Speaker and Language Independent Voice Quality Classification Applied to Unlabeled Corpora of Expressive Speech

John Kane, Stefan Scherer, Matthew Aylett, Louis-Philippe Morency and Christer Gobl

Motivation:

- Voice quality plays a pivotal role **in speech style variation**.
- **Synthesis of voice quality changes** without compromising naturalness is a key objective for current speech technology.
- Separating **within- and across-speaker differences** requires the analysis of large labeled corpora.
- We apply state-of-the-art voice quality analysis to **large speech corpora** that are readily available such as audiobook data.

Speech Data:

- **CereVoice Subcorpora:** acted lax, modal and tense voices; 8 hours 11 minutes; 15 different speakers; single utterances
- **Audiobook data:** *A Tramp Abroad* and *Pride and Prejudice*; one male one female speaker; approx. 2 hours of data each and about 2000 utterances each

Extracted Features:

The three below parameters are derived from the glottal source signal as estimated by Iterative Adaptive Inverse Filtering (IAIF).

- **NAQ**- normalised amplitude quotient is a correlate of the glottal closing quotient.
- **QOQ** - the quasi-open quotient is another amplitude based glottal measurement and is a correlate of the standard Open Quotient (OQ).
- **H1-H2** - the difference in amplitude between the first two harmonics measured from the narrowband glottal spectrum.
- **peakSlope** - is a spectral slope correlate and is derived by decomposing the speech signal into different frequency bands using wavelet analysis.
- **MDQ** - the same wavelet based decomposition as is used in peakSlope is applied to the Linear Prediction residual in the calculation of the Maxima Dispersion Quotient.
- **f0** – fundamental frequency and **MFCC** features

Experimental Protocol:

- We utilize **fuzzy support vector machines (FSVM)** for the classification.
- The output of the is a three dimensional vector of **membership assignments** to the three classes of lax, modal and tense.
- Protocol:
  - Training of FSVM on **CereVoice subcorpora**.
  - **Objective evaluation** of performance in subject independent leave one speaker out experiment.
  - **Subjective evaluation** of k-means clustering.

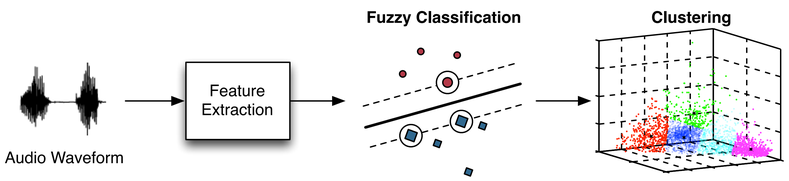

Figure 1: Processing pipeline of the presented approach. After feature selection memberships are assigned to samples for three dimensional clustering.

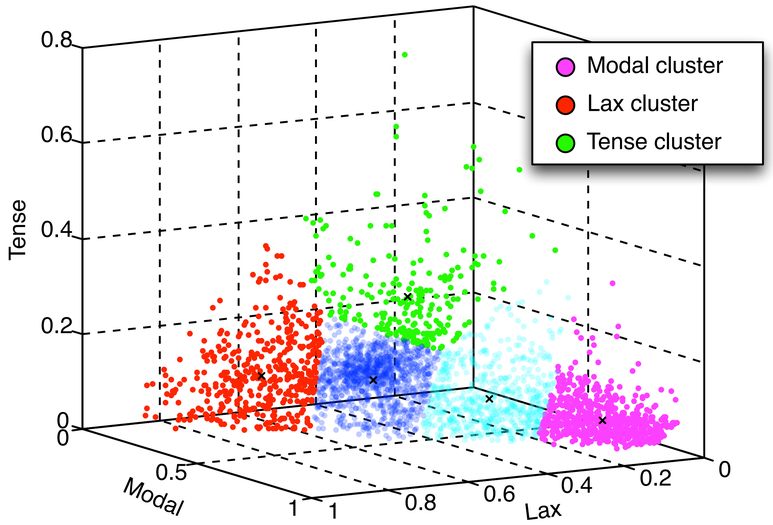

Figure 2: Sample k-means clustering of audiobook *Pride and Prejudice*. The outer clusters in the corners are corresponding to the clear voice qualities (i.e. lax, modal and tense). The inner two clusters are mixed and not used in the perception test.

Objective Evaluation:

- We observe a performance of 56.78% (std. 12.57) accuracy for the FSVM based on MFCCs and f0 only.
- When including **VQ features** the performance significantly improves accuracy to 64.40% (std. 13.96; pairwise t-test: $T(18) = 3.46$, $p < .003$) .

	MFCC+f_0			MFCC+f_0+VQ		
	L	**M**	**T**	**L**	**M**	**T**
Lax	60.93	26.61	12.44	**66.82**	24.62	8.54
Modal	29.60	41.02	29.36	23.62	**50.86**	25.50
Tense	13.62	17.98	68.39	7.89	16.58	**75.51**

Table 1: Confusion matrix of speaker independent VQ classification experiments.

- The audiobooks contain samples with varying voice quality. The predicted voice quality of all the speech samples in the **audiobook data are clustered into five distinct clusters**, as visualised in Figure 2. The corners of the triangle shaped space represent the areas where the classifier clearly could identify a dominant voice quality to be present.

	Pride and Prejudice			A Tramp Abroad		
	L	**M**	**T**	**L**	**M**	**T**
C-Lax	**0.67**	0.22	0.11	**0.80**	0.18	0.02
C-Modal	0.14	**0.80**	0.06	0.17	**0.80**	0.03
C-Tense	0.37	0.35	**0.28**	0.08	0.24	**0.68**

Table 2: Cluster centers of lax modal and tense samples for both audiobooks.

Subjective Evaluation:

- 30 participants completed the subjective evaluation and display **high inter-rater reliability**, as assessed using Krippendorff's alpha (All samples: 0.87, Acted samples: 0.95, Audiobook samples: 0.75).
- We further conduct several ANOVA in order to identify the sources of disagreement:
  - the mean difference between human ratings and classifier output **is significantly** [$F(1,118) = 26.42$, $p <$ 0.0001] higher **for female speakers**
  - the mean difference between human ratings and classifier output **is not significant** [$F(1,118) = 0.44$, $p =$ 0.507]

USC Institute for Creative Technologies

Dr. Stefan Scherer
scherer@ict.usc.edu