DETECTING A TARGETED VOICE STYLE IN AN AUDIOBOOK USING VOICE QUALITY FEATURES

Éva Székely[1] John Kane[2] Stefan Scherer[2,3] Christer Gobl[2] Julie Carson-Berndsen[1]

[1]CNGL, School of Computer Science and Informatics, University College Dublin, Ireland [2]Centre for Language and Communication Studies, Trinity College Dublin, Ireland
[3]Institute for Creative Technologies, University of Southern California, Los Angeles

Audiobooks are known to contain a variety of expressive speaking styles that occur as a result of the narrator mimicking a character in a story, or expressing affect. An accurate modeling of this variety is essential for the purposes of speech synthesis from an audiobook. Voice quality differences are important features characterizing these different speaking styles, which are realized on a gradient and are often difficult to predict from the text. The present study uses a parameter characterizing breathy to tense voice qualities using features of the wavelet transform, and a measure for identifying creaky segments in an utterance. Based on these features, a combination of supervised and unsupervised classification is used to detect the regions in an audiobook, where the speaker changes his regular voice quality to a particular voice style. The target voice style candidates are selected based on the agreement of the supervised classifier ensemble output, and evaluated in a listening test.

Aim:
To find all utterances in a corpus that are realized with a targeted voice style characterized by the following features:

- ✓ tense voice quality;
- ✓ occasional creaky segments;
- ✓ relatively low mean f0.

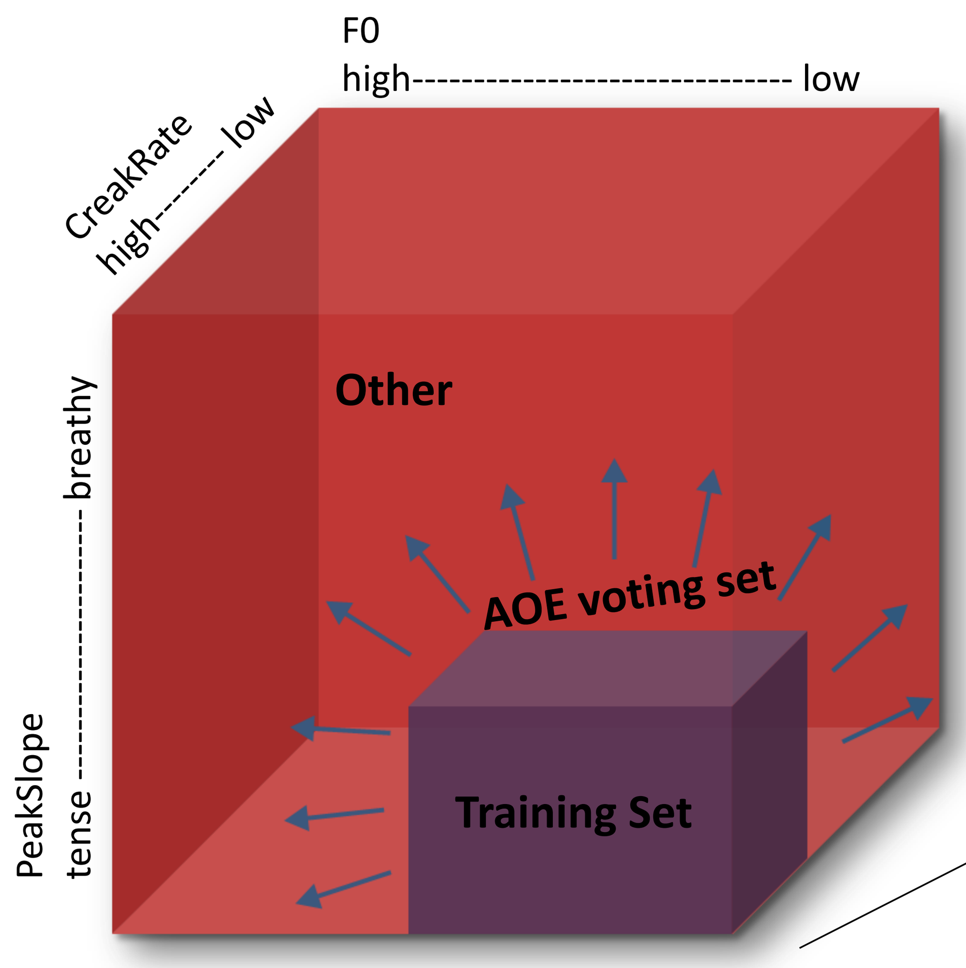


Illustration of the method of finding all utterances featuring a targeted voice style in a corpus

Features:

PeakSlope: a measure discriminating voice qualities on a breathy-to-tense continuum, derived following a wavelet-decomposition of the speech signal.

$$g(t) = -cos)(2\pi f_n t) \cdot exp(\frac{-t^2}{2T^2})$$

Where:

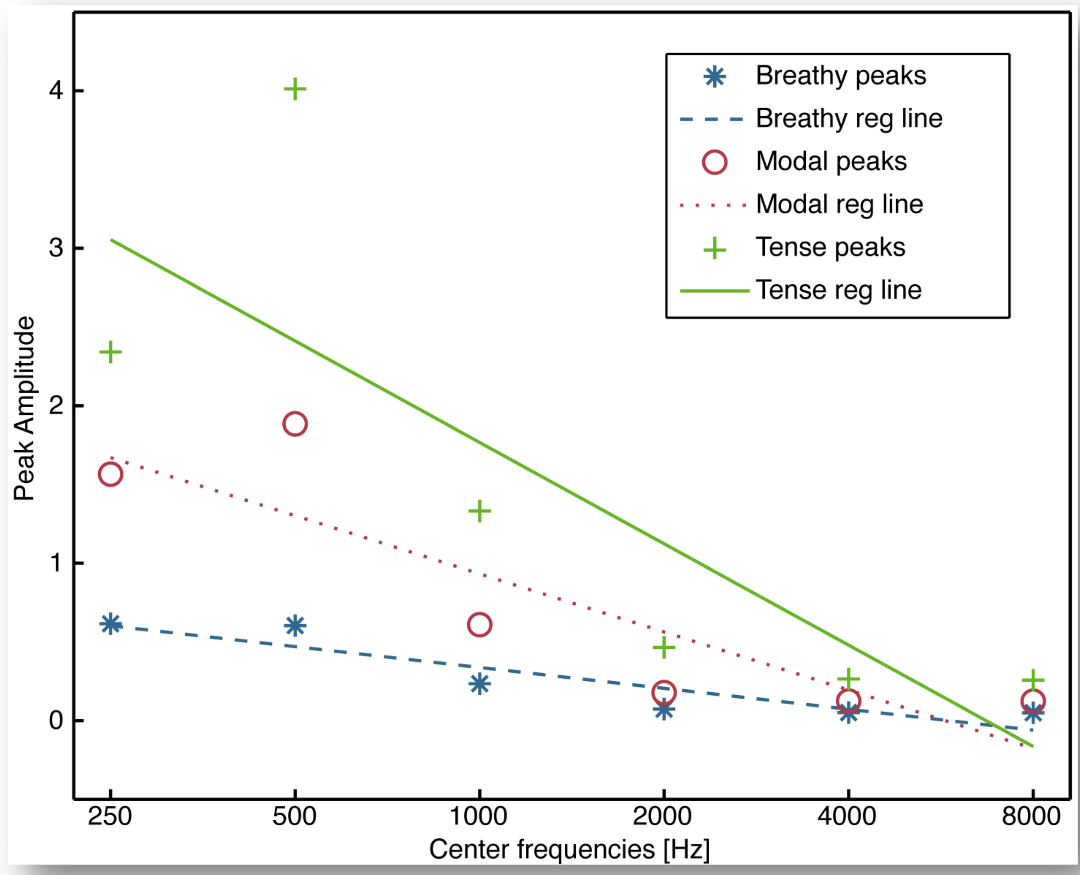
fs = 16kHz, $f_n = \frac{f_s}{2}$ and $T = \frac{1}{2f_n}$

CreakRate: the number of creaky segments divided by the number of voiced segments in an utterance. Parameters contributing to creak decision:

- Power peaks (PwP)
- Intra-Frame Periodicity (IFP)
- Inter-Pulse Similarity (IPS)

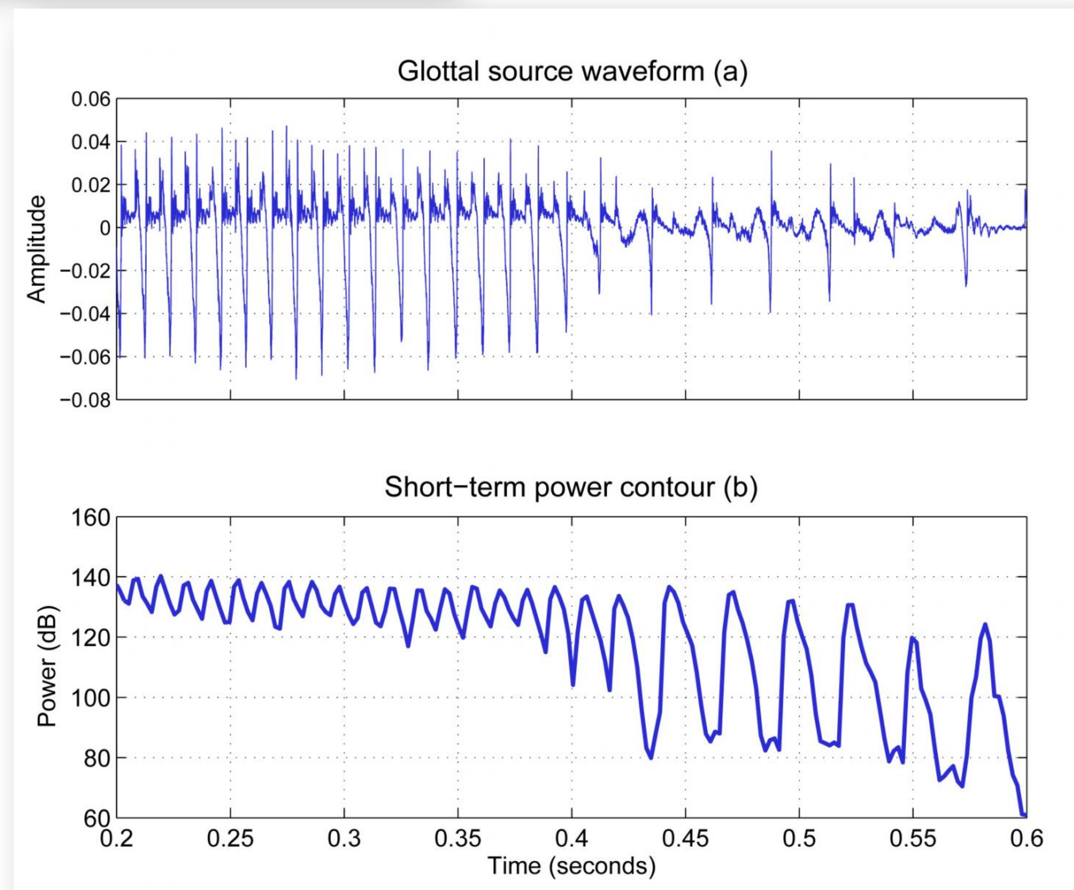
F0: the mean F0 over an utterance.

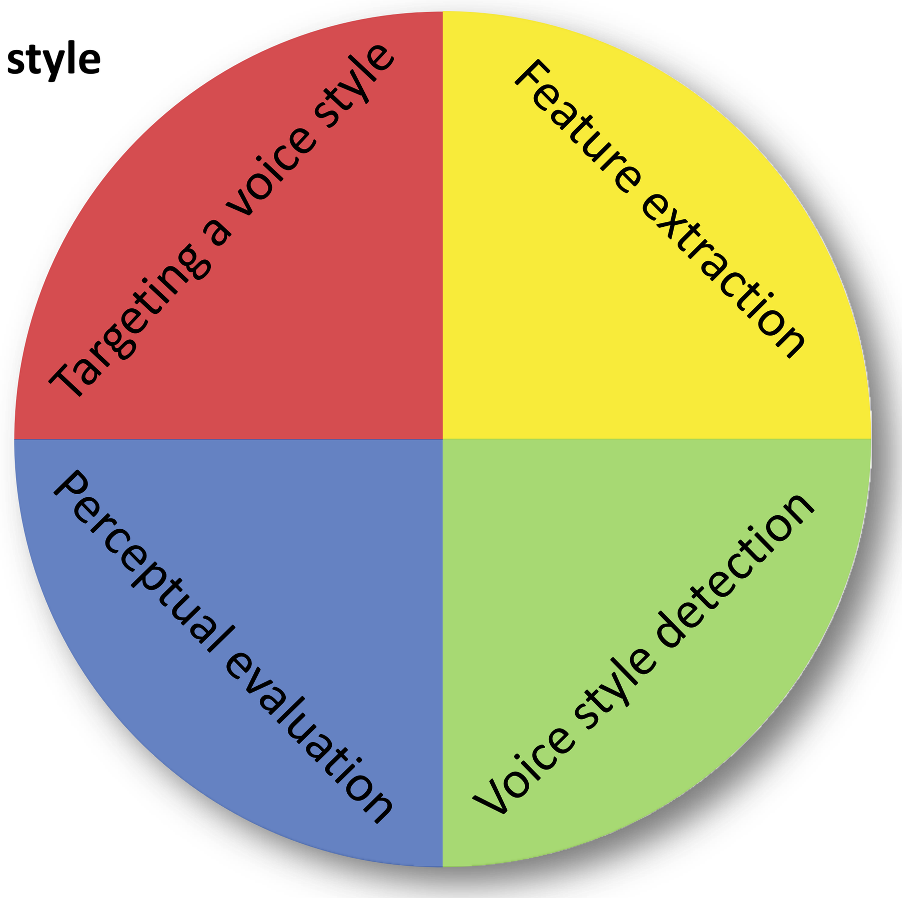
Targeted voice style



Wavelet peak amplitudes with regression lines for the center of an /o/ vowel produced by a male speaker in breathy, modal and tense voice qualities.

Glottal source waveform (a), estimated by inverse filtering and the very short term power contour (b) of an /a/ vowel produced by a male speaker which begins in a modal voice quality but changes into creak from around 0.4 seconds.





Agreement Optimized Ensemble voting: an ensemble of two classifiers: a fuzzy-output support vector machine (FSVM), and a Gaussian mixture model (GMM)

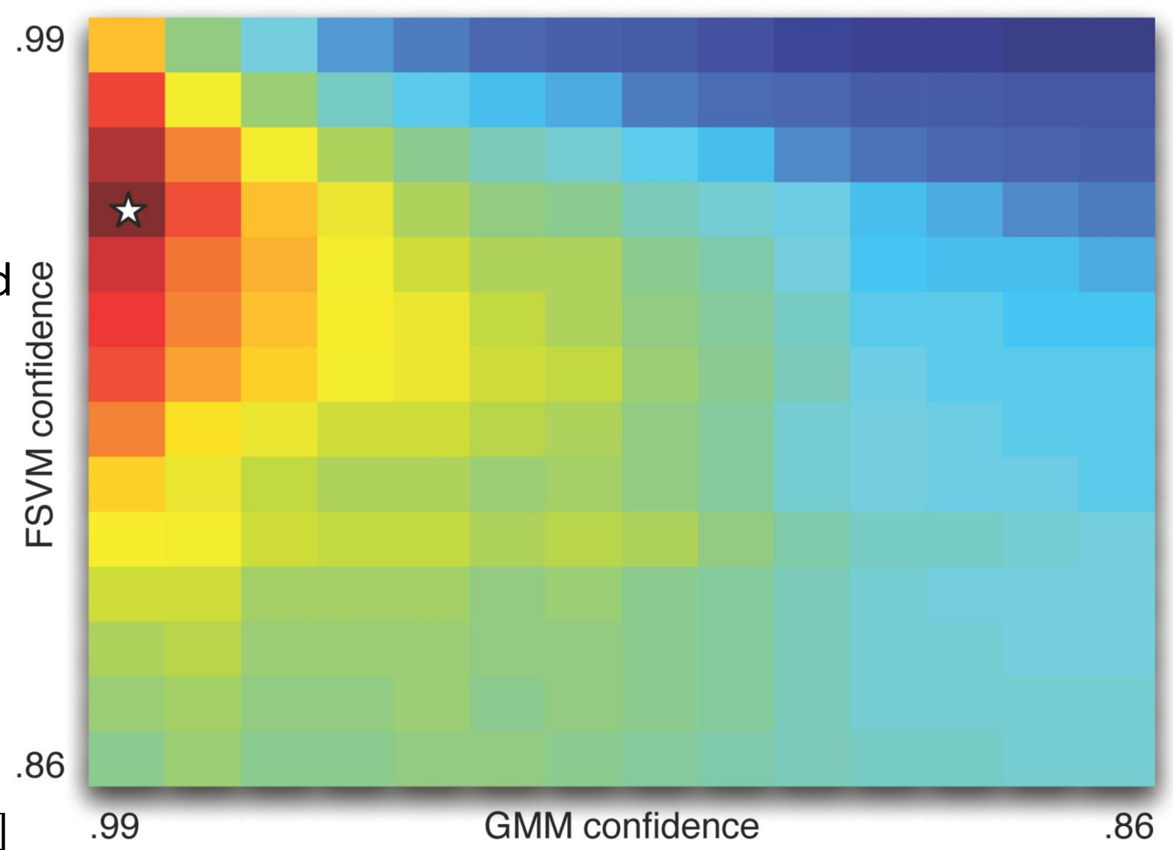
The speech samples outside the training set are classified using a trained ensemble based on the confidence of the output of the two classifiers.

FSVM confidence: the distance $d(x)$ of sample x to the separating hyperplane normalized using the sigmoid function:

$$c_{fsvm}(d(x)) = \frac{1}{1 - exp^{-d(x)}} \in [0,1]$$

GMM confidence: the posterior probability of sample x given model m_j:

$$c_{gmm} = P(x|m_j) \in [0,1]$$



Ensemble voting heat map of the confidence measures c_{fsvm} and c_{gmm} ranging between [0:99; 0:86]. Warm colors indicate good overlap measure and the star indicates the optimal value.

The optimal confidence thresholds for the two classifiers are identified using a measure of relative agreement $relA$:

$$relA(c_{fsvm}, c_{gmm}) = \frac{1}{|cand_{all}|}\left(\frac{|cand_{en}|}{|cand_{fsvm}|} + \frac{|cand_{en}|}{|cand_{gmm}|}\right)$$

Where:

$cand_{en}$ = agreement between the classifiers' output candidates: $cand_{fsvm} \cap cand_{gmm}$

$cand_{all}$ = overall number of selected candidates: $cand_{fsvm} \cup cand_{gmm}$

Evaluation: A-B listening test competed by 27 participants.

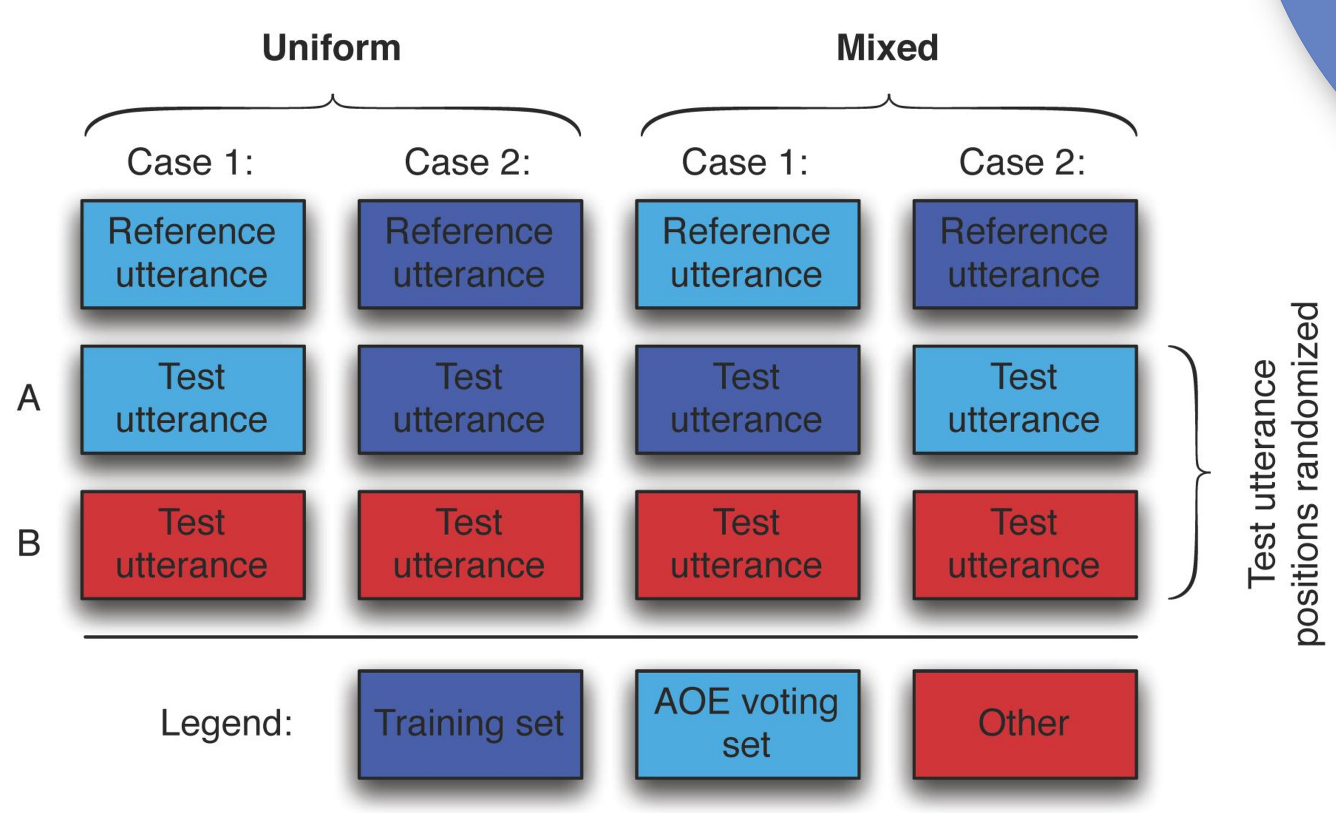


Illustration of various trial setups in the perception test. Stimuli: 60 randomly selected utterances: 20 from each set.

Results		
Group	Accuracy (%)	Standard deviation (%)
Uniform	88.48	10.44
Mixed	85.86	10.50
Combined	**87.04**	**7.88**

Aims of the evaluation:

- ❑ Do listeners perceive the utterances in the target voice style to sound similar to each other?
- ❑ Did the method select the majority of the utterances in the target voice style?
- ❑ Is there a significant difference between the training set and the AOE voting candidates?

Results

- ✓ Yes. We found that listeners' judgments were in 87 % agreement with the classification.
- ✓ Yes. The random selection of the stimuli would have detected target voice style samples remaining in the 'Other' set.
- ✓ Independent t-tests on 'Uniform' and 'Mixed' groups revealed no significant difference (t = -0.919, and p = 0.3623).

This research is supported by Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at UCD. This work was further supported as part of the FASTNET project - Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631