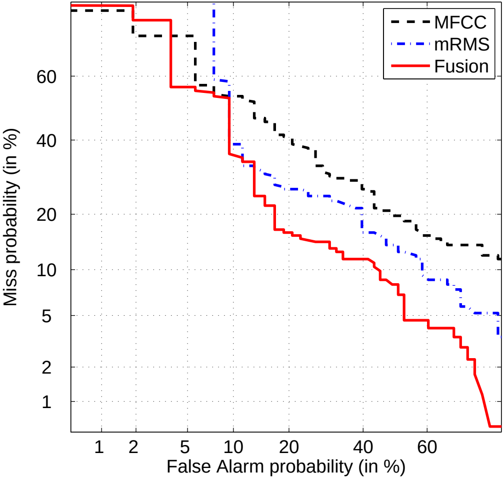
PdA: fusion



MFCC
mRMS
Fusion
Miss probability (in %)
60
40
20
10
5
2
1
False Alarm probability (in %)
1
2
5
10
20
40
60

Cross-database experiment

Train on PdA, test on MEEI



Cross-database experiment

	MFCC	MRMS	Fusion
MEEI	8.47	6.29 (125)	3.63
PdA	22.86	17.67 (125)	12.15
PdA-MEEI	28.24	24.40 (125)	16.87
MEEI-PdA	30.97	26.07 (450)	21.86

References for the work on Modulation Spectra

1. Maria Markaki and Yannis Stylianou:
Voice Pathology Detection and Discrimination based on Modulation Spectral Features.
IEEE Trans. on Audio, Speech and Language Processing. TASL.2010.2104141, Jan 2011
2. J.D. Arias-Londono, J.I. Godino-Llorente, M. Markaki, and Y. Stylianou:
On combining information from Modulation Spectra and Mel-Frequency Cepstral coefficients for Automatic Detection of Pathological Voices.
Logopedics, Phoniatrics, Vocology (LPV), Nov 2010.
3. Maria Markaki and Yannis Stylianou:
Discrimination of Speech from Nonspeeech in Broadcast News Based on Modulation Frequency Features
Speech Communication, Special Issue on Speech Communication on Perceptual and Statistical Audition,

Define Vocal Tremor

- Vocal Tremor: Involuntary modulations of frequency and/or amplitude in sustained phonation.
- Pathological & Physiological Vocal Tremor.
  - Pathological Tremor: From diseases like Parkinson, essential tremor, etc. $\Rightarrow$ Strong motor synchronization.
  - Physiological Tremor: Natural stochastic modulations in the interval $[2, 15]Hz$ with low amplitude.
- Acoustic Vocal Tremor Attributes:
  - Modulation Frequency: How fast are the modulations.
  - Modulation Level: How strong are the modulations.

Define Vocal Tremor

- Vocal Tremor: Involuntary modulations of frequency and/or amplitude in sustained phonation.
- Pathological & Physiological Vocal Tremor.
  - Pathological Tremor: From diseases like Parkinson, essential tremor, etc. $\Rightarrow$ Strong motor synchronization.
  - Physiological Tremor: Natural stochastic modulations in the interval $[2, 15] Hz$ with low amplitude.
- Acoustic Vocal Tremor Attributes:
  - Modulation Frequency: How fast are the modulations.
  - Modulation Level: How strong are the modulations.

Define Vocal Tremor

- Vocal Tremor: Involuntary modulations of frequency and/or amplitude in sustained phonation.
- Pathological & Physiological Vocal Tremor.
  - Pathological Tremor: From diseases like Parkinson, essential tremor, etc. $\Rightarrow$ Strong motor synchronization.
  - Physiological Tremor: Natural stochastic modulations in the interval $[2, 15]Hz$ with low amplitude.
- Acoustic Vocal Tremor Attributes:
  - Modulation Frequency: How fast are the modulations.
  - Modulation Level: How strong are the modulations.

Define Vocal Tremor

- Vocal Tremor: Involuntary modulations of frequency and/or amplitude in sustained phonation.
- Pathological & Physiological Vocal Tremor.
  - Pathological Tremor: From diseases like Parkinson, essential tremor, etc. ⇒ Strong motor synchronization.
  - Physiological Tremor: Natural stochastic modulations in the interval $[2, 15]Hz$ with low amplitude.
- Acoustic Vocal Tremor Attributes:
  - Modulation Frequency: How fast are the modulations.
  - Modulation Level: How strong are the modulations.

Define Vocal Tremor

- Vocal Tremor: Involuntary modulations of frequency and/or amplitude in sustained phonation.
- Pathological & Physiological Vocal Tremor.
  - Pathological Tremor: From diseases like Parkinson, essential tremor, etc. ⇒ Strong motor synchronization.
  - Physiological Tremor: Natural stochastic modulations in the interval $[2, 15] Hz$ with low amplitude.
- Acoustic Vocal Tremor Attributes:
  - Modulation Frequency: How fast are the modulations.
  - Modulation Level: How strong are the modulations.

Define Vocal Tremor

- Vocal Tremor: Involuntary modulations of frequency and/or amplitude in sustained phonation.
- Pathological & Physiological Vocal Tremor.
  - Pathological Tremor: From diseases like Parkinson, essential tremor, etc. $\Rightarrow$ Strong motor synchronization.
  - Physiological Tremor: Natural stochastic modulations in the interval $[2, 15]Hz$ with low amplitude.
- Acoustic Vocal Tremor Attributes:
  - Modulation Frequency: How fast are the modulations.
  - Modulation Level: How strong are the modulations.

Define Vocal Tremor

- Vocal Tremor: Involuntary modulations of frequency and/or amplitude in sustained phonation.
- Pathological & Physiological Vocal Tremor.
  - Pathological Tremor: From diseases like Parkinson, essential tremor, etc. $\Rightarrow$ Strong motor synchronization.
  - Physiological Tremor: Natural stochastic modulations in the interval $[2, 15]Hz$ with low amplitude.
- Acoustic Vocal Tremor Attributes:
  - Modulation Frequency: How fast are the modulations.
  - Modulation Level: How strong are the modulations.

Vocal Tremor Estimation

Use of an AM-FM decomposition algorithm based on the adaptive time-varying quasi-harmonic model for speech.

- High resolution in Time-Frequency plane.
- Estimation of Vocal Tremor for any sinusoidal component of speech.
- Time dependent Vocal Tremor estimations.

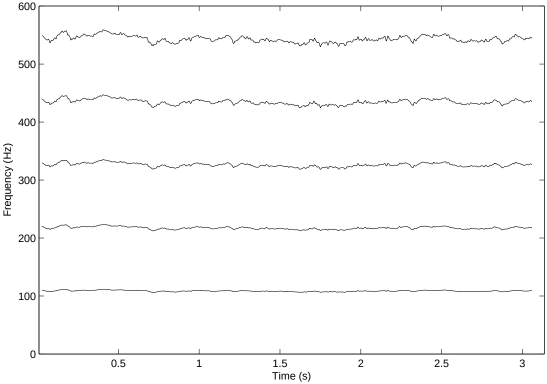
AM-FM Decomposition using aQHM

- Speech is modeled as a sum of AM-FM sinusoids:

$$s(t) = \sum_{k=1}^{K} a_k(t) cos(\phi_k(t))$$

  - K is the number of components,
  - $a_k(t)$ is the instantaneous amplitude of the k^{th} sinusoid,
  - $\phi_k(t)$ is the instantaneous phase of the k^{th} sinusoid, and
  - $f_k(t) = \frac{1}{2\pi} \frac{d\phi_k(t)}{dt}$ is the instantaneous frequency of the k^{th} sinusoid.
- AM-FM decomposition algorithm tries to estimate the instantaneous components.

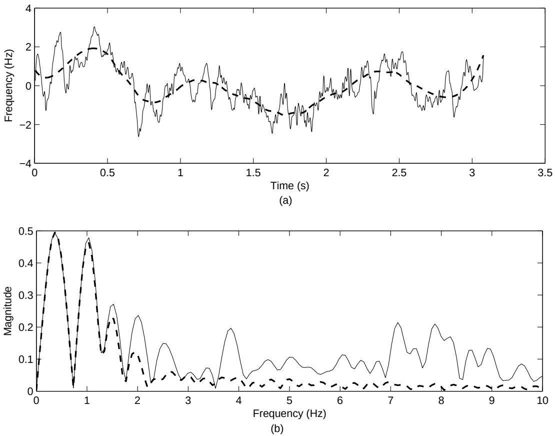
Example of AM-FM decomposition on Speech



Preprocessing of Inst. Component

- Downsample inst. component to $f_s = 1000Hz$
- Remove the very slow ($< 2Hz$) modulations of the instantaneous component.
- This is performed by Savinzky-Golay smoothing filter.
  - S-G smoothing filter performs a local polynomial regression.
  - S-G filter parameters: 4th order polynomial & $1sec$ frame size.
  - Advantage: Preserve features of the time-series such as relative maxima, minima and width.

S-G Filter Output



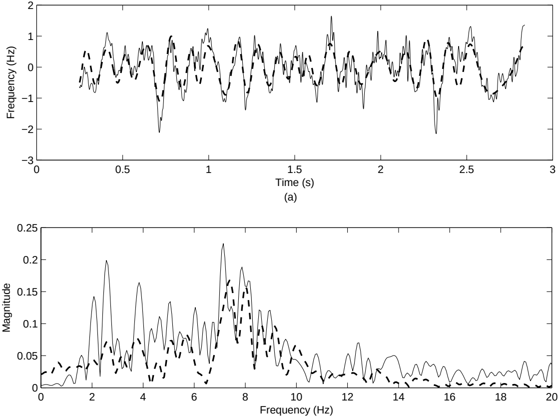
Compute Modulation Frequency & Level

- Assuming that the processed inst. component has a single but time-varying modulation frequency and modulation level.

$$x(t) = m(t)cos(\psi(t))$$

- Apply for second time the AM-FM dec. alg. to the processed inst. component.
- Thus,
  - Modulation frequency, $\frac{1}{2\pi}\frac{d\psi(t)}{dt}$, is estimated from the FM component of AM-FM dec. alg.
  - Modulation level, $m(t)$, is estimated from the respective AM component.

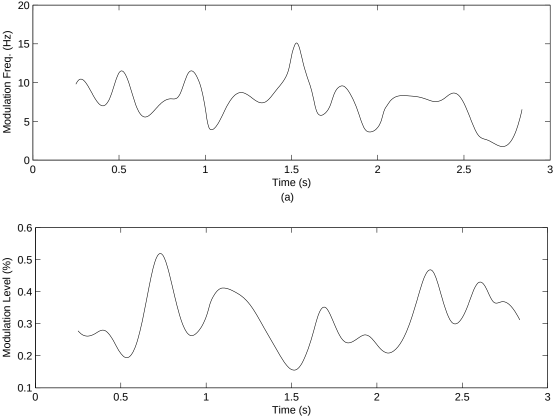
Compute Modulation Frequency & Level



(a)

(b)

Compute Modulation Frequency & Level



(a)

(b)

Voice Fatigue and Acoustic Features of Vocal Loading

- Voice Fatigue
  - Strain of the laryngeal tissues.
  - Relation between occupational voice fatigue and voice pathologies.
- Acoustic Features
  - Fundamental frequency raise.
  - Sound pressure raise.
  - Vocal tremor attributes raise (Boucher et 2008)
    - strain of the laryngeal muscles may affect the speaker's ability to sustain constant tension of the vocal folds.

Examining the Relationship between Vocal Loading and Tremor Attributes

- Estimating vocal tremor attributes:
  - extract instantaneous frequency and instantaneous amplitude.
- Comparing vocal tremor attributes before and after vocal loading:
  - compare the modulation frequencies and the modulation levels of two voiced signals of the same speaker before and after vocal loading.

Definitions

- Vocal Loading Amplitude Indicator (VLAI) $=$ Mean modulation level after loading - Mean modulation level before loading.
- Vocal Loading Frequency Indicator (VLFI) $=$ Mean modulation frequency after loading - Mean modulation frequency before loading.
  - positive value: increase of vocal tremor attributes $\Rightarrow$ possible degradation of voice.
  - negative value: decrease of vocal tremor attributes $\Rightarrow$ possible enhancement of voice.

DB1: Comparing VLAI and VLFI to Subjective Evaluations (SE)

	Speakerid	*VLAI*	*VLFI*	*Student SE*	*Trainer SE* Pre:Post
Female	1	-0.03	0.78	-1	0:1
	2	0.06	-0.47	-1	-1:-1.5
	5	-0.17	0.12	-3	-1.5:-1.5
Male	4	-0.03	-0.07	0	-1:-0.5
	3	-0.01	-0.21	-3	-2:-2.5
	6	0.08	-0.01	0	-1:-2

Reminder:

- Student SE: 0 (no tired) to -3 (very tired).
- Trainer SE: -3 (being very poor voice) to +3 (being excellent).

Summary for this task

- No relation seems to be between vocal loading and voice tremor.
- There is a correlation between objective and subjective evaluations for voice quality assessment.

References for the work on Tremor

1. Yannis Pantazis, Maria Koutsoyannaki and Yannis Stylianou:
   A novel method for the extraction of vocal tremor, MAVEBA-2009, Florence, Italy, 14-16 Dec, 2009.
2. Maria Koutsoyannaki, Yannis Pantazis, Yannis Stylianou, and Philippe Dejonckere:
   Tremor in speakers with spasmodic dysphonia, MAVEBA-2011, Florence Italy, Aug 2011

Acknowledgments

- My students: Maria Markaki, Maria Koutsoyannaki
- My ex-student: Yannis Pantazis.
- Prof. Juan Ignacio Godino-Llorente, and J.D. Arias-Londono (PhD) (UPM, Spain)
- Prof. Anne-Maria Laukkanen (Un. of Tampere, Finland) for providing the database with vocal fatigue examples.

THANK YOU

for your attention