
Voice Characteristics of MARSEC Speakers



Eric Keller

LAIP - IMM - Lettres, Université de
Lausanne, 1015 Lausanne, Switzerland,
eric.keller@imm.unil.ch

Supported by OFES under COST 277

How can voice styles be classified?



Sports Commentary



Prepared News



Telling Tales



On-the-spot Interviews

Voice quality is:

- situation-linked
- differential
- in part stable
- perceptually identifiable

Thus VQ should be:

- somewhere in the acoustics
- classifiable by statistical techniques

An exploratory study:

- spectral profiles of vowels
- global “source” + “filter” investigation
- large corpus with varied speakers and speech styles

MARSEC: Machine-Readable Spoken English Corpus

- Numerous speakers (56/65 produced more than 25 vowels)
 - All RP British English speakers
 - All practiced adult speakers
 - A good spectrum of speech styles
 - Recording quality generally quite good (BBC, Open University)
 - Several speakers contributed extended material → intra-/inter-subject comparisons.
 - Prosodics has been examined by other authors
- A large, naturalistic corpus such as MARSEC permits to initiate the examination of the statistical effects of the finer shades of gender, attitude and thematic coloration of voice that have largely been left aside in studies bearing on the voice quality of emotion portrayals.
-

What was done, in short

1. Data gathering:

- Developed a **vowel detector**
- Obtained some 30'000 vowel nuclei from 56 MARSEC speakers
- Extracted a 1024-pt log magnitude spectrum from each nucleus
- Averaged spectral rays over each bark band [BB]
- Linearized BB responses with a square root transformation
- Standardized linearized BB responses for each subject (Z-score)
→ a **30'453*22-cell matrix of BB responses**

2. Factor analysis:

→ **Areas of common spectral response**

3. Difference/predictive analysis:

→ **Gender and speech style differences in spectral response**

4. Cluster analysis:

→ **Underlying factors of spectral responses**

Vowel Detector - Method

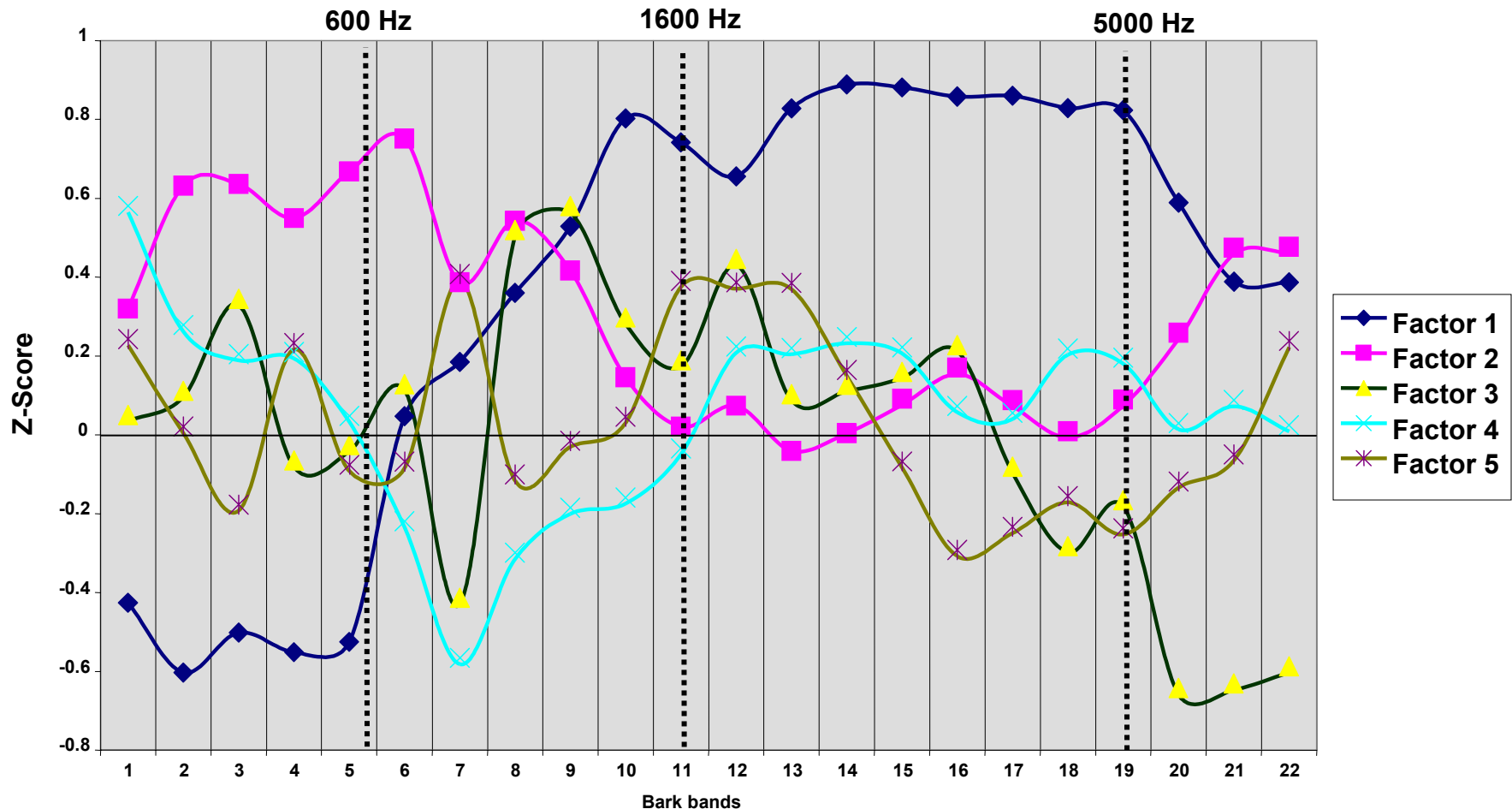
- Linearized BB responses → **two multiple linear regression models**:
 - Used unstandardized BB responses to predict the **presence of vowels** in 87 sentences with manually verified segmentation, taken from a speaker in the corpus (Brian Perkins from BBC-4 News).
 - Predicted the **absence of silent periods** in similar fashion.
 - Vowel nuclei were accepted if...
 - *low probability* of silent period
 - *high probability* of vowel nucleus
 - Pearson correlation $r = .889$ for vowel/non-vowel judgement on 1067 vowels.
 - **Errors (N=120)**: 69 spectra from voiced segments, 17 from silent periods and 34 (or about 3.2%) from unvoiced spectra.
 - **30'453 vowel nuclei** identified in this manner.
-

Factor analysis:

Areas of common spectral response

- **Initial question:** What level of grain is required to analyze differential spectral responses?
 - **Spectral ray** (e.g., 1024 data points from 0 – 8'000 Hz)?
 - **Bark band** (i.e, 22 data points from 0 – 8'000 Hz)?
 - **Larger regions** with even fewer data points?
 - **Factor analysis** can be used advantageously: identifies same-directional responses in a matrix, e.g., it can identify groups of spectral ray or BB responses acting in concert.
 - **Factor analysis** performed on 22 BB responses for 30'453 vowel nuclei.
-

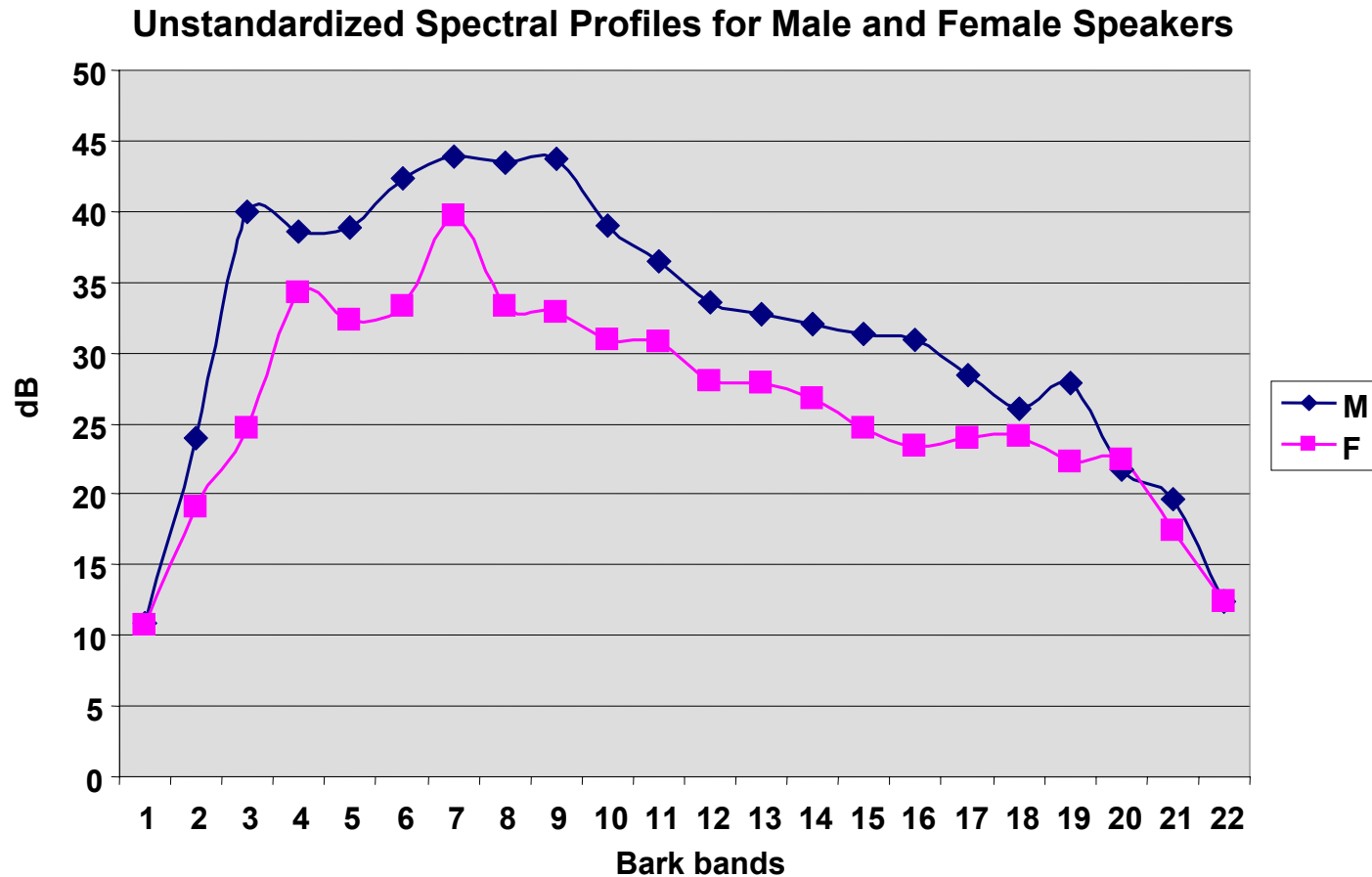
Factor analysis: Areas of common spectral response



Five factors with eigenvalues ≥ 1.0 : about 42% of total variation (all 30'453 vowel nuclei):
Four regions of roughly same-directional variation, delimited by 600, 1600, 5000 and 8000 Hz.

Difference/predictive analysis — visual:

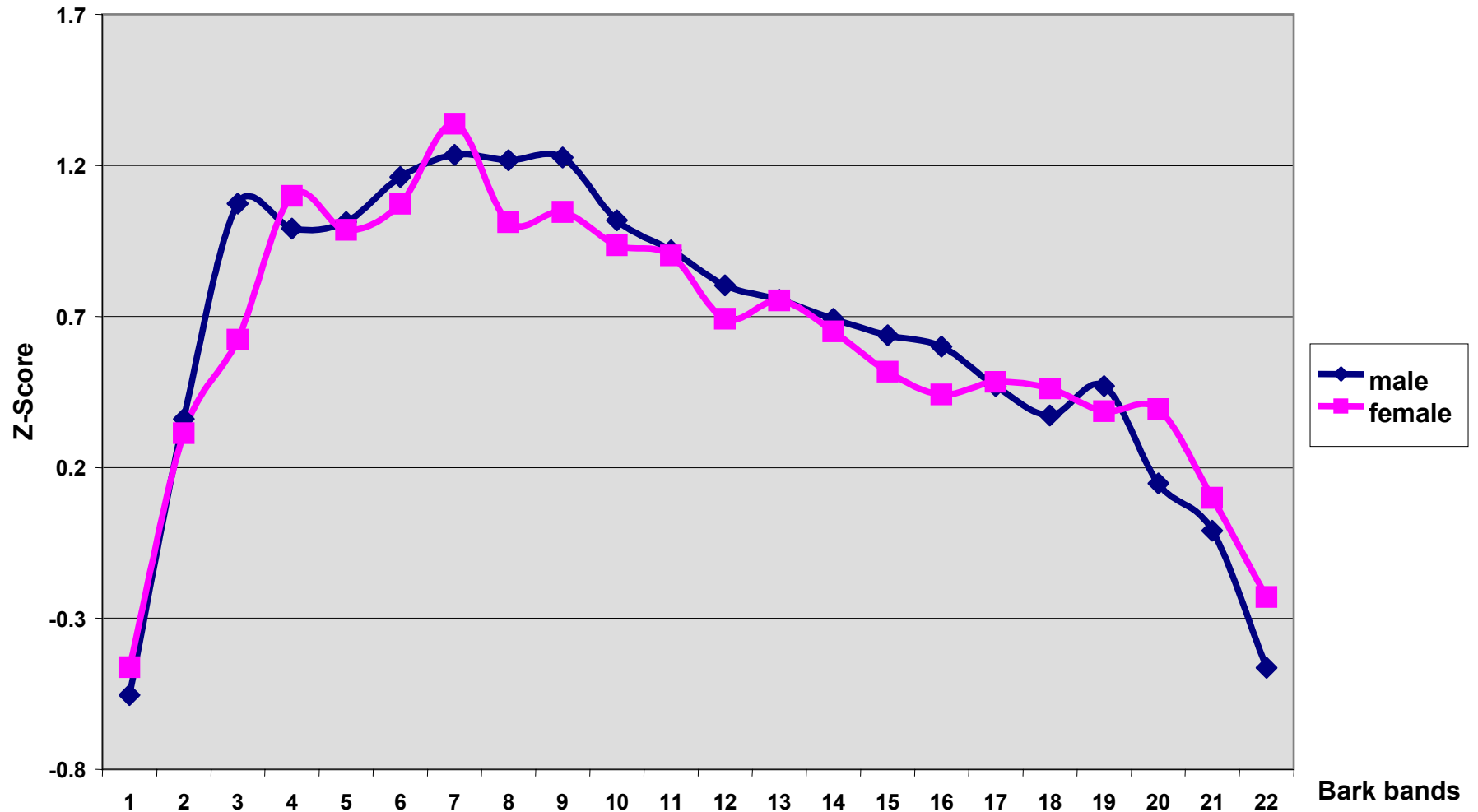
1. Gender differences (non-standardized)



Female speakers show globally lower volume than male speakers in unstandardized data.

Difference/predictive analysis — visual:

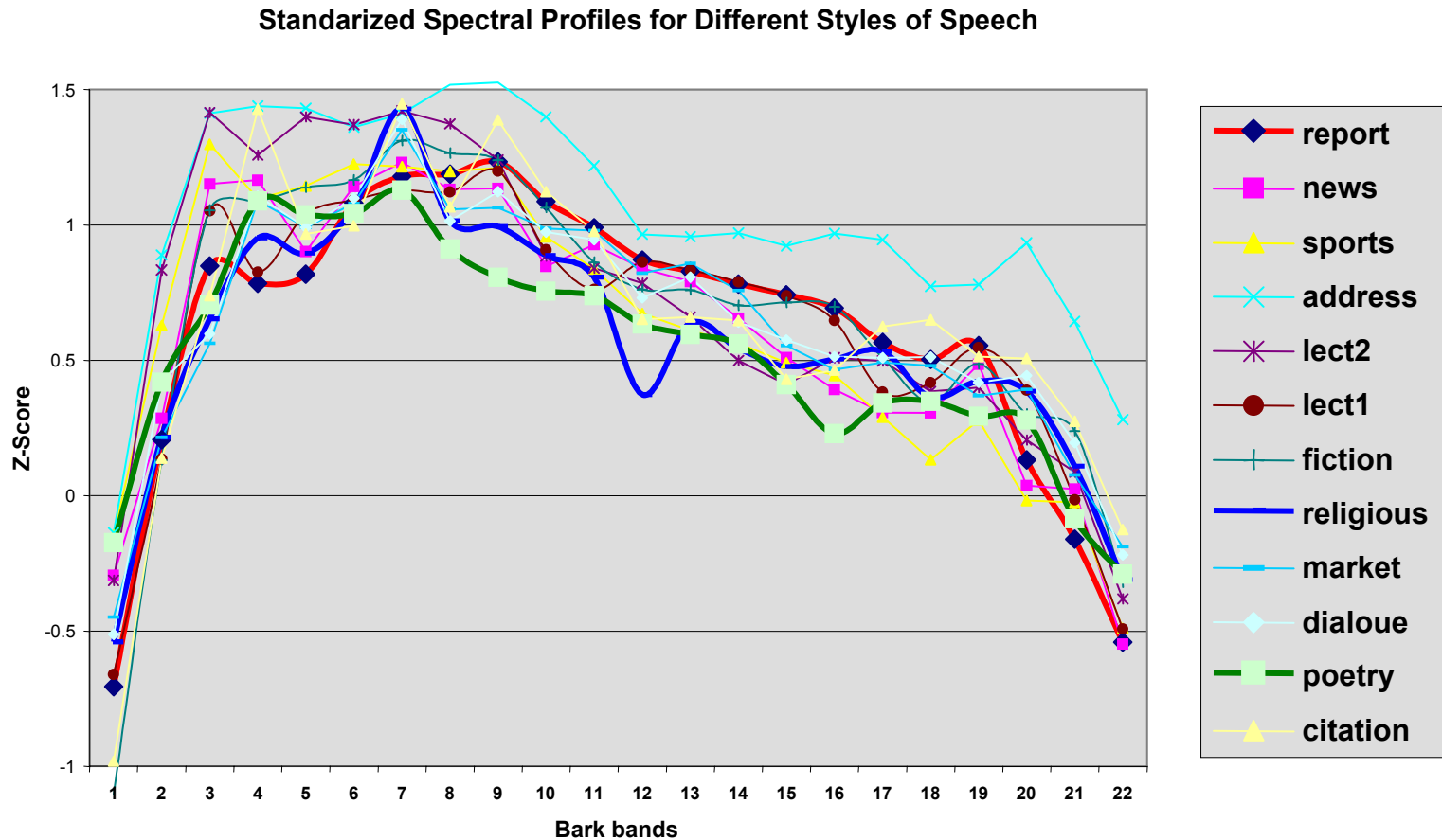
1. Gender differences (standardized)



In standardized data, the only major gender differences occur around Bark bands 3 and 8-9.

Difference/predictive analysis — visual:

2. Style differences (standardized)



There were complex, relatively minor spectral variations as a function of style of speech

Difference/predictive analysis — statistical (1):

- Multiple regression of type
Intercept + STYLE + SEX + STYLE*SEX
to predict response in **four spectral regions**:
 - Wilk's Lambda significant at $p = .000$ for:
 - **Fstyle = 262.610**
 - **Fsex = 56.106**
 - **Fstyle*sex = 176.130**
 - **Modelling: $r = 0.647$, **0.210**, 0.421 and 0.366 for the four spectral regions (average: $r = .411$, 16.9% of total variation).**


- Multiple regression on the **22 Bark bands**
 - Wilk's Lambda significant at $p = .000$ for:
 - **Fstyle = 294.537**
 - **Fsex = 38.674**
 - **Fstyle*sex = 220.857**
 - **Modelling: average $r = 0.402$, 16.2% of total variation**

Difference/predictive analysis — statistical (2):

■ **Summary:**

- Gender and speech style have significant effects on the overall spectral profile of vowels, particularly so in the first and third spectral regions.
 - Of the two examined predictors, style appears to be a better predictor than gender in the standardized data.
 - Average correlations show that the strongest predictions are made for the first spectral region.
 - The similarity in the prediction of spectral magnitude values in the four spectral regions and in Bark bands provides an initial validation of the four-part subdivision of the spectral domain.
-

Cluster analysis (see handout)

- **Cluster analysis** groups speakers according to similarity in spectral behaviour
 - Loose groupings for:
 - Sports reports – KG2 
 - Market reports
 - Fiction and poetry – JH 
 - Telephone reports by international correspondents
 - Liturgy and lectures – DG 
-

Conclusion

- Average vowel spectral profiles are differentially related to gender and speech style (and underlying speech context)
- Average spectral variation probably involves four regions of differential spectral response in the 0-8'000 Hz range.
- **In interaction with**
 - **speaker-dependent source activity,**
 - **language-dependent formant activity,**
 - **as well as prosodic information,**the parameters identified here may well play a useful role in the individuation of voice quality.

Epilogue (suggested by Dr. Fourcin)

- Subsequent to the submission to VOQUAL, I implemented averaged mean envelope changes in an HNM (Harmonics & Noise Modelling) framework.
 - Using signals from Brian Perkins, I modified all vowel envelopes in accordance with BB differences between BP's mean BB envelope and that of a number of other MARSEC speakers.
 - The results were disappointing. I could just barely hear the differences.
 - Unfortunately, I did not save any of the files. But the result motivated me to use the rather conservative interpretation of the results I just presented.
-