# Evaluating expressive speech synthesis from audiobooks in conversational phrases

**Éva Székely, João P. Cabral, Mohamed Abou-Zleikha, Peter Cahill, Julie Carson-Berndsen**

CNGL, School of Computer Science and Informatics, University College Dublin
Dublin, Ireland
{eva.szekely|mohamed.abou-zleikha}@ucdconnect.ie, {joao.cabral|peter.cahill|julie.berndsen}@ucd.ie

**Abstract**

Audiobooks are a rich resource of large quantities of natural sounding, highly expressive speech. In our previous research we have shown that it is possible to detect different expressive voice styles represented in a particular audiobook, using unsupervised clustering to group the speech corpus of the audiobook into smaller subsets representing the detected voice styles. These subsets of corpora of different voice styles reflect the various ways a speaker uses their voice to express involvement and affect, or imitate characters. This study is an evaluation of the detection of voice styles in an audiobook in the application of expressive speech synthesis. A further aim of this study is to investigate the usability of audiobooks as a language resource for expressive speech synthesis of utterances of conversational speech. Two evaluations have been carried out to assess the effect of the genre transfer: transmitting expressive speech from read aloud literature to conversational phrases with the application of speech synthesis. The first evaluation revealed that listeners have different voice style preferences for a particular conversational phrase. The second evaluation showed that it is possible for users of speech synthesis systems to learn the characteristics of a certain voice style well enough to make reliable predictions about what a certain utterance will sound like when synthesised using that voice style.

**Keywords:** audiobooks, expressive speech synthesis, conversational speech

## 1. Introduction

Expressive synthetic speech is a desirable feature in human-robot interaction, speech-to-speech translation and in applications of augmentative and alternative communication. Beyond expressiveness of speech, for such applications, it is desirable for the synthetic voice to suit user preferences of age, gender, accent and character. In order to produce this variety of synthetic voices, large amounts of speech corpora are needed, which require substantial human and financial resources.

In this study, we evaluate the usability of open source audiobooks for the purpose of synthesising conversational speech. Open source audiobooks are a rich, free language resource, and if the expressive variety of voice styles is handled correctly, they can become a very valuable source of expressive speech. Examples of speech synthesis from audiobooks include (Zhao et al., 2006) and (Breuer et al., 2006).

We use the term *voice style* in this work to describe the different ways a speaker produces an utterance in terms of voice quality characteristics e.g. tenseness 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. Due to this variability they cannot be described in the same way as labelled emotional speech. Therefore, the transition between expressive speech styles occurring in read aloud literature and expressive speech styles used in other genres, for example conversational speech, is not straightforward. In this work we evaluate expressiveness of synthetic speech on sentences that commonly occur in conversations and focus only on the expressiveness of conversational speech as a result of the speaker's intention. We do not model other characteristics of conversational speech such as dis-fluencies and pronunciation variation. Perceptual tests featuring utterances of conversational speech are carried out to show the usability of synthetic voices built from audiobooks on the genre of conversational speech.

## 2. Separating the audiobook corpus into subsets of different voice styles

### 2.1. 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 the Blizzard Challenge 2012 by 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 (Braunschweiler et al., 2010). Two of the four available Mark Twain books, *A Tramp Abroad* and *The Man That Corrupted Hadleyburg* were selected for this experiment. This was necessary to eliminate the effect of changes of the recording environment on the resulting synthetic speech. A corpus containing a variety of highly expressive speech styles was formed from the utterances of the books that were no longer than 5 seconds, in order to obtain utterances which have small variation of voice style within the utterance. Based on informal listening tests it was assumed that the vast majority of these utterances did not contain abrupt changes of voice style.

### 2.2. Separation of voice styles

To identify the variety of voice styles in the audiobook corpus, a method described in (Székely et al., 2011) was applied: Self-Organizing Feature Map was used for clustering

with input features calculated from glottal source parameters (Cabral et al., 2007) and fundamental frequency per speech segment. This method has been previously proven to create clusters of in terms of voice style similarly sounding utterances when applied to audiobook recordings containing expressive speech. The audiobook corpus was then separated into 3 subcorpora, featuring broad categories of three different speaking styles. The most extreme expressions of voice styles were excluded from the subcorpora. A summarised perceptual characterisation of these subcorpora is as follows:

- *Subcorpus A*: Soft, lax voice, featuring relatively low pitch ranges
- *Subcorpus B*: Tense, louder voice, with wide pitch ranges
- *Subcorpus C*: Very expressive, intense voice, with mid to high pitch ranges

We have deliberately not given descriptive names to the subcorpora, because the aim of the evaluation is to assess the way in which people use them based on how they sound rather than on whatever explicit knowledge is available about them.

## 3. Evaluation 1

### 3.1. Purpose

The aim of this evaluation was to show that: a) people prefer different voice styles for different utterances, b) there are significant differences amongst individual preferences of the subjects. We used unit-selection speech synthesis to create the expressive speech samples for this evaluation. Unit selection speech synthesis is expected to replicate the original voice quality of the speech corpus well, because it is based on concatenation of units of recorded speech. Therefore, by using unit-selection voices built from the different subcorpora respectively, were able to synthesise speech with significantly different styles.

### 3.2. Voices

Three different voices were created from the subcorpora described in Section 2, using the MUSE Open Source Speech Technology Research Platform (Cahill and Carson-Berndsen, 2010). The labels provided with the voice data were used without any further HMM re-estimation iterations. The system architecture and configuration was the same as was used in the Blizzard Challenge 2011 (Cahill et al., 2011). A classification and regression tree was trained from the labelled voice data to estimate durations using the *wagon* tool in the Edinburgh Speech Tools, and phonetic labels were kept in the supplied X-SAMPA phone set.

### 3.3. Evaluation design

Figure 1 shows the flowchart of Evaluation 1. The subjects were presented with 30 stimuli: 10 sentences in 3 different voice styles. The order of the sentences was randomised. The participants were asked to select the utterance they preferred considering which utterance sounded most appropriate for the content of the sentence in terms of voice style.

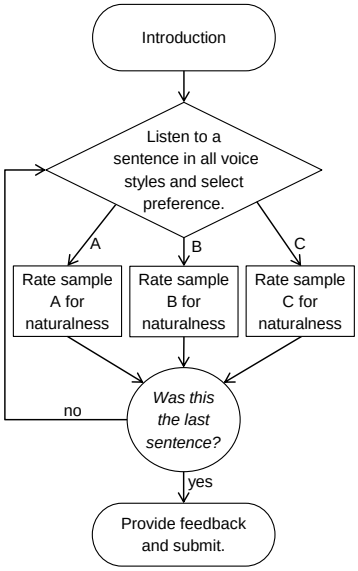


Figure 1: Block diagram of Evaluation 1. *A diagram describing the different steps to guide the participant in the experiment.*

Subsequently, the listeners rated their preferred utterance on a scale of 1-5 for naturalness. After completing the evaluation, the participants answered a question asking whether thought they selected different speaking styles at different times, or they had an overall preference for a particular speaking style. They were also asked whether they would prefer the option of customising the speaking style of a synthetic voice if they were to use one in an application.

### 3.4. Results

The evaluation was completed by 45 participants. Only one participant showed a strong preference to one particular voice (8/10), the remaining subjects selected different voices in at least 4 out of the 10 times. The frequency with which each voice style was preferred is close to one third for each voice. This shows that there are significant differences amongst subject's individual preferences (with the exception of some sentences e.g. s6 and s7). The average rating of the utterances regarding naturalness was 3.6, which means a reasonably good result for expressive speech synthesis. The question about preferring customisable synthetic voices in an application was answered with yes by all of the participants. On Figure 2 it can be seen that samples with higher naturalness ratings have been chosen somewhat more often than samples with low naturalness scores. The slight correlation of frequency of preference and naturalness score could be due to listeners being influenced by the naturalness of utterances when indicating their preference for a particular voice style. This is a limitation of evaluation methods that involve making a choice of preference among several samples. In the next section we introduce an evaluation method that overcomes this limitation.

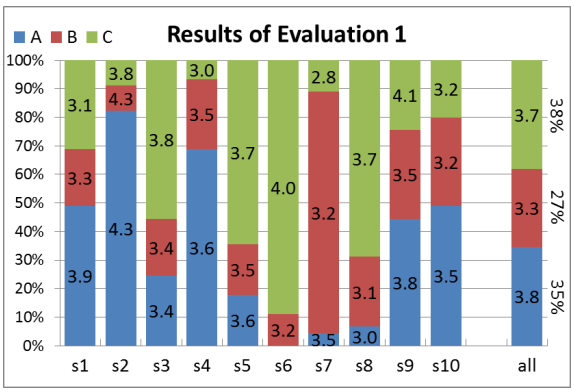


Figure 2: Results of Evaluation 1. *The voices built from the different subcorpora A, B and C are represented in blue, red and green respectively. Each bar represents an evaluated sentence. The colors show how often a particular voice style was preferred for that sentence. The values on each coloured bar show the average naturalness rating of that sample.*

# 4. Evaluation 2

## 4.1. Purpose

The purpose of the second evaluation was to assess the predictability of the synthetic voice styles. This evaluation was built upon the findings of the first evaluation that indicated that the preference for voice styles varied among participants. This experiment was conducted to investigate if after a brief familiarisation with the three synthetic voices subjects were able to predict how a particular utterance was going to sound in a given voice style. It is essential for the usability of expressive speech synthesis in conversational phrases, that the outcome be consistent and roughly predictable. Therefore, we used HMM-based speech synthesis to prepare the expressive speech samples for our second evaluation. HMM-based speech synthesis describes a parametric model of speech by averaging over the acoustic characteristics of similarly sounding speech segments. When we are dealing with expressive speech corpora, this means that the variation within a corpus is smoothed out, making the resulting synthetic speech more consistent and predictable. In other words, one of the main differences to unit selection is that the speech samples within one voice style sound more similar to each other. There is still a clear audible difference amongst the three different voice styles.

## 4.2. Voices

The HTS voices were built using the HTS-2.1 toolkit. The HMM-based speech synthesiser used in the experiment is similar to the speaker-dependent HMM-based speech synthesiser called Nitech-HTS 2005 (Zen et al., 2007). The system uses the STRAIGHT vocoder (Kawahara et al., 1999) to extract the spectrum and aperiodicity parameters from the speech signal, during analysis. Meanwhile, F0 is estimated using the function $get_f0$ from Entropic Speech Tools. For acoustic modelling, the system uses a five-state HMM structure. The F0 parameter vector (including its delta and delta-delta features) is modelled by multi-space probability distribution HMM (MSD-HMM), whereas the spectrum and aperiodicicty streams (including dynamic features) are modelled by HMM using continuous distributions respectively. The F0, spectrum and aperiodicity parameters are clustered using different decision trees, because these parameters have their own contextual factors. During synthesis, the speech parameters are generated from the input sentence and trained HMMs using a parameter generation algorithm based on the maximum likelihood criterion. Finally, the speech waveform is produced from the speech parameters using the STRAIGHT vocoder.

## 4.3. Evaluation design

Figure 3 shows the flowchart of Evaluation 2. The evaluation consisted of two parts: in the first part the subjects were introduced to the three different voice styles by listening to several samples from each voice. The aim of this component was to let the users familiarise themselves with the way the three synthetic voices sound. In the second part they were presented with written stimuli, conversational phrases, and they were asked to select which voice they think would suit this phrase best. In the next step, they listened to the utterance from the selected voice and they were asked to answer a yes or no question whether the expressive synthetic speech sample has met their expectations. In the case of a negative answer they were asked to make a second choice of voice for that utterance and see if that one met their expectations better. In the case of a second negative answer, the participants were asked to rate the last speech sample of that sentence. 23 conversational phrases were used in this evaluation, 5 in the familiarisation part and 18 in the testing part. The phrases were obtained from a collection of context specific conversational messages composed by AAC specialists (University of Nebraska, 2011).

## 4.4. Results

The evaluation was completed by 12 participants. In 71.1% of the cases the subjects answered ”yes” to their first selected voice style. The distribution of the answer patterns is shown in Table 1.

| Answer patterns | Percentage |
|---|---|
| 1st choice / 2nd choice / 3rd choice | |
| yes | 71.0% |
| no / yes | 21.0% |
| no / no / yes | 3.1% |
| no / no / no | 4.9% |

Table 1: Answer patterns in Evaluation 2.

Figure 4 shows the confusion matrix that displays the rate of first choices for the different voice styles compared with the rate of ultimate voice choices. The last column shows for each voice what percentage of the times a voice style was the first choice compared to how often this voice was ultimately chosen. This shows how well the initial choice of the listeners reflects their desired result. Overall, 71% of the desired samples could be selected without previous listening, while 5% of the times a listener could not find

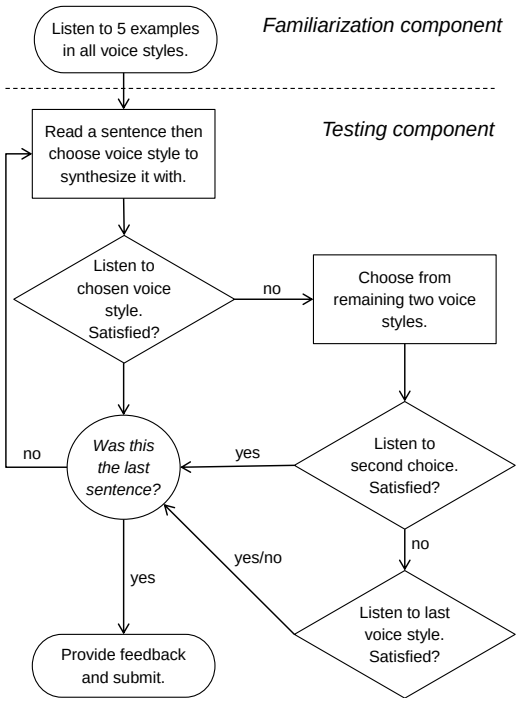


Figure 3: Block diagram of Evaluation 2. *A diagram describing the different steps to guide the participant in the experiment.*

| ultimate positive answer | A | B | C | Overall |
|---|---|---|---|---|
| A | 20.8% | 1.7% | 0.3% | 90.9% |
| B | 5.2% | 24.7% | 6.6% | 67.6% |
| C | 1.4% | 5.9% | 28.5% | 79.6% |
| none | 0.7% | 1.7% | 2.4% | 0% |
| | A | B | C | Overall |
| | first choice of voice style | | | 74.0% |

Figure 4: Results of Evaluation 2.

a suitable voice style. For the A and the C voices 85.7% and 80.3% respectively arrive at the desired voice style on the first attempt. For these voice styles, the listeners are well able to select the desired style on the first attempt. This is confirmed by the low number of people switching to the voice on the opposite side of the spectrum (only 1.2% switched from A to C or vice versa). This strongly supports the hypothesis that listeners are, after some familiarisation, quite capable of deciding what voice style they want to use, without having to listen to the results first. For the B voice some more familiarisation may be necessary, as a number of times this result was derived after first selecting A (6.2%) or C (7.4%). The overall results show that after a very short familiarisation phase, listeners were able to make a reasonably good prediction of how a sample will sound from a particular voice.

## 5. Conclusion and future work

The aim of this study was to evaluate expressive speech synthesis from audiobook subcorpora with different voice styles, and to investigate the suitability of audiobooks as a language resource for expressive speech synthesis used in conversational phrases. The perceptual evaluations show that listeners prefer voice styles of synthetic speech to be customisable. Moreover, after a short listening introduction to the synthetic voices, participants seemed to be able to predict which voice style suited their preferences for a given sentence. Future work involves conducting an evaluation that lets pairs of subjects type into the synthesiser simulating a conversation, in order to examine the frequency with which the different voice styles being used, and the context in which they are preferred.

## 6. 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.

## 7. References

N. Braunschweiler, M. J. F. Gales, and S. Buchholz. 2010. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *INTERSPEECH*, pages 2222–2225. ISCA.

S. Breuer, S. Bergmann, and R. Dragon. 2006. Set-up of a Unit-Selection Synthesis with a Prominent Voice. In *LREC*.

J.P. Cabral, S. Renals, K. Richmond, and J. Yamagishi. 2007. Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In *6th ISCA Workshop on Speech Synthesis*.

P. Cahill and J. Carson-Berndsen. 2010. Muse: An open source speech technology research platform. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 115–120. IEEE.

P. Cahill, U. Ogbureke, J. Cabral, É. Székely, M. Abou-Zleikha, Z. Ahmed, and J. Carson-Berndsen. 2011. UCD Blizzard Challenge 2011 Entry. In *Blizzard Challenge*.

Esps programs version 5.3.

HMM-based speech synthesis system version 2.1, http://hts.sp.nitech.ac.jp, 2008.

H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based $f_0$ extraction: Possible role of a repetitive structure in sounds. In *Speech Communication, Vol. 27, pp. 187–207.*

É. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen. 2011. Clustering expressive speech styles in audiobooks using glottal source parameters. In *INTERSPEECH*, pages 2409–2412. ISCA.

P. Taylor, R. Caley, A. Black, and S. King. 1999. Edinburgh Speech Tools library with documentation edition1.2., Technical report, The University of Edinburgh.

AAC center University of Nebraska. 2011. Context Specific Messages 1-4: Suggested by AAC Specialists. (A. a. University of Nebraska, Ed.) Retrieved 3.10. 2011, from http://aac.unl.edu/vocabulary.html.

H. Zen, T. Toda, M. Nakamura, and K. Tokuda. 2007. Details of Nitech HMM-based Speech Synthesis System for the Blizzard Challenge 2005. In *IEICE Tran., V.E90-D, No1, pp. 325–333*.

Y. Zhao, D. Peng, L. Wang, M. Chu, Y. Chen, P. Yu, and J. Guo. 2006. Constructing stylistic synthesis databases from audio books. In *INTERSPEECH*. ISCA.