

Clustering Expressive Speech Styles in Audiobooks Using Glottal Source Parameters

Éva Székely, João P. Cabral, Peter Cahill, Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin, Dublin, Ireland

eva.szekely@ucdconnect.ie, {joao.cabral|peter.cahill|julie.berndsen}@ucd.ie

Abstract

A great challenge for text-to-speech synthesis is to produce expressive speech. The main problem is that it is difficult to synthesise high-quality speech using expressive corpora. With the increasing interest in audiobook corpora for speech synthesis, there is a demand to synthesise speech which is rich in prosody, emotions and voice styles. In this work, Self-Organising Feature Maps (SOFM) are used for clustering the speech data using voice quality parameters of the glottal source, in order to map out the variety of voice styles in the corpus. Subjective evaluation showed that this clustering method successfully separated the speech data into groups of utterances associated with different voice characteristics. This work can be applied in unit-selection synthesis by selecting appropriate data sets to synthesise utterances with specific voice styles. It can also be used in parametric speech synthesis to model different voice styles separately.

Index Terms: expressive speech, voice quality, audiobook, speech synthesis

1. Introduction

A high variability in voice characteristics needs to be considered when synthesising speech from highly expressive corpora. In this work we are interested in using audiobooks in which a range of different voice styles are represented. Whether recorded by a professional or not, most read aloud literature contains a significant amount of acting from the reader. Imitating the voice of different characters in a book requires the reader to go beyond prosodic variations of their natural voice and intentionally change their voice quality to mimic different characters or express emotions. In unit-selection speech synthesis, the different voice styles within a corpus, if not handled separately, usually cause distortion in the synthesised speech. Similarly, in statistical parametric speech synthesis, separate modelling of the different voice styles in the corpus contributes to a more accurate modelling of them. One of the challenges in detecting the different voice styles is that the type of variety within a corpus is, to a large extent, dependent on the voice characteristics and the acting performance of the reader. A highly flexible method is therefore needed for an accurate and reliable representation of this variety.

One way of modelling different speaking styles in an expressive speech corpus is to predict from the text, which parts of the audio belong to prompts of certain characters in a book, and treat the speech material from the character roles separately [1]. This is a rational method when dealing with monitored and professionally recorded material. However, the speaker may use the same voice style to represent different characters, especially in the case of freely available audiobooks recorded by

non-professionals. In addition, expression of emotions add an extra layer of variability that is difficult to monitor with text-based methods. Another problem is to find the best size of the speech segments that should be used in separating the different voice styles. It is useful to consider that the segmentation of large audio files on exact sentence level is not a straight-forward task [2]. Moreover, to limit the amount of changes of voice style within an utterance, shorter segments of speech are desirable.

In this work we use a clustering method to group the different expressive voice styles of short speech segments occurring in a corpus using voice quality parameters. The method aims to create clusters of short utterances that are homogeneous in terms of the voice style of the speaker. Due to its unsupervised nature, this method is highly flexible and thus suitable for use on expressive corpora such as audiobooks in which little prior knowledge about the content of the corpus exists. The parameters used to detect the different voice styles in the corpus are the voice quality parameters of the Liljencrants-Fant (LF) acoustic model of the glottal source [3].

2. Glottal Source Parameters

2.1. Liljencrants-Fant Model

The Liljencrants-Fant (LF) model is a popular acoustic model of the glottal source derivative, which is shown in Figure 1. Each pitch cycle of the glottal signal with duration equal to the fundamental period (T_0) starts at the opening instant of the vocal folds, t_o , and ends at the instant of maximum negative amplitude, t_e . The amplitude of the glottal signal is zero when the vocal folds are completely closed (from t_c to the end of the glottal period). The parameter $T_a = t_a - t_p$ measures the abruptness of the closure. The other two parameters are the instant of maximum airflow t_p and the excitation amplitude E_e .

2.2. Voice Quality Parameters of the LF-model

The LF-model can also be described by other parameters which are correlated with voice quality characteristics. The most relevant parameters are the open quotient OQ, speed quotient SQ, and the return quotient RQ, which can be calculated from the basic time domain parameters as follows [4]:

$$OQ = \frac{t_e + T_a}{T_0} \quad (1)$$

$$SQ = \frac{t_p}{t_e - t_p} \quad (2)$$

$$RQ = \frac{T_a}{T_0} \quad (3)$$

Several studies can be found in the literature, e.g. [6, 7], which show the correlation between these parameters and different

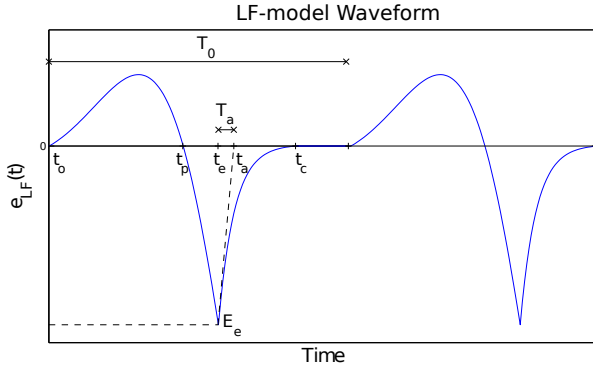


Figure 1: Example of the LF-model waveform which represents the glottal source derivative.

types of voice quality, such as breathy, tense and creaky. The characteristics of these parameters are summarised as follows:

- *OQ*: measures the relative duration of the glottal pulse to T_0 and is related to the pressed-lax dimension of the glottis. For example, *OQ* is typically high for breathy (lack of tension in glottis) and low for tense and creaky.
- *SQ*: relates to the asymmetry of the glottal pulse and it increases with the tension of the vocal folds and vocal effort. For example, *SQ* is usually low for breathy and high for tense and creaky.
- *RQ*: Measures the abruptness of the glottal closure and it tends to increase with the loudness of the voice (vocal effort dimension). For example, it is typically high for breathy and low for tense and creaky.

3. Method for Separating Speech with Different Voice Styles

An overview of the proposed method of clustering utterances on voice quality parameters is shown in Figure 2. The different parts of the method are described in this section.

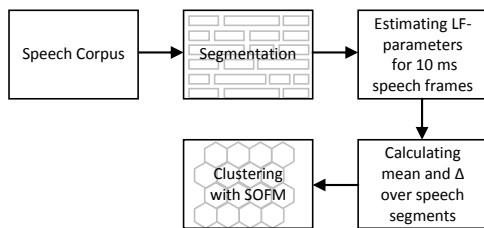


Figure 2: Diagram of the steps involved in the clustering of utterances with different voice styles

3.1. Segmentation

In the process of grouping utterances according to voice style, the size of the utterance is of high importance. Significant voice quality and reading style changes might occur within a single sentence. It is therefore desirable to use shorter segments of speech, where the assumption can be made that the same voice style is used across the whole segment. The segmentation technique used in this work consists of splitting an audio

into smaller chunks with a pause detection method that identifies regions of silence. In order to obtain better results, this process was done with the help of manual tuning: changing the threshold for the length and level of silence after the first segmentation step in cases of longer speech segments where further segmentation is needed. The resulting short utterances do not always fully correspond to prosodic phrases, but abrupt changes of voice style within a speech segment can be avoided.

3.2. Speech Feature Analysis

The parameters T_0 , t_e , T_a and t_p of the LF-model, which was described in Section 2.1, are estimated from the speech signal using the techniques described in [5]. In this method, the glottal source derivative is estimated by inverse filtering (with pre-emphasis) the speech signal using the linear prediction (LP) coefficients. Then, the LF-parameters are estimated from the resulting LP residual using amplitude-based measurements. Finally, the voice quality parameters *OQ*, *RQ* and *SQ* are calculated from the time domain parameters of the LF-model using (1) to (3).

Assuming that the parameters are approximately constant over a short speech segment, the average values of the LF-parameters were used as a meaningful indicator of the speaking style of the actor in that segment. Considering the average voice quality parameter values over a whole segment instead of looking at the values for each individual frame ensures that we focus on the high level changes in voice style (such as cases when the speaker changes his voice quality to imitate a character) and avoid the effect of variations in voice quality due to natural prosody. As the voice quality parameters are defined for voiced speech only, zero values at the unvoiced segments were ignored when calculating the mean.

Apart from the mean values, the measure of the variation of the voice quality parameters within a short-time speech segment can also bear characteristics of a particular speaking style. We therefore calculated the changes of a voice quality parameter (Δ) within one utterance. For example, the Δ of the voice quality parameter *RQ*, is calculated for each speech frame i , as follows:

$$\Delta_{RQ}(i) = |RQ(i) - RQ(i-1)| \quad (4)$$

Accordingly, the input features for the clustering are the following, for each speech segment:

- mean values of the open quotient (*OQ*), the return quotient (*RQ*) and the speed quotient (*SQ*), respectively,
- the mean of the Δ values of the *OQ*, *RQ* and the *SQ*, respectively,
- mean value of the fundamental frequency ($F_0 = 1/T_0$).

3.3. Self-Organising Feature Maps

In this work, the method used to cluster different voice styles in a corpus consists of using Self-Organising Feature Maps (SOFM) [8]. SOFMs are particularly suitable for this task, because they contain both a topology and a distribution. Due to the unsupervised nature of the method, no assumptions need to be made on the input vectors. Each cluster in a SOFM is represented by a neuron with a weight and location in the network. The weight is the physical location, expressed as a vector in the space of features, and the position is defined by the neighbours to which the neuron is connected to in the network.

The application of SOFMs follows three stages: training, mapping and testing, the third of which is optional. In the train-

ing phase, the positions of the neurons in the network are initialised based on a topological function, connecting it with the physically closest neighbours in the network. The network is then presented with input vectors from the training set. Using a competitive learning algorithm, developed by Kohonen in [9], the weight of the neuron is optimised by the smallest Euclidean distance to the input vector, as well as the neighbours of this neuron that lie within a specified area. Through many iterations, the network will cover the topology input space, while retaining a meaningful relation between neighbouring neurons. An additional useful feature of SOFM is that multidimensional features can be illustrated in a two dimensional map.

Figure 4 represents a distance map of the neighbouring clusters. The hexagonals are the clusters resulting from training on the input variables, in this case the voice quality parameters. Connections between clusters that are further apart are represented in darker colours. Outliers in the input data and groups of clusters can be identified using this representation. After clustering, each input vector is associated with its cluster, the size of that cluster, and the location of that cluster in relation to any other cluster, in terms of both Euclidean distance and in number of intermediary clusters.

4. Experiments Using Audiobooks

4.1. Clustering of voice styles

4.1.1. Corpus

For this experiment we used a 50 minute long freely available audiobook from Librivox.org (using 128kbps MP3-file), read by a North-American male speaker [10]. The genre of the audiobook is drama (read by one person only), which ensures a significant amount of variability in voice style, as the speaker expresses emotions and imitates different characters. The corpus was recorded in one session, and was originally one single file. With the pause detection method described in Section 3.1, we segmented the recording into 1736 utterances, where the average duration was 1.6 seconds.

4.1.2. Results

For these utterances, the voice quality parameters were estimated for each 10 ms frame of speech. The F0 values were extracted using the ESPS pitch tracker *get.f0* [11]. The estimated LF-model parameters and the fundamental frequency were used to calculate the 7 parameters described in Section 3.2. Then, they were used for clustering the corpus with the SOFM method described in Section 3.3.

Figure 3 displays the distribution of the utterances across the clusters. The distances between the clusters (hexagonals) are illustrated in Figure 4 using different colors. By looking at the neighbouring distances of the clusters, assumptions can be made about how different the utterances are in terms of voice quality. In this case, the clusters positioned on the upper right side of the network show more similarity to each other, whereas on the lower left side of the network, the clusters are further apart. It can be assumed therefore, that clusters on the upper right side of the network are the ones that contain the utterances spoken with a more neutral voice style, while the rest of the clusters contain the more expressive parts of the corpus. This assumption was confirmed by our informal listening test, where it was clear that the clusters on the upper right contained the more neutral sounding utterances (circled in black). A significant amount of acting takes place in the recording, because of

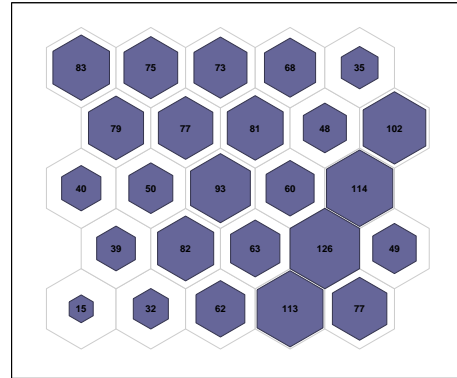


Figure 3: *Distribution of utterances in the SOFM. The numbers correspond to the number of speech segments associated with a cluster.*

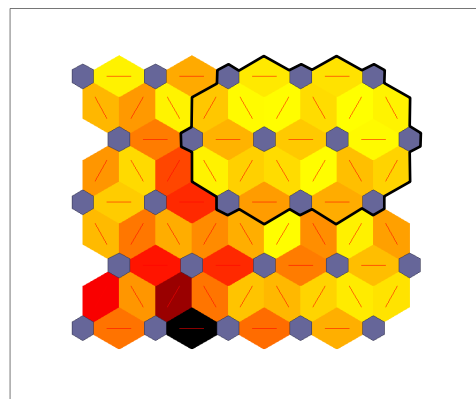


Figure 4: *Neighbouring distances in the SOFM. The larger distances between the clusters are represented with darker colours connecting the hexagonals. The lighter colours represent smaller distances between the clusters.*

the genre of the corpus (drama). Therefore, the more neutral clusters are not significantly larger than the ones containing the different expressive voice styles.

For this corpus, 25 clusters were suitable, but the method is not bound to the number of clusters: they can be changed according to need. By increasing the number of clusters we will have more similar sounding clusters, while lowering the number of clusters will result in more variation within a cluster.

4.2. Separation of Speech from Different Speakers

To provide an objective measure showing that the different clusters stand for different voice styles, an additional experiment was carried out, applying the proposed clustering method for the task of separating speech from two different speakers. For this purpose, two sets of 750 utterances were selected randomly from two different audiobooks, (also originating from Librivox.org) both read by North American male speakers. The same clustering method was applied to these corpora, with the objective of separating the utterances originating from the different speakers. The only difference in the clustering method was that the mean of the fundamental frequency was not included in the input feature vector. The explanation for this is

that the fundamental frequency is not a good feature for differentiating between the two speakers due to the highly expressive nature of the corpora (both audiobooks). That is, the F0 variations due to expressive speech styles are relatively high compared to the F0 range characteristic of the speakers for neutral speech. The SOFM with 25 neurons succeeded in this task of separating the utterances of different speakers with 95.2% accuracy. This accuracy was calculated by assigning each cluster to a speaker and counting the number of utterances attributed to the wrong speaker. The distribution of the utterances of the two speakers across the clusters is shown in Figure 5.

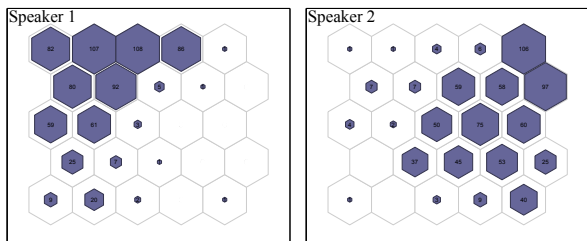


Figure 5: Illustration of the clusters obtained using SOFM for a corpus containing speech from two speakers. The two images show the distribution of the speech segments from each speaker along the clusters.

Similarly to the results presented in Section 4.2.2, the results of this experiment support the hypothesis that the clusters obtained using voice quality parameters over short speech segments do in fact represent differences in voice characteristics.

5. Subjective Evaluation

5.1. Evaluation Design

The goal of the subjective evaluation was to assess whether listeners perceive speech segments from the same cluster to be similar in voice style. In other words, we were interested in how different utterances from different clusters sounded to a listener. Five clusters were selected for the evaluation. The criterion for this selection was to choose clusters that were far apart from each other. This was done by consulting the matrix of neighbouring distances. The evaluation set consisted of ten randomly selected utterances from each of these clusters. The evaluation was an AB listening test. It consisted of 40 trials (20 trials repeated twice in random order). In each trial, the participants were presented with three stimuli: a reference sample and two samples (A and B). One of the A-B samples originated from the same cluster as the reference, the other from a different cluster. The subjects were asked to judge which of the two test utterances 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.

5.2. Results of the Evaluation

The subjective evaluation was completed by 26 participants. The listeners selected the A/B sample from the same cluster as the reference in 81.4% of the cases. The result is statistically significant with a 95% confidence interval of [78.9%, 83.8%] and p -value ≤ 0.0001 . This result shows that the clustering technique used in this work can be used to successfully separate groups of sentences associated with different voice styles. Also, the voice style of the utterances seems to be fairly con-

sistent within a cluster. The utterances used in the listening test are available at: <http://muster.ucd.ie/~eva/Interspeech2011>

6. Conclusions and Future Work

In this paper we grouped utterances from an expressive speech corpus into clusters associated with different voice styles. This was performed using SOFMs for clustering and voice quality parameters of the LF-model as speech features. An objective experiment showed that the mean values of the voice quality parameters and the mean values of the respective deltas over a short speech segment were indicative of the different voice styles in the corpus. This was confirmed by a perceptual evaluation. Results showed that utterances originating from different contexts were often grouped together based on the highly similar voice styles with which they were spoken. This would have been difficult to achieve with text-based classification techniques commonly used when dealing with audiobooks. We conclude that the flexible and unsupervised clustering method handles this side of the problem well. The main application area of the clusters in different voice styles in open source audiobook recordings is speech synthesis. In future work, we plan to implement the information from the voice quality clusters in both unit-selection and statistical parametric speech synthesizers.

7. Acknowledgements

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). 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. Many thanks to Dr. Jie Jiang for providing the tool for segmentation.

8. References

- [1] Zhao, Y., Peng, D., Wang, L., Chu, Min., Chen, Y., Yu, P. and Guo, J., "Constructing stylistic synthesis databases from audio books", Interspeech, 2006.
- [2] Braunschweiler, N., Gales, M. J. F. and Buchholz, S. "Lightly supervised recognition for automatic alignment of large coherent speech recordings", Interspeech, 2010.
- [3] Fant, G., Liljencrants, J. and Lin, Q., "A four-parameter model of glottal flow", STL-QPSR, 26(4), pp. 001-013, 1985.
- [4] Fant, G. and Lin, Q., "Frequency domain interpretation and derivation of glottal flow parameters", STL-QPSR, 29(2-3), pp. 1-21, 1988.
- [5] Cabral, J. P., Renals, S., Richmond, K. and Yamagishi, J., "Towards an Improved Modeling of the Glottal Source in Statistical Parametric Speech Synthesis", Proc. of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, 2007.
- [6] Childers, D. G. and Ahn, C., "Modeling the glottal volume-velocity waveform for three voice types", J. Acoust. Soc. Am., 97(1), pp. 505-519, 1995.
- [7] Gobl, C., "A preliminary study of acoustic voice quality correlates", STL-QPSR, Royal Institute of Technology, Sweden, 1989.
- [8] Bealen M.H., Hagan, M.T. and Demuth, H.B., Neural Network Toolbox, Revised for Version 7.0 (Release 2010b), 2010.
- [9] Kohonen, T., Kaski, S. and Lappalainen, H., "Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM", Neural Computation, vol. 9, no. 6, pp. 1321-1344, 1997.
- [10] <http://librivox.org/one-act-play-collection-001/> accessed on 31.03.2011
- [11] Entropic Research Laboratory, Washington, D.C., ESPS Version 5.0 Programs Manual, Aug., 1993.