Expressive speech synthesis in MARY TTS using audiobook data and EmotionML

Marcela Charfuelan, Ingmar Steiner

DFKI GmbH, Language Technology Lab
Berlin and Saarbrücken, Germany
firstname.lastname@dfki.de

Abstract

This paper describes a framework for synthesis of expressive speech based on MARY TTS and Emotion Markup Language (EmotionML). We describe the creation of expressive unit selection and HMM-based voices using audiobook data labelled according to voice styles. Audiobook data is labelled/split according to voice styles by principal component analysis (PCA) of acoustic features extracted from segmented sentences. We introduce the implementation of EmotionML in MARY TTS and explain how it is used to represent and control expressivity in terms of discrete emotions or emotion dimensions. Preliminary results on perception of different voice styles are presented.

Index Terms: speech synthesis, unit selection, parametric speech synthesis, expressive speech, EmotionML, signal processing

1. Introduction

Audiobooks are becoming a popular and challenging resource for creating synthetic expressive speech [1]. The expressivity variation in audiobooks imposes several challenges, in the sense that although it is rich, it is difficult to handle. In the traditional unit selection technique, for example, it is very difficult to use all audiobook data unless the data has been previously labelled or clustered according to voice style, emotion, impersonated characters, or in general, according to some level of expressivity. Some techniques that have been proposed to split or cluster audiobook data include manual selection of neutral data [2], manual annotation of interpreted characters [3], and unsupervised clustering using prosody descriptors [4] or glottal source features [5].

Whether the audiobook data is split into subsets or not, the next challenge is to create expressive voices that can render several styles, including neutral narrative style. Several approaches can be followed: create or train one voice for each particular style in addition to neutral [6]; create only one voice that is capable of rendering expressive narrative style, using style related context features [7, 8]; or create a universal or neutral model that is adapted to the styles of small expressive sets [4, 5].

The final challenge is, given an arbitrary text or paragraph, how to automatically select an appropriate expressive style for synthesis. This is also related to the problem of representing those styles, emotions or levels of expressivity in a standard way, so that they can be used, for example, in emotion-related applications. One option to tackle this problem is the Emotion Markup Language (EmotionML) standard [9, 10], a World Wide Web Consortium (W3C) candidate recommendation flexible enough to represent emotions and related states. With respect to the challenge of automatically predicting the style of arbitrary text, some techniques used in recognition of emotions in text can be used to train prediction models. For example, linguistic information has been used in [11] to recognise affect in a 3D continuous space. Also, emotional salience of words has been used in [12] to detect emotions in text; in [6] we addressed this issue with sentiment analysis, where different sentiment scores are extracted from audiobook text sentences and used to train a voice style prediction model.

In this paper, we extend the work in [6] to first of all create only one voice capable of rendering several styles, instead of creating several voices in different styles. This idea is in part motivated by the work in [13], where a style context feature is also used, but in our case the style is determined automatically by splitting audiobook data using principal component analysis (PCA) of acoustic features. We show how this style feature can be used to create expressive unit selection and/or HMM-based voices in the MARY TTS framework. Secondly, we extend the current implementation of EmotionML in MARY TTS to encode the audiobook style as the arousal dimension in a 3D pleasure-arousal-dominance (PAD) representation. The rest of the paper is organised as follows. In Section 2, we explain the creation of expressive unit selection and HMM-based voices from audiobook data in the MARY TTS framework. In Section 3, control of expressivity in MARY TTS using EmotionML is introduced. Preliminary listening test results on perception of different voice styles are presented in Section 4 and conclusions are drawn, and future work outlined, in Section 5.

2. Building expressive voices in MARY TTS

The MARY TTS platform is an open-source, modular architecture for building text-to-speech systems, including unit selection and statistical parametric waveform synthesis technologies [14]. The code in the latest release, MARY TTS 5.0, has been thoroughly restructured; the main new features include:

- Simplified installation and voice distribution
- Agile build management and integration of MARY TTS into other projects (using Apache Maven [15])
- New MaryInterface API
- Emotion Markup Language (EmotionML) support

Details about these new features and the new modularised code can be found in the new development repository [16].

In this section, we describe the pre-processing of audiobook data, which is the same for building unit selection or HMM-based voices. The fundamental idea is to use PCA to partition the data into expressive sets and add a style label to each one. Afterwards this style label is used as a context feature for creating unit selection or HMM-based voices.

2.1. Audiobook data

The audiobook data used in this paper is “The Adventures of Tom Sawyer” released in the Blizzard Challenge 2012 [1]. The audiobook data was already split into prosodic phrase level chunks. The sentence segmentation and orthographic text alignment of the audiobook has been performed using an automatic sentence alignment method – LightlySupervised – as described in [17]. From the selected audiobook, we discarded the sentences with confidence value $< 100\,\%$, as well as sentences with more than 30 words. The number of sentences used was 3676, corresponding to 17 chapters and approximately 6.6 h of recordings at 44.1 kHz.

2.2. Data partitioning

As in [18], we used the value of the first principal component (PC1) as a measure of expressivity, after performing a PCA of acoustic features extracted from all audiobook sentences; this measure is used to split the data. As acoustic features, we extracted well-known acoustic correlates of emotional speech: mainly prosody or fundamental frequency (F0) related features, some intonation-related measures (F0 contour measures) and voicing strength features, used in vocoded speech and parametric synthesis to model excitation. The following acoustic features are extracted at frame level and averaged per sentence:

- F0 and F0 statistics; mean, max., min., and range. F0 values were extracted with the Snack Toolkit [19].
- Number of words.
- Average energy, calculated as the short term energy averaged by the duration of the sentence in seconds.
- Voicing rate calculated as the number of voiced frames per time unit.
- F0 contours, as in [20] we extracted slope (a_1), curvature (b_2), and inflection (c_3); these measures are estimated by fitting a first-, second- and third-order polynomial to the voiced F0 values extracted from each sentence:

$$y = a_1 * x + a_0 \tag{1}$$

$$y = b_2 * x^2 + b_1 * x + b_0 \tag{2}$$

$$y = c_3 * x^3 + c_2 * x^2 + c_1 * x + c_0 \tag{3}$$

  These values are calculated for the whole F0 contour and for each voiced region; a mean value is obtained from all voiced regions in a sentence.
- Voicing strengths estimated with peak normalised cross correlation of the input signal [21]. Seven bandpass voicing strengths are calculated, that is, the input signal is filtered into seven frequency bands; mean statistics of these measures are extracted (mean_str1-7) per sentence.

The feature loadings and percentage of variance explained by the first two principal components (PCs) are presented in Table 1. Here we can observe that PC1 explains nearly one third of the variance (0.31%); also according to the loadings we can conclude that in this data voicing_rate and voicing strengths (mean_str1-7) are the features that contribute most to the variance. This is in line with the results obtained in [22] and [4], where better clustering of voice styles in audiobook data is obtained with excitation-related features like glottal source parameters, voicing probability, jitter, and shimmer. It is interesting to note that mean_F0 and F0 contours are not highly loaded in PC1, which might indicate that the audio data analysed contains more variation of speaking styles (voice quality) than extreme emotions.

	PC1		PC2
features	loadings	features	loadings
voicing_rate	−0.272	num_words	−0.407
curvature	0.001	mean_str1	−0.339
mean_curvature	0.012	mean_str3	−0.165
inflexion	0.017	curvature	−0.147
mean_inflexion	0.019	mean_str4	−0.134
slope	0.027	mean_str2	−0.064
num_words	0.034	mean_curvature	−0.057
mean_slope	0.053	mean_inflexion	−0.037
avg_energy	0.101	inflexion	−0.016
min_f0	0.130	mean_str6	0.014
range_f0	0.242	slope	0.044
max_f0	0.259	mean_str5	0.049
mean_f0	0.261	mean_str7	0.054
mean_str1	0.270	range_f0	0.132
mean_str7	0.308	mean_slope	0.148
mean_str4	0.316	max_f0	0.210
mean_str2	0.321	voicing_rate	0.335
mean_str5	0.322	mean_f0	0.351
mean_str3	0.328	min_f0	0.373
mean_str6	0.337	avg_energy	0.418
% Variance	0.31		0.15

Table 1: *PCA of audiobook acoustic features: feature loadings and percentage of variance explained by the first two principal components PC1 and PC2.*

The PC1 of each sentence was calculated and used to split the data into several sets. Quartile statistics of PC1 were used for partitioning the data into the following sets:

$$\textit{veryhigh} \;:\; k2 \times Q_3 <= \mathrm{PC1} \tag{4}$$

$$\textit{high} \;:\; k1 \times Q_3 < \mathrm{PC1} < k2 \times Q_3 \tag{5}$$

$$\textit{centre} \;:\; k1 \times Q_1 <= \mathrm{PC1} <= k1 \times Q_3 \tag{6}$$

$$\textit{low} \;:\; k2 \times Q_1 < \mathrm{PC1} < k1 \times Q_1 \tag{7}$$

$$\textit{verylow} \;:\; \mathrm{PC1} <= k2 \times Q_1 \tag{8}$$

where Q_1 and Q_3 are the first and the third quartiles of PC1, and k_1 and k_2 are constants empirically designed to generate similar densities for levels in the centre and the extremes, where the data is more sparse. Informal listening tests of sentences in the different sets were performed, and perceptual differences were found among the different sets that appear to correspond to variation in the “arousal” dimension, more details on this topic can be found in [6, 18].

2.3. Building voices using a style context feature

The building of unit selection voices or HMM-based voices using the voice import tools in the MARY TTS framework has been described elsewhere [14, 23]. Here, we describe the additional steps that are performed in order to take into account the expressive labels of the data.

As explained above, we split the audiobook data into five sets and labelled the corresponding audio files in each set with prefixes: *veryhigh*, *high*, *center*, *low*, and *verylow*. This information is passed to the `AllophonesExtractor` component of the voice building tools in the form of a style definition configuration file. This component, together with the MARY text analyser, generates MaryXML files [24] containing a style parameter in the prosody element. Later, context features are extracted from each MaryXML file using the

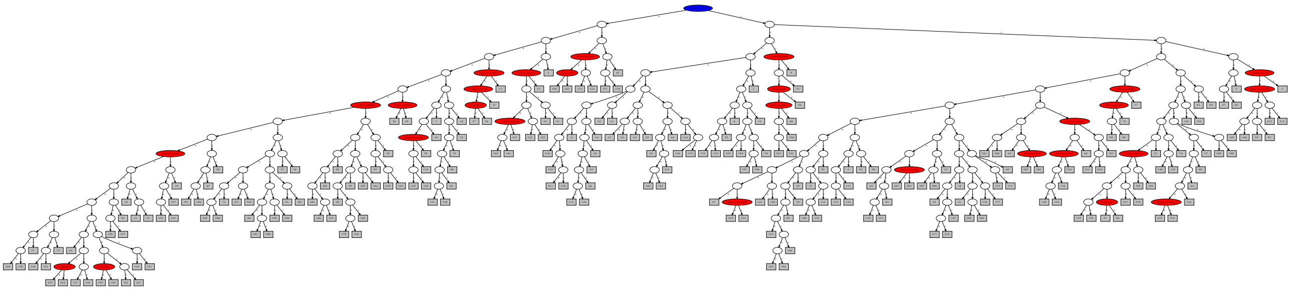

Figure 1: *Voicing strengths decision tree for one state of a HMM-based voice created using a style context feature. Due to the size of this kind of trees, just nodes of questions about style are highlighted in red to show at which level in the tree they are used. Nodes in grey are other context linguistic questions and boxes are leaves which corresponds to PDF distributions.*

`PhoneUnitLabelComputer` and the MARY text analyser. Phone feature vectors are calculated which include context features such as phonological features, linguistic and prosody-related features like part of speech, sentence punctuation, lexical stress, as well as rule-based predictions of tones and break indices (ToBI) accents and phrases. In addition, each phone is assigned a style value which depends on the style of the sentence. Once the acoustic parameters and context features have been extracted according to the voice building procedure described in [23], we continue with the normal voice building procedure.

For unit selection voice building, the acoustic models for prosody prediction are trained by integrating style in addition to the other context features.

For training HMM-based voices, we use the scripts provided by HTS [25]; specifically, MARY TTS 5.0 includes the HTS 2.2 for HTK 3.4.1 training scripts, which have been modified to:

- Use monophone and full context feature labels extracted with the MARY text analyser,
- Generate a questions file for tree building, depending on the MARY context features selected for training the HMMs,
- Generate and use band-pass voicing strengths during training for mixed excitation generation.

As a result of training the HMMs, we have a set of decision trees and their corresponding probability density function (PDF) distributions for fundamental frequency, voicing strengths, and mel-generalised cepstrum (MGC). We have found that the decision trees, created by context clustering, contain more questions related to the extreme styles (*veryhigh* and *verylow*) than the other styles, and style questions mostly appear in fundamental frequency and voicing strengths decision trees. We also observed that the questions about style normally appear early in the tree (cf. the example for voicing strengths in Figure 1), which indicates that the style feature works very well for clustering the data when there are clear acoustic differences among the style sets. This has also been observed in decision trees created using the style-mixed modelling method, where emotional data annotated in several styles and tree-based context clustering is applied to all styles at the same time [13].

For run-time synthesis using HMM-based voices, MARY TTS includes a version of the hts_engine API (1.05) ported to Java. This Java HMM-based synthesiser is fully integrated into MARY TTS and has additional possibilities such as support for explicit prosody specification using the `prosody` element of the Speech Synthesis Markup Language (SSML) W3C recommendation [26].

3. Control of expressivity in MARY TTS using EmotionML

MARY TTS 5.0 implements preliminary support for requesting expressive synthetic speech using EmotionML. The request can be formulated in terms of discrete emotions such as angry, happy, or sad (cf. Listing 1), or in terms of continuous values for emotion dimensions (PAD, cf. Listing 2).

Control of expressivity in terms of discrete emotions is possible for voices that have been trained with data in different emotional styles. Currently, there is an example in the online MARY TTS demo [27] of a male German unit selection voice (dfki-pavoque-styles) with which it is possible to realise angry, happy and sad emotional styles. The current implementation of EmotionML for emotion dimensions allows to map the “big six” or the seventeen “everyday categories” of emotions [28] into the three PAD dimensions. This is done by following the rules described in [29] to render a position in a three-dimensional emotion space. Emotion dimension values are in turn implemented through modifications of pitch contour, pitch level, and speaking rate.

Since the HMM-based voice that we have trained using audiobook data is capable of rendering five arousal levels, we have extended the current implementation of EmotionML to map these levels into the arousal dimension and pass the arousal style as a context feature for realisation. Basically the internal arousal value between [0, 1] is split into 5 thresholds, that will correspond to the five styles, from *verylow* to *veryhigh*. So for example worried, in the “everyday categories”, is internally mapped into arousal=0.2, this means that for rendering this emotion category the *verylow* voice style will be triggered.

Control of expressivity in terms of emotion dimensions is mainly available in HMM-based voices (even if they were not trained with the arousal style), since in the parametric framework it is easier to generate prosody modifications while maintaining the same quality. In unit selection voices, this feature is limited, due to the difficulty of performing prosody modifications without compromising speech quality by introducing artifacts.

4. Voice style perception

As described above, we have created one expressive unit selection voice and another HMM-based voice with the same audiobook data, labelled according to expressive styles. We have performed informal listening tests with the unit selection voice and found that for sentences in the training corpus the system is able to synthesise extreme emotions using EmotionML, but not

```
<emotionml version="1.0" ...
    xmlns="http://www.w3.org/2009/10/emotionml">
  <emotion category-set="http://www.w3.org/TR/
  emotion-voc/xml#everyday-categories">
    <category name="angry"/>
    What was that all about?
  </emotion>
  <emotion>
    <category name="happy"/>
    Nice to see you again!
  </emotion>
  <emotion>
    <category name="sad"/>
    I also had something else in mind than this.
  </emotion>
</emotionml>
```

Listing 1: *EmotionML example where the style is determined by a category name: angry, happy or sad.*

```
<emotionml version="1.0" ...
    xmlns="http://www.w3.org/2009/10/emotionml">
  <emotion dimension-set="http://www.w3.org/TR/
  emotion-voc/xml#pad-dimensions">
      I'm calm.
    <dimension name="arousal" value="0.3"/>
    <dimension name="pleasure" value="0.9"/>
    <dimension name="dominance" value="0.8"/>
  </emotion>
</emotionml>
```

Listing 2: *EmotionML example where the style is determined by a small number of continuous scales in three dimensions: pleasure (or valence), arousal (or activity/activation), and dominance (or control, power, or potency).*

text independent. So in future experiments with unit selection we will need to use more audiobook data to be able to perform a listening test like the one presented below.

For the HMM-based voice, we have performed an experiment similar to the one presented in [6], with the difference that here, we have trained only one voice in different styles and generate samples in extreme emotions using EmotionML. As in the previous experiment, we are interested to know whether users perceive different styles in the samples, in particular extreme ones.

In each trial of the experiment, users were presented with a sentence synthesised in three emotions: pleased, excited and worried. These emotions were chosen because in the internal EmotionML implementation they are mapped into the middle and extreme arousal levels respectively, see Table 2. Users were asked to select from the three synthetic samples the one closest to the original sentence spoken by the audiobook reader. The text sentences used were the same as in the listening test in [6, Table 6], that is, ten sentences from extreme sets, plus ten from the centre. In the experiment, users were also given the opportunity to select "none", if they could not decide and the text was presented on the screen.

Eight users, non-native speakers of English participated in the experiment, two of the listeners are speech experts. The users listened to ten sentences of each style in random order. There was no training phase, so the users were not familiar with the three voice styles before the test, this was to avoid influencing any preference. We have obtained similar results to the ones reported in [6, Table 5 (b)], as can be seen in Table 2. That is, the different styles were perceived by the users, again the extreme styles seem to be easier to identify with 65.0% for veryhigh style and 81.2% for verylow style. These results also confirm the findings in [13], where HMM-based voices are created using style-dependent modeling and style-mixed modeling and the styles are almost equally perceived in both systems. Our results are lower than the ones presented in [13] though, in part because we use audiobook data, not recordings of professional speakers in each particular style; also because we are using the same parametric synthesiser used in the Blizzard Challenge 2012. Currently we are working on improving the quality of our parametric synthesiser by incorporating a glottal source model instead of the current mixed excitation.

	Perceived style by users %			
	Generated emotion			
	excited	pleased	worried	none
PC1 style	(a=0.8)	(a=0.5)	(a=0.2)	
veryhigh	65.0	13.8	2.5	18.8
center	2.5	47.5	47.5	2.5
verylow	2.5	16.2	81.2	0.0

Table 2: *Perception of a style, diagonal agreement: 64.6%. a: Arousal level in the EmotionML "everyday categories".*

5. Conclusions

We have described a framework for synthesis of expressive speech based on MARY TTS and EmotionML. We have explained how an expressive style label on the data can be used to create expressive unit selection or HMM-based voices in the MARY TTS framework, and how that expressivity can be represented in terms of EmotionML. We introduce the implementation of EmotionML in MARY TTS and explain how it is used to represent and control expressivity in terms of discrete emotions or emotion dimensions.

We have also proposed a method for splitting and labelling different expressive styles in audiobook data using PCA of acoustic features. In this respect we have found that voicing strengths extracted in several bands and voicing rate are very good correlates of expressivity in the audiobook data analysed. The levels of expressivity in which we split the data correspond to variations in the "arousal" dimension, therefore we were able to map the audiobook styles to the arousal dimension of EmotionML. The procedure we use in MARY TTS to create expressive voices, using an expressive or emotion label, is general enough to be used with explicitly recorded data in expressive emotions or audiobook data labelled according to styles using other clustering methods.

In the listening test we obtained similar results as in [13], confirming that expressive HMM-based voices can be created separately using data in diferent styles, as we have done in [6], or using all the data in several styles at the same time, as we have done in this paper. In the case of audiobook data, the important issue is to be able to separate (acoustically) clear style clusters in the data, so the questions about style appear early in the decision trees.

In future work we will use more audiobook data to make experiments with unit selection synthesis; we also have plans to incorporate glottal source parameters into the HMM-based synthesis and into the clustering of audiobook data.

6. Acknowledgements

This work is supported by the EU project SSPNet (FP7/2007-2013). The authors would like to thank Marc Schröder for the restructuring and implementation of the new features in the MARY TTS 5.0 system.

7. References

[1] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Blizzard Challenge Workshop*, Portland, OR, USA, Sep. 2012.

[2] N. Braunschweiler and S. Buchholz, "Automatic sentence selection from speech corpora including diverse speech for improved hmm-tts synthesis quality," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1821–1824.

[3] L. Wang, Y. Zhao, M. Chu, Y. Chen, F. K. Soong, and Z. Cao, "Exploring expressive speech space in an audio-book," in *Speech Prosody*, Dresden, Germany, May 2006, p. 182.

[4] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. F. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4009–4012, doi:10.1109/ICASSP.2012.6288797.

[5] É. Székely, T. G. Csapó, B. Tóth, P. Mihajlik, and J. Carson-Berndsen, "Synthesizing expressive speech from amateur audio-book recordings," in *IEEE Spoken Language Technology Workshop (SLT)*, Miami, Florida, USA, Dec. 2012, pp. 297–302, doi:10.1109/SLT.2012.6424239.

[6] M. Charfuelan, "MARY TTS HMM-based voices for the Blizzard Challenge 2012," in *Blizzard Challenge Workshop*, Portland, OR, USA, Sep. 2012.

[7] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM entry for Blizzard Challenge 2012: Hybrid approach," in *Blizzard Challenge Workshop*, Portland, OR, USA, Sep. 2012.

[8] S. Takaki, K. Sawada, K. Hashimoto, K. Oura, and K. Tokuda, "Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2012," in *Blizzard Challenge Workshop*, Portland, OR, USA, Sep. 2012.

[9] M. Schröder, P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter, and E. Zovato, "EmotionML – an upcoming standard for representing emotions and related states," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer, 2011, vol. 6974, pp. 316–325.

[10] P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter, and E. Zovato. (2012) Emotion markup language (EmotionML) 1.0. [Online]. Available: http://www.w3.org/TR/emotionml/

[11] B. Schuller, "Recognizing affect from linguistic information in 3D continuous space," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 192–205, Oct.-Dec. 2011, doi:10.1109/T-AFFC.2011.17.

[12] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, Mar. 2005, doi:10.1109/TSA.2004.838534.

[13] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, pp. 502–509, Mar. 2005, doi:10.1093/ietisy/e88-d.3.502.

[14] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, "Open source voice creation toolkit for the MARY TTS platform," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 3253–3256.

[15] Sonatype, Inc., *Maven: The Definitive Guide*. O'Reilly, 2008.

[16] MARY TTS development repository. [Online]. Available: https://github.com/marytts/marytts

[17] N. Braunschweiler, M. J. F. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Interspeech*, Makuhari, Japan, Sep. 2010, pp. 2222–2225.

[18] M. Charfuelan and M. Schröder, "Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives," in *4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES3)*, Istanbul, Turkey, May 2012, pp. 99–103.

[19] K. Sjölander. The Snack sound toolkit. [Online]. Available: http://www.speech.kth.se/snack/

[20] C. Busso, S. Lee, and S. S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009, doi:10.1109/TASL.2008.2009578.

[21] W. C. Chu, "Mixed excitation linear prediction," in *Speech coding algorithms: Foundations and Evolution of Standardized Coders*. Wiley, 2003, ch. 17, pp. 454–485.

[22] É. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 2409–2412.

[23] MARY TTS. VoiceImportTools tutorial. [Online]. Available: https://github.com/marytts/marytts/wiki/VoiceImportToolsTutorial

[24] MaryXML. [Online]. Available: http://mary.dfki.de/documentation/maryxml

[25] K. Tokuda, K. Oura, K. Hashimoto, S. Shiota, S. Takaki, H. Zen, J. Yamagishi, T. Toda, T. Nose, S. Sako, and A. W. Black. HMM-based Speech Synthesis System (HTS). [Online]. Available: http://hts.sp.nitech.ac.jp/

[26] P. Baggia, P. Bagshaw, M. Bodell, D. Z. Huang, L. Xiaoyan, S. McGlashan, J. Tao, Y. Jun, H. Fang, Y. Kang, H. Meng, W. Xia, X. Hairong, and Z. Wu. (2010) Speech synthesis markup language (SSML) version 1.1. [Online]. Available: http://www.w3.org/TR/speech-synthesis11/

[27] MARY TTS demo. [Online]. Available: http://mary.dfki.de:59125/

[28] K. Ashimura, P. Baggia, F. Burkhardt, A. Oltramari, C. Peter, and E. Zovato. Vocabularies for EmotionML. [Online]. Available: http://www.w3.org/TR/emotion-voc/

[29] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," Ph.D. dissertation, Saarland University, 2004.