A COMPARATIVE EVALUATION OF VOCODING TECHNIQUES FOR HMM-BASED LAUGHTER SYNTHESIS

Bajibabu Bollepalli[1]*, Jérôme Urbain*[2]*, Tuomo Raitio*[3]*, Joakim Gustafson*[1]*, Hüseyin Çakmak*[2]

[1]Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
[2]TCTS Lab – University of Mons, Belgium
[3]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

ABSTRACT

This paper presents an experimental comparison of various leading vocoders for the application of HMM-based laughter synthesis. Four vocoders, commonly used in HMM-based speech synthesis, are used in copy-synthesis and HMM-based synthesis of both male and female laughter. Subjective evaluations are conducted to assess the performance of the vocoders. The results show that all vocoders perform relatively well in copy-synthesis. In HMM-based laughter synthesis using original phonetic transcriptions, all synthesized laughter voices were significantly lower in quality than copy-synthesis, indicating a challenging task and room for improvements. Interestingly, two vocoders using rather simple and robust excitation modeling performed the best, indicating that robustness in speech parameter extraction and simple parameter representation in statistical modeling are key factors in successful laughter synthesis.

Index Terms— Laughter synthesis, vocoder, mel-cepstrum, STRAIGHT, DSM, GlottHMM, HTS, HMM

1. INTRODUCTION

Text-to-speech (TTS) synthesis systems have already reached high degree of intelligibility and naturalness, and they can be readily used in reading aloud a given text. However, applications such as human-machine interaction and speech-to-speech translation require that the synthetic speech includes more expressiveness and conversational characteristics. To bring expressiveness into speech synthesis systems, it is not sufficient to only concentrate on improving the verbal signals alone, since non-verbal signals also play an important role in expressing emotions and moods in human communication [1].

Laughter is one such non-verbal signal playing a key role in our daily conversations. It conveys information about emotions and fulfills important social functions, such as back-channeling. Integrating laughter into a speech synthesis system can bring the synthesis closer to natural human conversation [2]. Hence, the research on analysis, detection, and synthesis of laughter signals has seen a significant increase in the last decade. In this paper, we focus on acoustic laughter synthesis, and explore the role of vocoder techniques in statistical parametric laughter synthesis.

The paper is organized as follows. Section 2 gives the background of work done in laughter processing and laughter synthesis in particular. Section 3 describes the different vocoders compared in this work. Section 4 focuses on the perceptual evaluation experiment carried out to compare the vocoders in their capabilities to produce natural laughter. The results of these experiments are discussed in Section 5. Finally, Section 6 presents the conclusions of this work.

The research leading to these results has received funding from the Swedish research council project InkSynt (VR #2013-4935) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 270780 (ILHAIRE) and n° 287678 (Simple[4]All). H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

2. BACKGROUND

In the last decade, a considerable amount of research has been done on the analysis and detection of laughter (see e.g. [3]), whereas only a few studies have been conducted for synthesis. The characteristics of laughter and speech are slightly different. Formant frequencies in laughter have been reported to correspond to those of central vowels in speech, but acoustic features like fundamental frequency (F_0) has been shown to have higher variability in laughter than in speech [4]. Importantly, the proportion of fricatives in laughter has been reported to be about 40–50% [5], which is much higher than in speech. Despite the differences, the same speech processing algorithms have been applied for laughter analysis as for speech analysis.

As the acoustic behavior of laughter is different from speech, it is relatively easy to discriminate laughter from speech. Classification usually depends upon various machine learning methods, such as Gaussian mixture models (GMMs), support vector machines (SVMs), multi-layer perceptrons (MLPs), or hidden Markov models (HMMs), which all use traditional acoustic features (MFCCs, PLP, F_0, energy, etc.). Equal error rates (EER) vary between 2% and 15% depending on the data and classification method used [6, 7, 8].

On the other hand, acoustic laughter synthesis is an almost unexplored domain. In [9], Sundaram and Narayanan modeled the temporal behaviour of laughter using the principle of a damped simple harmonic motion of a mass-spring model. Laughs synthesized with this method were perceived as non-natural by naive listeners (average naturalness score of 1.71 on a 5-point Likert scale [10]. ranging from 1 (very poor) to 5 (excellent)). Lasarcyk and Trouvain [11] compared two laughter synthesis approaches: articulatory synthesis resulting from a 3D modeling of the vocal organs and diphone concatenation (obtained from a speech database). The 3D modeling led to the best results, but laughs could still not compete with natural human laughs in terms of naturalness. Recently two other methods have been proposed. Sathya et al. [12] synthesized voiced laughter bouts by controlling several excitation parameters of laughter vowels: pitch period, strength of excitation, amount of frication, number of laughter syllables, intensity ratio between the first and the last syllables, duration of fricative and vowel in each syllable. The synthesized laughs reached relatively high scores in perceived quality and acceptability, with values around 3 on a scale ranging from 1 to 5. However, it must be noted that no human laugh was

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

included in the evaluation, which might have had a positive influence on the scores obtained by the synthesized laughs (as there is no "perfect" reference to compare with in the evaluation). Also, the method only enables the synthesis of voiced bouts (there is no control over unvoiced laughter parts). Finally, Urbain et al. [13] used HMMs to synthesize laughs from phonetic transcriptions, similar to the traditional methods used in statistical parametric speech synthesis. Models were trained using the HMM-based speech synthesis system (HTS) [14] on a range of phonetic clusters encountered in 64 laughs from one person. Subjective evaluation resulted in an average naturalness score of 2.6 out of 5 for the synthesized laughs.

From this brief review of the literature, it is clear that the research on HMM-based laughter synthesis is scarce – there exists only one study on HMM-based laughter synthesis using a single vocoder. In this work, we report the role of four state-of-the-art vocoders commonly used in statistical parametric speech synthesis for the application of HMM-based laughter synthesis.

3. VOCODERS

The following vocoders were chosen for comparison: 1) Impulse train excited mel-cepstrum based vocoder, 2) STRAIGHT [15, 16] using mixed excitation, 3) Deterministic plus stochastic model (DSM) [17], and 4) GlottHMM vocoder [18]. All the vocoders use the source-filter principle for synthesis, and thus there are two components that mostly differ among the systems: the type of spectral envelope extraction and representation, and the method for modeling and generating the excitation signal. The vocoders are depicted in Table 1 and described in more detail in the following sections.

3.1. Impulse train excited mel-cepstral vocoder

The impulse train excited mel-cepstrum based vocoder (denoted in this work as MCEP) describes speech with only two acoustic features: F_0 and speech spectrum. The speech spectrum is estimated using the algorithm described in [19]. Mel-cepstral coefficients are commonly used as the spectral representation of speech as they provide a good approximation of the preceptually relevant speech spectrum. By changing the values of α (frequency warping) and γ (factor defining generalization between LP and cepstrum), various types of coefficients for spectral representation can be obtained [19]. Here, we use $\alpha = 0.42$ and $\gamma = 0$ which correspond to simple mel-cepstral coefficients. Both F_0 and mel-cepstrum are estimated using the pitch function in speech signal processing toolkit (SPTK) [20], which uses the RAPT method [21]. Speech is synthesized by exciting the mel-generalized log spectral approximation (MGLSA) filter [22] with either simple impulse train for voiced speech or white noise for unvoiced speech. This simple excitation method has an effect that the synthesized signal often sounds buzzy.

System	Parameters	Excitation
MCEP	mcep: 35 + F_0: 1	Impulse + noise
STRAIGHT	mcep: 35 + F_0: 1 band aperiodicity: 21	Mixed excitation + noise
DSM	mcep: 35 + F_0: 1	DSM + noise
GlottHMM	F_0: 1 + Energy: 1 + HNR: 5 + source LSF: 10 + vocal tract LSF: 30	Stored glottal flow pulse + noise

Table 1. Vocoders in test and their parameters and excitation type.

3.2. STRAIGHT

STRAIGHT [15, 16] was proposed mainly for the high quality analysis, synthesis, and modification of speech signals. However, more often STRAIGHT is used as a reference for comparing between different vocoders in HMM-based speech synthesis, since it is the most widely used vocoder, is robust and can produce synthetic speech of good quality [23]. STRAIGHT decomposes the speech signal into three components: 1) spectral features extracted using pitch-adaptive spectral smoothing and represented as mel-cepstrum, 2) band-aperiodicity features which represent the ratios between periodic and aperiodic components of 21 sub-bands, and 3) F_0 extracted using instantaneous-frequency-based pitch estimation. In synthesis, STRAIGHT uses mixed excitation [24] in which impulse and noise excitations are mixed according to the band-aperiodicity parameters in voiced speech. The excitation of unvoiced speech is white Gaussian noise. Overlap-add is used to construct the excitation, which is then used to excite a mel log spectrum approximation (MLSA) filter [25] corresponding to the STRAIGHT mel-cepstral coefficients.

3.3. Deterministic plus stochastic model (DSM)

The deterministic plus stochastic model (DSM) of the residual signal [26] first estimates the speech spectrum, and uses the inverse of the filter to reveal the speech residual. Glottal closure instant (GCI) detection is used to extract individual GCI-centered residual waveforms, which are further resampled to fixed duration. The residual waveforms are then decomposed into the deterministic and stochastic parts in frequency domain, separated by the *maximum voiced frequency* F_m fixed at 4 kHz. The deterministic part is computed as the first principal component of a codebook of residual frames centered on glottal closure instants and having a duration of two pitch periods. The stochastic part consists of a white Gaussian noise filtered with the linear prediction (LP) model of the average high-pass filtered residual signal, and time-modulated according to the average Hilbert envelope of the stochastic part of the residual. White Gaussian noise is used as excitation for unvoiced speech. The DSM excitation is then passed through the MGLSA filter. The DSM vocoder has been shown to reduce buzziness and to achieve comparable synthesis quality as that of STRAIGHT [26]. DSM vocoder was also used in the previous HMM-based laughter synthesis work [13]. In this paper, STRAIGHT is used to extract F_0 and mel-cepstrum for the DSM analysis, but the extraction of voice source features and synthesis is performed using the DSM vocoder.

3.4. GlottHMM

The GlottHMM vocoder uses glottal inverse filtering (GIF) in order to separate the speech signal into the vocal tract filter contribution and the voice source signal. Iterative adaptive inverse filtering (IAIF) [27] is used for the GIF, inside which LP is used for the estimation of the spectrum. IAIF is based on repetitively estimating and canceling the vocal tract filter and voice source spectral contribution from the speech signal. The output of the IAIF are the LP coefficients, which are converted to line spectral frequencies (LSF) [28] in order to achieve a better parameter representation for the statistical modeling, and the voice source signal that is further parameterized into various features. First, pitch is estimated from the voice source signal using autocorrelation method. Harmonic-to-noise ratio (HNR) of five frequency bands is estimated by comparing the upper and lower smoothed spectral envelopes constructed from the harmonic peaks and the interharmonic valleys, respectively. In addition, the voice source spectrum is estimated with LP and converted to LSFs.

In synthesis, a pre-stored natural glottal flow pulse is used for creating the excitation. First, the pulse is interpolated to achieve a desired duration according to F_0, scaled in energy, and mixed with noise according to the HNR measures. The spectrum of the excitation is then matched to the voice source LP spectrum, after which the excitation is fed to the vocal tract filter to create speech.

4. EVALUATION

A subjective evaluation was carried out to compare the performance of the 4 vocoders in synthesizing natural laughs. For each vocoder, two types of samples were used: a) copy-synthesis, which consists of extracting the parameters from a laugh signal and re-synthesizing the same laugh from the extracted parameters; b) HMM-based synthesis, where HMM-based system is trained from a laughter database and laughs are then synthesized using the models and the original phonetic transcriptions of a laughter. Copy-synthesis can be seen as the theoretically best synthesis that can be obtained with a particular vocoder, while HMM-based synthesis shows the current performance that can be achieved when synthesizing new laughs. Human laughs were also included in the evaluation for reference.

Our initial hypotheses were the following:

- H1: Human laughs are more natural than copy-synthesis and HMM laughs.
- H2: Copy-synthesis laughs are more natural than HMM laughs, as they omit the modeling stage.
- H3: All vocoders are equivalent for laughter synthesis.

The third hypothesis concerns the comparison of the vocoders among themselves, which is the main objective of this work. The way this hypothesis is formulated illustrates the fact that we do not have a priori expectations that one vocoder would be better suited for laughter than other vocoders.

4.1. Data

For the purpose of this work, two voices from the AVLaughterCycle database [29] were selected: a female voice (subject 5, 54 laughs) and a male voice (subject 6, the same voice as in previous work [13], 64 laughs). As in [13], phonetic clusters were formed by grouping acoustically close phones found in the narrow phonetic annotations of the laughs [30]. This resulted in 10 phonetic clusters used for synthesis: 3 for consonants (nasals, fricatives and plosives), 4 for vowels (ə, a, ɪ and o), and 3 additional clusters were formed with typical laughter sounds: grunts, cackles, and nareal fricative (noisy airflow expelled through the nostrils). Inhalation and exhalation phones are distinguished and form separate clusters. Hence there are 20 clusters in total when considering both inhalation and exhalation clusters. For each voice, the phonetic clusters that did not have at least 11 occurrences were assigned to a garbage class.

For each voice and each of the considered vocoders and extracted parameters (see Table 1), HMM-based systems were trained using the standard HTS procedure [14, 31] using all the available laughs. For the test, five laughs lasting at least 3.5 seconds were randomly selected for each voice. For each vocoder, these laughs were synthesized from their phonetic transcriptions (HMM synthesis) as well as re-synthesized directly from their extracted parameters (copy-synthesis). The 5 original laughs were also included in the evaluation. This makes a total of 5 (original laughs) + 5 $\times$ 2 (HMM and copy-synthesis) $\times$ 4 (number of vocoders) = 45 laughs in the evaluation set for both voices.

4.2. Evaluation setup

A subjective evaluation was carried out using a web-based listening test, where listeners were asked to rate the quality of synthesized laughter signals on a 5-point Likert scale [10]. Participants were suggested to use headphones, and were then presented one laugh at a time. Participants could listen to the laugh as many times as they wanted and were asked to rate its naturalness on a 5-point Likert scale where only the highest (completely natural) and lowest (completely unnatural) options were labeled. The 45 laughter signals were presented in random order. 18 participants evaluated the male voice while 15 evaluated the female one. All listeners were between 25–35 years of age, and some of them were speech experts.

5. RESULTS

Figure 1 shows the means and 95% confidence intervals of the naturalness ratings for copy-synthesis (right) and HMM synthesis (left) of the male (upper) and female (lower) voices. The pairwise p-values (using the Bonferroni correction) between vocoders are shown in Table 2 for copy-synthesis and in Table 3 for HMM synthesis.

As expected (H1), original human laughs were perceived as more natural than all other laughs (copy-synthesis and HMM). In addition, H2 was also confirmed: for each vocoder, the naturalness achieved with copy-synthesis was significantly higher than with HMM synthesis. The most interesting is the comparison between the vocoders (H3). In copy-synthesis, GlottHMM was rated as less natural than all other vocoders (for both female and male), MCEP and DSM obtained similar naturalness scores, while STRAIGHT was slightly preferred for female laughs (but not for male laughs). This may indicate that STRAIGHT is potentially the most suitable vocoder for laughter synthesis with the female voice, while MCEP, DSM, and STRAIGHT are equivalently good for the male voice. This trend is generally confirmed when looking at HMM-based laughter synthesis (right plots), where it appears that MCEP obtained the best results for the female voice, followed by DSM, STRAIGHT and finally GlottHMM. For the male laughs, DSM achieved the best results, slightly over STRAIGHT and finally MCEP and GlottHMM, which were rated as similar. However, the only statistically significant differences with HMM synthesis were for the female voice with MCEP (significantly more natural than STRAIGHT and GlottHMM) and DSM (significantly better than GlottHMM).

These results indicate that MCEP and DSM are in general good choices for laughter synthesis. Both vocoders use simple parameter representation in statistical modeling: only F_0 and spectrum are

Female	System	DSM	Glott	MCEP	STR	Nat
	DSM	–	**0.006**	1	1	**0**
	Glott	**0.006**	–	**0.04**	**0.002**	**0**
	MCEP	1	**0.04**	–	1	**0**
	STR	1	**0.002**	1	–	**0**
	Nat	**0**	**0**	**0**	**0**	–
Male	**System**	**DSM**	**Glott**	**MCEP**	**STR**	**Nat**
	DSM	–	**0.003**	1	1	**0**
	Glott	**0.003**	–	**0**	**0.002**	**0**
	MCEP	1	**0**	–	1	**0.027**
	STR	1	**0.002**	1	–	**0**
	Nat	**0**	**0**	**0.027**	**0**	–

Table 2. Pairwise p-values between the vocoders copy-synthesis and natural laughs. Statistically significant results are marked in bold.

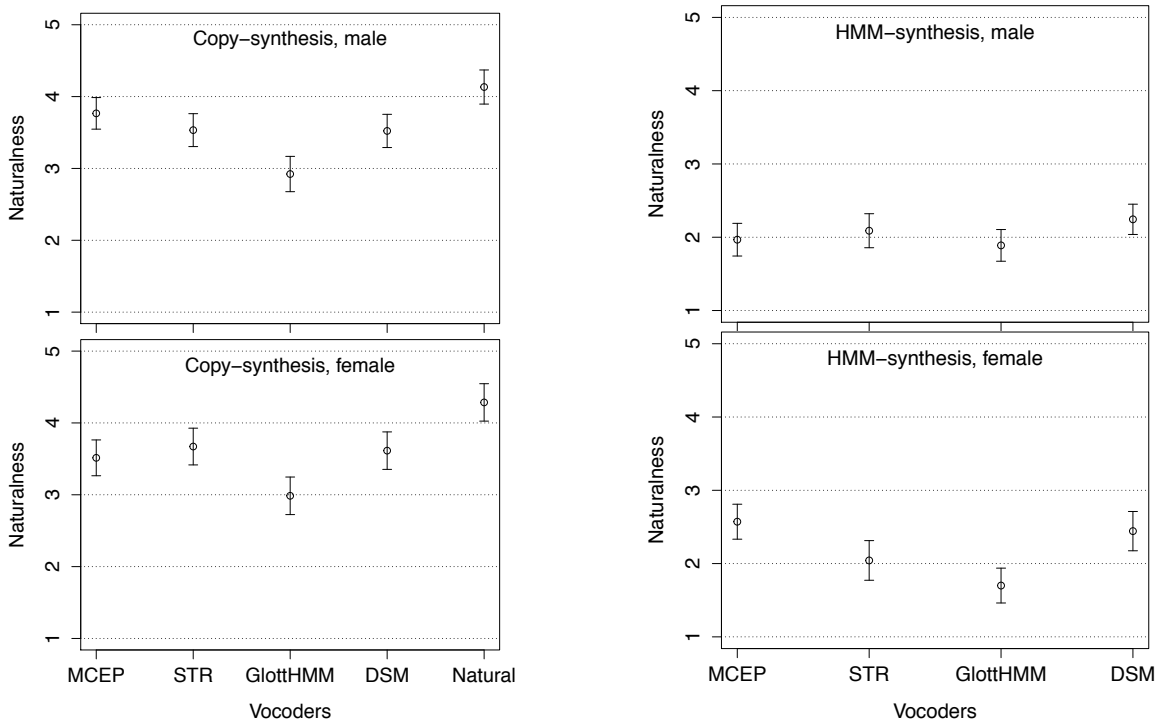

Fig. 1. Naturalness scores for copy-synthesis (left) and HMM synthesis (right) for the male (upper) and female (lower) speakers.

modeled and all other features are fixed. Accordingly, the synthesis procedure of these vocoders is very simple: the excitation generation depends only on the modeled F_0. In DSM, F_m, residual waveform, and noise time envelope are fixed and thus they cannot produce additional artefacts beyond possible errors in F_0 and spectrum. MCEP obtained the best naturalness scores for the female voice, although the known drawback of this method is its buzziness. This was likely not too disturbing as the the female voice used few voiced segments. The buzziness could, however, explain why male laughs synthesized with MCEP were perceived as less natural than female laughs, since the male laughs contained more and longer voiced segments.

STRAIGHT performed better in copy-synthesis with a female voice but cannot hold this advantage in HMM-based laughter synthesis, when statistical modeling is involved. This may well be due to the modeled aperiodicity parameters, which are difficult to estimate from the challenging laughter signals, consisting a lot of partly voiced sounds. Moreover, STRAIGHT pitch estimation is known to be unreliable with non-modal voices (see e.g. [32]), which is very often the case with laughter. Thus, the estimated aperiodicity parameters may have a lot of inconsistent variation, thus degrading the statistical modeling of the parameters. Therefore, in HMM synthesis, the mixed excitation may fail to produce an appropriate excitation.

GlottHMM also suffers occasionally from pitch estimation errors, especially if the voicing settings are not accurately set or speech material is challenging. At least the latter is true with laughter, in which the vocal folds do not reach a complete closure as in modal speech [33]. Pitch estimation errors are even more harmful for the GlottHMM vocoder than the other vocoders since the analysis of voiced and unvoiced sounds is treated completely in a different manner. Thus, voicing errors generate severe errors in the output parameters of GlottHMM. GlottHMM is also considerably more complex than the other systems, thus making the statistical modeling of all the parameters challenging with small amount of data.

Finally, the role of the training material was not studied in this experiment, but it is expected that it also has a significant effect, especially when dealing with challenging material such as laughter.

Female	System	DSM	Glott	MCEP	STR
	DSM	–	**0.003**	1	0.16
	Glott	**0.003**	–	**0**	0.34
	MCEP	1	**0**	–	**0.02**
	STR	0.16	0.34	**0.02**	–
Male	System	DSM	Glott	MCEP	STR
	DSM	–	0.14	0.46	1
	Glott	0.14	–	1	1
	MCEP	0.46	1	–	1
	STR	1	1	1	–

Table 3. Pairwise p-values between HMM synthesis of different vocoders. Statistically significant results are marked in bold.

6. SUMMARY AND CONCLUSIONS

This paper presented an experimental comparison of four vocoders for HMM-based laughter synthesis. The results show that all vocoders perform relatively well in copy-synthesis. However, in HMM-based laughter synthesis, all synthesized laughter voices were significantly lower in quality than in copy-synthesis. The evaluation results revealed that two vocoders using rather simple and robust excitation modeling performed the best, while two other vocoders using more complex analysis, parameter representation, and synthesis suffered from the statistical modeling. These findings suggest that the robustness of parameter extraction and representation is a key factor in laughter synthesis, and increased efforts should be directed on enhancing the robust estimation and representation of the acoustic parameters of laughter.

7. REFERENCES

[1] J. Robson and J. MackenzieBeck, "Hearing smiles-perceptual, acoustic and production aspects of labial spreading," in *Proc. of Inter. Conf. of the Phon. Sci. (ICPhS)*, San Francisco, USA, 1999, pp. 219–222.

[2] N. Campbell, "Conversational speech synthesis and the need for some laughter," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 14, no. 4, pp. 1171–1178, 2006.

[3] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 216–234, 2011.

[4] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *J. Acoust. Soc. Am.*, vol. 110, no. 3, pp. 1581–1597, 2001.

[5] J.-A. Bachorowski and M. J. Owren, "Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect," in *Psychological Science*, 2001, vol. 12, pp. 252–257.

[6] K. P. Truong and D. A. van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Commun.*, vol. 49, pp. 144–158, 2007.

[7] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2973–2976.

[8] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, 2004, pp. 118–121.

[9] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter," *J. Acoust. Soc. Am.*, vol. 121, no. 1, pp. 527–535, 2007.

[10] R. Likert, "A technique for the measurement of attitudes," *Archives of psychology*, 1932.

[11] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," in *Proc. of Interdisciplinary Workshop on the Phonetics of Laughter*, Saarbrücken, Germany, 2007, pp. 43–48.

[12] T. Sathya Adithya, K. Sudheer Kumar, and B. Yegnanarayana, "Synthesis of laughter by modifying excitation characteristics," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3072–3082, 2013.

[13] J. Urbain, H. Cakmak, and T. Dutoit, "Evaluation of hmm-based laughter synthesis," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, Vancouver, Canada, 2013, pp. 7835–7839.

[14] [Online], "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/.

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[16] H. Kawahara, Jo Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

[17] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 20, no. 3, pp. 968–981, 2012.

[18] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "Hmm-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 1, pp. 153–165, 2011.

[19] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – A unified approach to speech spectral estimation," in *Proc. ICSLP*, 1994, vol. 94, pp. 18–22.

[20] [Online], "Speech signal processing toolkit (SPTK) v. 3.6," 2013.

[21] D. Talkin, "A robust algorithm for pitch tracking (rapt)," in *Speech Coding and Synthesis*, W. B. Klein and K. K. Palival, Eds. Elsevier, 1995.

[22] T. Kobayashi, S. Imai, and T. Fukuda, "Mel generalized log spectrum approximation (MGLSA) filter," *Journal of IEICE*, vol. J68-A, no. 6, pp. 610–611, 1985.

[23] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of nitech hmm-based speech synthesis system for the blizzard challenge 2005," in *IEICE Trans. Inf. and Syst.*, 2007, vol. E90-D, pp. 325–333.

[24] T. Yoshimura, K. Tokuda, T. Masuko, and T. Kitamura, "Mixed-excitation for HMM-based speech synthesis," *Proc. Eurospeech*, pp. 2259–2262, 2001.

[25] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, 1992, vol. 1, pp. 137–140.

[26] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 20, no. 3, pp. 968–981, 2012.

[27] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2–3, pp. 109–118, 1992.

[28] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Proc. (ICASSP)*, Mar. 1984, vol. 9, pp. 37–40.

[29] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The AVLaughterCycle database," in *Proc. of Seventh conference on Intl Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010, pp. 2996–3001.

[30] J. Urbain and T. Dutoit, "A phonetic analysis of natural laughter, for use in automatic laughter processing systems," in *Proc. of 4^{th} bi-annual Intl Conf. of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII2011)*, Memphis, Tennesse, 2011, pp. 397–406.

[31] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[32] T. Raitio, J. Kane, T. Drugman, and C. Gobl, "HMM-based synthesis of creaky voice," in *Proc. Interspeech*, 2013, pp. 2316–2320.

[33] Wallace Chafe, *The Importance of not being earnest. The feeling behind laughter and humor.*, vol. 3 of *Consciousness & Emotion Book Series*, John Benjamins Publishing Company, Amsterdam, The Nederlands, paperback 2009 edition, 2007.