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

eva.szekely@ucdconnect.ie, kanejo@tcd.ie, scherer@ict.usc.edu, cegobl@tcd.ie, julie.berndsen@ucd.ie

ABSTRACT

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.

Index Terms— voice quality, audiobooks, expressive speech, fuzzy support vector machines, speech synthesis, classifier ensemble

1. INTRODUCTION

Audiobooks contain a variety of different expressive speech styles. These speech styles are non prompted, and can sound very natural, as they reflect the speakers own decision and represent the variety of ways the speaker is comfortable using their own voice. This makes an audiobook an attractive corpus for expressive speech synthesis.

We use the term voice style in this work to describe the different ways a speaker produces an utterance in terms of changes in voice quality combined with certain prosodic variation over the course of the entire utterance. The voice styles occurring in audiobooks are not only direct expressions of emotion and affect, but often a result of the speaker deliberately changing their voice quality to imitate different characters. These voice quality changes are highly speaker specific, and they need to be modeled accurately to avoid distortions in the resulting synthetic speech. Moreover, the similar voice styles in the corpus need to be detected and grouped to serve as homogeneous sub-corpora for expressive speech synthesis. Voice quality features have previously been shown to be important in creating speech synthesis platforms of this kind [1].

The traditional way of detecting different voice styles in an audiobook is to use predictions from text. This is often unreliable because the speaker can use different narrator styles, or they can use the same voice style to imitate different characters or express different emotions. Parameters of voice quality are effective indicators of voice style changes within an audiobook. A mixture of voice quality parameters need to be used to reflect the different dimensions and volume of these voice style changes. A voice style detection method using a combination of voice quality parameters and single corpus-based classification is beneficial, because it takes into account the speakers own variety of voice styles [2].

The work presented here is motivated by the results of a previous study [2] where glottal source parameters are used to identify the variety of speaking styles in an audiobook, placing the similar utterances on a continuum of neighboring clusters of a Self-Organizing Feature Map. In the present study, we investigate the possibility of using prior knowledge to target a specific voice style present in the audiobook, and detect the utterances featuring that voice style.

The voice style detection method described here uses a combination of unsupervised and supervised learning to identify the similar sounding utterances to a pre-defined target voice style group. The selection of target voice style candidates is based on agreement optimized multiple classifier system voting, using fuzzy support vector machines and GMMs. In the experiment, we are aiming to detect a particular voice style featuring tense voice quality with a relatively low f_0 and occasional creaks. Informal listening tests reported that the speaker often uses this deviation from his modal voice style, to express affect and involvement. Detecting the utterances using this voice style can help in building a suitable sub-corpus for expressive speech synthesis.

2. VOICE QUALITY MEASUREMENTS

We selected acoustic measurements which were both suitable for characterizing the voice qualities used by the speaker and have also been shown to be useful at discriminating voice qualities in less than ideal recording conditions.

2.1. Breathy to tense - *PeakSlope*

Breathy and tense voice qualities are the most common, and hence the most studied non-modal voice qualities. Breathy voice qualities are characterized by a smooth glottal closure, which results in

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin (UCD) and Trinity College Dublin (TCD). The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland. 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

strong attenuation of higher harmonics [3]. Breathiness also involves a jet of air which becomes turbulent as it passes through the glottis and manifests itself in the speech spectrum as noise, typically from around the third formant region [4]. Tense voice qualities, on the other hand, involve sharp glottal closure, with strong higher harmonics compared with modal or breathy speech. There is no audible aspiration noise in tense voice qualities and the type of phonation is one in which there is a relatively short open phase in the glottal vibration cycle.

A new parameter, name *PeakSlope*, was recently described for discriminating voice qualities on a breathy-to-tense continuum [5]. It performed better than other voice quality measurements even in the presence of simulated noisy conditions. *PeakSlope* is derived following wavelet-based decomposition of the speech signal. This is done using the mother wavelet:

$$g(t) = -cos(2\pi f_n t) \cdot exp(\frac{-t^2}{2\tau^2}) \quad (1)$$

where the sampling frequency, fs = 16 kHz, $f_n = \frac{fs}{2}$ and $\tau = \frac{1}{2f_n}$. The speech signal, $x(t)$, is the convolved with the scaled version of the wavelet, $g(\frac{t}{s_i})$, where $s_i = 2^i$ and $i = 0, 1, 2, ..., 5$. This results in having a filter bank with the frequency responses of the scaled wavelet having center frequencies at different octave bands: 8 kHz, 4 kHz, 2 kHz, 1 kHz, 500 Hz and 250 Hz. Absolute amplitude maxima are then measured from each of the outputted waveforms. It was previously shown that if a regression line was fit to these maxima, the slope of the line was able to robustly discriminate breathy-to-tense voice qualities. This is further highlighted in Fig. 1.

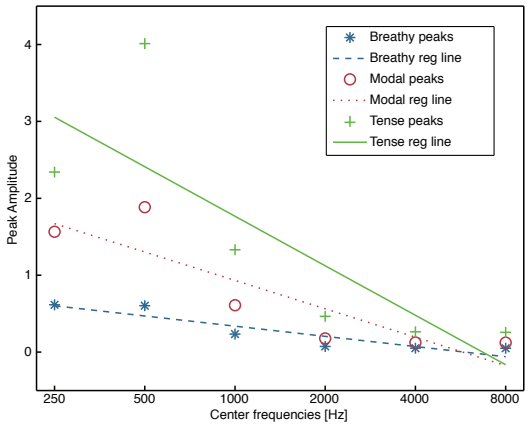

Fig. 1. Wavelet peak amplitudes with regression lines for the center of an /o/ vowel produced by a male speaker in breathy, modal and tense voice qualities.

For the current study wavelet based decomposition is carried out on the whole speech utterance and then measures of *PeakSlope* are carried out using a 32 ms rectangular window, on the outputted waveforms, and a 10 ms frame shift is used.

2.2. Creak detection

Creaky voice qualities (also called vocal fry) are characterised by an extremely low f_0, irregular pulsing and at times the presence of secondary and even tertiary excitations [6]. These acoustic characteristics are the result low levels of longitudinal and high levels of adductive vocal fold tension combined with low levels of subglottal pressure [3]. Creaky voice qualities can occur from a phenomenon called *ventricular incursion* [10] where the ventricular folds adduct on the *true* vocal folds resulting in additional mass (leading to lower frequency of vibration) [10]. This can provide conditions for laryngeal vibration also occurring above the level of the glottis, which may account for the secondary and tertiary excitations. For detecting speech segments containing creak we used a set of parameters described in [7]. It was shown in the original paper that a very short term power contour, with 4 ms frame length and 2 ms shift, could be used for detecting candidate creak pulses. One can observe in Fig. 2 (panel B) the very large fluctuations in power towards the end of the contour, where the speaker produces creaky phonation. This clearly corresponds to the change in the glottal source waveform at the same time point (Fig. 2, panel A). A power peaks (PwP) parameter is used to quantify the amplitude of these fluctuations and highlight candidate creak pulses.

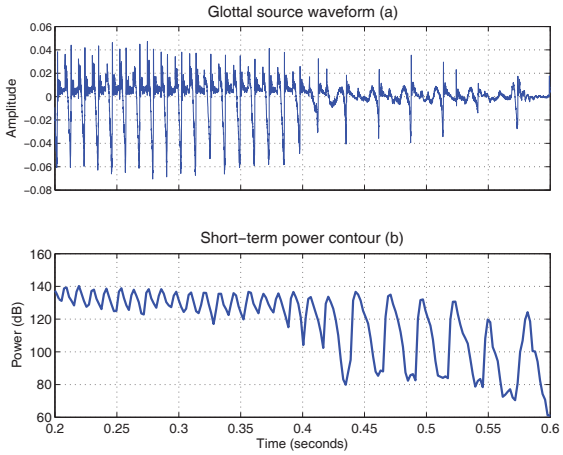

Fig. 2. Glottal source waveform (a), estimated by inverse filtering and the very short term power contour (b) of an /a/ vowel produced by a male speaker which begins in a modal voice quality but changes into creak from around 0.4 seconds.

Then, in order to differentiate between creaky segments and 'normal' voiced segments an Intra-Frame Periodicity (IFP) measure is used. This involves using a normalized autocorrelation function of a 32 ms windowed frame of the speech signal. One would expect strong repeating peaks above zero-lag if the frame contains normal voiced speech. But for creaky segments, as they typically either display very long or irregularly spaced glottal pulse lengths one would expect relatively weaker peaks in the autocorrelation function. A further measure is used to discriminate between creaky segments and unvoiced speech. The Inter-Pulse Similarity (measure) involves taking a cross-correlation function from speech segments around consecutive creak candidate locations (determined from the power contour). It is likely that consecutive creak segments would display a reasonable degree of similarity compared with unvoiced segments. Hence, this would show comparably stronger peaks in the cross-correlation function for consecutive creak segments. Finally, in order to make the binary decision on the presence or absence of creak we use the suggested parameter thresholds given in the original papers, i.e. PwP $\geq$ 7 dB, IFP $\leq$ 0.5, IPS $\geq$ 0.5, for a segment to be considered to contain creak.

3. SPEECH SAMPLE CANDIDATE SELECTION BASED ON AGREEMENT OPTIMIZED ENSEMBLE VOTING (AOE VOTING)

Based on the voice quality measures described in Section 2, we se-

lect a voice style group as a narrow target style (see Section 4.3), that we want to widen with similar speech samples, and one opposing group, that did not fulfill the criteria. Based on this selection, we trained an ensemble of two classifiers, namely a fuzzy-output support vector machine (FSVM), and a Gaussian mixture model (GMM) [8, 9].

The AOE voting is conducted as follows: The remaining unlabeled samples $x \in X$ are classified using the trained ensemble and selected to be either in the broadened target group (i.e. candidates) or not, based on the confidence of the classifiers' output. In the case of the FSVM the confidence c_{fsvm} is measured as the distance $d(x)$ of sample x to the separating hyperplane normalized using the sigmoid function $c_{fsvm}(d(x)) = \frac{1}{1-\exp^{-d(x)}} \in [0, 1]$. For the GMM the confidence c_{gmm} is measured as the a posterior probability of sample x given model m_j: $c_{gmm} = P(x|m_j) \in [0, 1]$.

The optimal confidence thresholds for the two classifiers are identified using a measure of the relative agreement $relA$, taking agreement (i.e. the number of agreeing candidates of the ensemble) between the classifiers' output candidates for the broadened target class $cand_{en} = cand_{fsvm} \cap cand_{gmm}$ and the overall number of selected candidates $cand_{all} = cand_{fsvm} \cup cand_{fsvm}$ into account:

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

As the candidate lists vary with respect to the confidence threshold of the classifiers, the confidence values c_{fsvm} and c_{gmm} are varied in order to find the maximal $relA$. The result and experimental setup for this study is described in Section 4.4.

4. EXPERIMENT

4.1. Corpus

The corpus used for the experiment is part of an open source audiobook originally published on librovox.org, read by John Greenman. The segmented audio was made available for Blizzard Challenge 2012 Toshiba Research Europe Ltd, Cambridge Research Laboratory. The method used to align the audio with the corresponding text and segment it into smaller utterances is described in [10]. One of the four available Mark Twain books, *A Tramp Abroad* was selected for this experiment. This was necessary to eliminate changes of the recording environment. A pilot corpus of 3017 utterances containing a variety of highly expressive speech styles was formed from the utterances of *A Tramp Abroad* that were no longer than 5 seconds. Based on informal listening tests it was assumed that the vast majority of these utterances contain no abrupt changes of voice style.

4.2. Preparation of parameters

After the voice quality parameters were estimated as described in sections 2.1 and 2.2, they were transformed into input features of the classification, one feature per parameter per utterance. This was necessary so that the length and content of the utterances would not influence the outcome of the classification. The input feature of the *PeakSlope* parameter was produced by calculating the mean of the minima of the curve of values per utterance. To indicate the presence of creak in an utterance, a *creak rate* feature was introduced, based on the creak decision values in each utterance. The *creak rate* was calculated by dividing the number of creaky segments by the number of voiced segments in an utterance. This method proved to be suitable in making up for the length differences across utterances. The third input feature was the mean fundamental frequency over an utterance. The f_0 values were extracted using the ESPS pitch tracker get f_0 [11].

4.3. Selecting the target voice qualities

We targeted a voice style produced using a tense voice quality and a relatively low f_0 with occasional creaky segments. Based on this limited knowledge, an initial dataset was selected by choosing the utterances with features located in the tensest 30 % of the overall number of the utterances (indicated by the *PeakSlope* values), and that had features displaying low f_0 and more creaky segments than the median values taken from the whole corpus. 6.9 % (i.e. 207 samples) of the corpus qualified for all of these requirements and 18.4 % (i.e. 552 samples) that represent the opposite speech styles (i.e. fulfilling none of the three criteria). Those two groups are utilized in the training of the classifier ensemble used in the AOE voting approach. The target selection is aimed to be broadened out with similar sounding utterances with the help of the AOE voting.

4.4. AOE voting

Figure 3, shows the generated heat map of the algorithm with the confidence measures c_{fsvm} and c_{gmm} ranging between $[0.99, 0.86]$. It is clearly seen that a peak (marked with a star in Figure 3) of $relA$ is found at $c_{fsvm} = 0.96$ and $c_{gmm} = 0.99$. For this point out of $|cand_{all}| = 370$ we find an overlap of $|cand_{en}| = 150$.

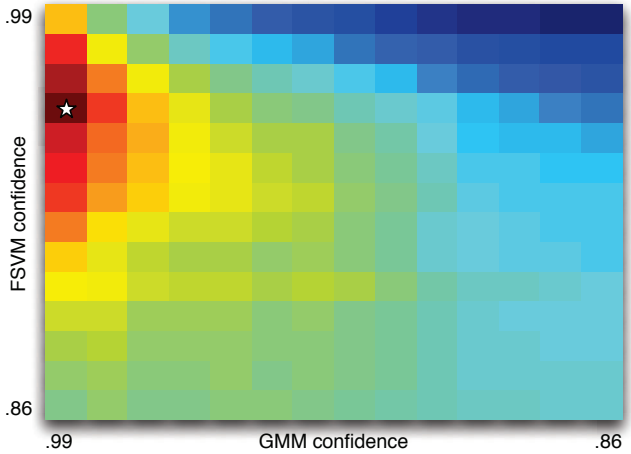

Fig. 3. Ensemble voting heat map. Warm colors indicate good overlap measure and the star indicates the optimal value.

5. PERCEPTUAL EVALUATION

5.1. Stimuli

The goal of the subjective evaluation was three-fold: firstly, we were to assess whether listeners perceive the utterances in the target voice style group to sound similar to each other. The second aim was to find out whether the method selected the vast majority of utterances in the target voice style. Thirdly, we were to show whether there was a significant difference between the training set of target voice style utterances and the group of candidates selected by the AOE voting. The evaluation set consisted of 60 randomly selected utterances: 20 from the **Training set** of the target voice style group, 20 from the samples selected by the **AOE voting**, and 20 from the rest of the corpus (**Other**).

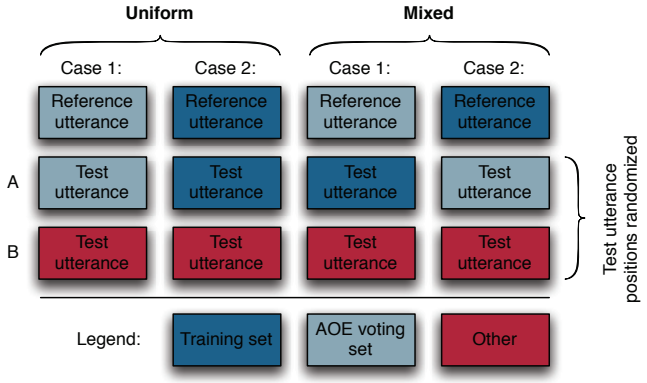

Fig. 4. Illustration of various trial setups in the perception test. There were equal quantities of each case. Both the order of the cases and the A-B position of the test utterances was randomized.

Table 1. *Results from the perception test. Percentages indicate agreement with the classification.*

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

5.2. Experiment

A perception test was carried out in the form of a web application where participants were encouraged to use headphones. The setup of the test is illustrated in Fig. 4. Each participant was presented with 20 sets each containing three utterances (one reference and two test utterances). The four cases, shown in Fig. 4, each appeared in quantities of 5 and were randomly presented to the participant. Furthermore, the A-B position was randomized. The reference sample was taken from either of the target voice style groups. One of the A-B samples originated from the target voice style groups, the other from the rest of the corpus. The listeners were asked to decide which of the two test samples sounded more similar to the reference utterance, in terms of voice characteristics. They were also asked to ignore the length and the content of the utterances. The utterances used in the listening test are available at: http://muster.ucd.ie/∼eva/voicestyletest

6. RESULTS

The subjective evaluation was completed by 27 participants. We found that listeners judgements were in 87 % agreement (see Table 1) with the classification. This shows that the utterances classified as belonging to the target voice style were perceived to be similar to each other and significantly different from the rest of the corpus. These results also show that the method selected the vast majority of utterances in the target voice style because the random selection of test utterances would have detected if there were many target voice style utterances remaining in the "Other" part of the corpus.

In order to test whether there was a perceivable difference between utterances in the training set and the utterances selected by the AOE voting, we considered participant ratings for presented stimuli where the reference utterance and one of the test utterances were either both from the training set or both from the AOE voting set (see Fig. 4). This group was called 'Uniform'. This 'Uniform' group was compared to the 'Mixed' group which was defined as participant ratings for stimuli where the reference utterance came not from the same set as either of test utterances.

Independent t-tests carried out on participant preferences for 'Uniform' and 'Mixed' groups revealed no significant difference ($t = -0.919$, $p = 0.3623$). This indicates that there was no perceivable difference between the training set and the utterances selected by the AOE voting.

7. CONCLUSIONS AND FUTURE WORK

This study describes a method to detect a targeted voice style in an audiobook. The experiment showed a successful separation of a particular voice style from the rest of the corpus. With the help of some prior knowledge in characterizing the desired voice styles in terms of voice quality features, parameters indicating voice quality can be effectively used to find the targeted utterances. Future work will involve including further parameters indicating breathiness and whisper in voice styles, as well as implementing the voice style detection method in expressive speech synthesis.

8. REFERENCES

[1] N. Campbell and P. Mokhtari, "Voice quality: The 4th prosodic dimension," in *ICPhS*, 2003, pp. 2417–2420.

[2] E. Szekely, J. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters," *Proceedings of Interspeech*, 2011.

[3] J. Laver, *The Phonetic Description of Voice Quality*, Cambridge University Press, 1980.

[4] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J Acoust Soc Am*, vol. 87, no. 2, pp. 820–857, 1990.

[5] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," *Proceedings of Interspeech*, 2011.

[6] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *J Acoust Soc Am*, vol. 103, no. 5, pp. 2649–2658, 1998.

[7] C. T. Ishi, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 1, 2008.

[8] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Technical report, International Computer Science Institute and Computer Science Division,, 1998.

[9] C. Thiel, S. Scherer, and F. Schwenker, "Fuzzy-input fuzzy-output one-against-all support vector machines," in *11th KES*. 2007, vol. 3 of *Lecture Notes in Artificial Intelligence*, pp. 156–165, Springer.

[10] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," *Proc. of Interspeech*, 2010.

[11] D.C. Entropic Research Laboratory, Washington, "ESPS version 5.0 programs manual," 1993.