SPEECH-LAUGHS: AN HMM-BASED APPROACH FOR AMUSED SPEECH SYNTHESIS

Kevin El Haddad, Stéphane Dupont, Jérôme Urbain, Thierry Dutoit

TCTS lab - University of Mons, Belgium

ABSTRACT

This paper presents an HMM-based synthesis approach for speech-laughs. The building stone of this project was the idea of the co-occurrence of smile and laughter bursts in varying proportions within amused speech utterances. A corpus with three complementary speaking styles was used to train the underlying HMM models: neutral speech, speech-smile, and finally laughter in different articulatory configurations. Two types of speech-laughs were then synthesized: one made by combining neutral speech and laughter bursts, and the other made by combining speech-smile and laughter bursts. Synthesized stimuli were then rated in terms of perceived amusement and naturalness levels. Results show the compound effect of both laughter bursts and smile on both amusement and naturalness and inspire interesting perspectives.

Index Terms— speech-laugh, HMM, speech-smile, synthesis

1. INTRODUCTION

Human-computer interaction through voice systems and virtual conversational agents is becoming more frequent and will eventually become more anchored in our daily habits. Improving those interactions will rely, in particular, on increasing the expressiveness of the voice and the repertoire of vocal sounds that can be generated by the computer. In this work, we propose and evaluate one of the first attempts towards the generation of speech-laughs. Speech-laughs are occurrences of laughter happening during a verbal utterance and intermingled with it, or vice-versa. This feature of our social spoken language communication derives directly from isolated laughter. Both have similar social purposes, although speech-laughs are probably more frequently spontaneous, actually disrupting the speech flow. Besides, according to Trouvain [1], speech-laughs appear in dialogs even more often than isolated laughter. Consequently, giving the computer the ability to make use of speech-laughs when appropriate could increase both the perceived naturalness of the interaction and the emotional engagement of the user [2]. This also presents several interesting challenges in terms of vocal signal processing and modeling.

Previous research on analyzing the acoustic characteristic of speech-laughs can be found in [1], [3], [4] and [5]. These studied the variation of pitch, formants and duration of speech sounds when laughter happens. But to the best of our knowledge, only one attempt was made at synthesizing them. Oh and Wang [6] tried real-time modulation of neutral speech to make it closer to speech-laugh, based on the variation of characteristics such as pitch, rhythm and tempo. No evaluation of the naturalness of that approach has been reported though.

More work can be found on the synthesis of isolated laughter. In [7], laughter was imitated using articulatory as well as diphone concatenation synthesis. Sundaram and Narayanan generated laughter using a parametric model based on a mass-spring system in [8]. In [9], Urbain et al. used Hidden Markov Models (HMM)-based synthesis to generate and synthesize laughter with various degrees of arousal, improving on naturalness compared to the previous state-of-the-art. In [10], Çakmak et al. also used HMM-based models with a forced duration approach to synthesize audio and visual laughter signal. The previous positive HMM-based synthesis attempts encouraged us to apply this statistical parametric modeling and synthesis approach to speech-laughs.

Another important perspective is the question regarding the continuum from smile to laugh. According to [1], people tend to reject the idea of a smile-laugh continuum. Ruch, though, proved in [11], based on facial expression comparisons that enjoyment smiles were involved in laughter. He also stated that there is a smooth transition between laughter and smile. In our paper, we nevertheless consider the co-occurrence of smile and laugh on top of speech signals, with varying proportions of smile and laughter. Speech-laughs are generated by inserting laughter sounds into neutral speech on one side, and into speech-smile (i.e. smiled speech) on the other side. Based on the HMM approach, speech-smiles are obtained after adaptation of a neutral speech GMM acoustic model onto a relatively small amount of smiled speech audio data. Laughter episodes within speech are themselves synthesized using the HMM-based approach, relying on a corpus of recorded laughs covering various vowel articulatory positions. The perceived levels of amusement and naturalness of these signal are then evaluated through mean opinion score (MOS) tests.

The paper is organized as follows. Section 2 presents our system. It first gives a general overview and then details the different HMM models used for synthesis. The evaluation protocol and results are exposed in Sections 3 and 4 respectively. Section 5 concludes and proposes perspectives for future work.

2. SPEECH-LAUGH SYNTHESIS SYSTEM

Speech-laughs are extremely variable and strongly dependent both on individual styles and in the context in which people interact. In consequence, our first work here relies on artificially constructed patterns of speech-laughs based on speech and speech-laughs from a single target person, for which a database has been collected and used for acoustic model training and adaptation.

Our HMM-based speech-laugh synthesis approach relies on first creating unified models for the acoustic features of pitch (F0), spectrum coefficients and phoneme durations during the acoustic model training step [12]. GMM (Gaussian Mixture Models), probability density function calculated as the weighted sum of Gaussian component densities, are being used for this purpose. Acoustic model

This work was partly supported by the Chist-Era project JOKER with contribution from the Belgian Fonds de la Recherche Scientique (FNRS), contract no. R.50.02.14.F. The research leading to these results has also received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n. 270780 (project ILHAIRE).

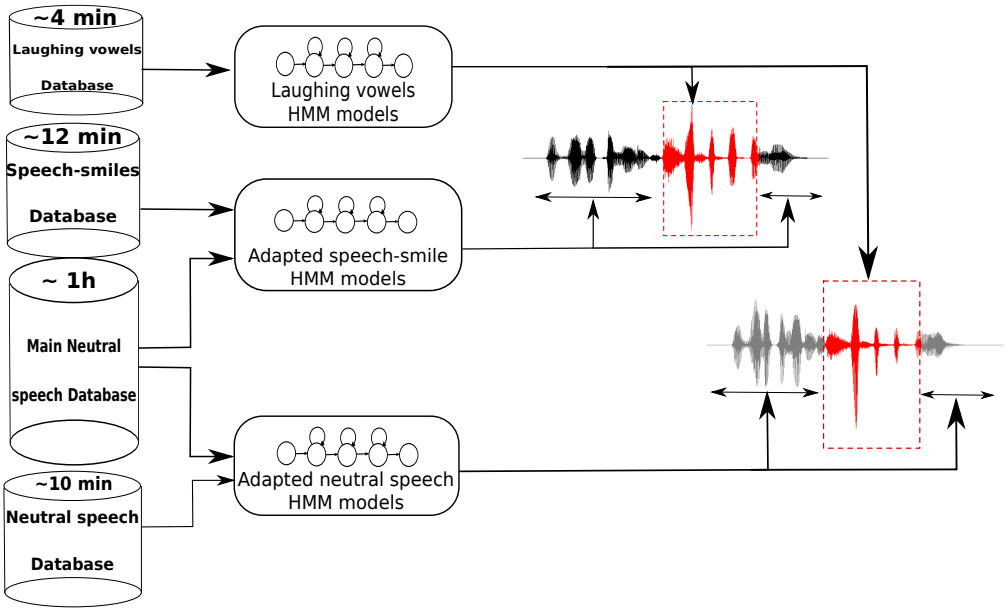

Fig. 1: HMM-based speech-laugh synthesis system pipeline

adaptation is also leveraged. Previously trained models can hence be transformed into adapted models of the target voice through the Constrained Maximum Likelihood Linear Regression (CMLLR) method [13] (a description of the CMLLR adaptation algorithm can be found in [14]). The CMLLR is an adaptation method in which the feature distributions mean and covariance matrices are adapted. CMLLR has been used here to adapt to a target person neutral speech and speech-smile voices.

During the synthesis step, and based on the sequence of phonemes to be produced, the best suited trajectories for the features above are predicted using a maximum-likelihood parameter generation algorithm making use of the HMM model previously made during training [15]. From those trajectories, a synthesizer can generate the waveform.

Both speech (including speech-smile) and laughter HMM models are hence created. Speech-laughs are then generated by replacing some of the vowels within the sequence of phonemes by one or a sequence of laughter "phonemes". Several basis for producing the individual speech and laughter HMM models could be used. Regarding the laughter "component", based on the fact that laughs within speech signals typically happen with articulatory positions similar to the ones of the underlying speech phonemes (at least when they start), we introduced the concept of "laughing vowels". These correspond to laughter occurring during the pronunciation of a vowel. Regarding the speech "component", inspired by the idea of smile, speech and laugh co-occurrence, we constituted models of both neutral speech but also of speech-smile. Note that some previous work has proven speech-smiles are still strongly phonetically recognizable [16].

This mixture of both laughing vowels and speech-smiles creates the effect of laughter mixed with speech and thus, speech-laughs. We will in particular show that the presence of speech-smile (as opposed to neutral speech) has a significant impact on the naturalness of the resulting signal. Fig.1 shows a general overview of our system.

The following sections give more information regarding these different speech-laugh "components", the training/adaption corpora they rely on, and the integration of both to generate speech-laughs.

2.1. Speech model

A main neutral speech model is built from data taken from [17]. In terms of volume, 1356 recorded sentences were used from the main speech database, making a total of 54 min and 23 seconds of speech recording. The recordings were made at 44.1 kHz and stored in a PCM 16 bits. An adapted speech model was then created using acoustic model adaptation of the previously obtained neutral speech model. This was done for compatibility with the target voice from which the laughing vowels and speech-smiles were recorded (see Sections 2.2 and 2.3). Picart used acoustic model adaption to transform a neutral speech model to models enabling the generation of hypo and hyper-articulated speech in [18]. Other examples of HMM-based adaptation systems are given in [19] and [20]. These positive results of acoustic model adaptation applied to speech synthesis encouraged us to use this approach in our system. The adaptation was made on a small database of 200 neutrally read sentences. These sentences were a subset of the ones used to record the previous database, and the recording was made at 48 kHz and stored in PCM 16 bits using a rode Podcaster USB microphone. Their total time is of 10 minutes and 22 seconds making it 19% of the database used for the main neutral speech model. The adapted neutral speech model will be, thus, the one representing the neutral speech cathegory in our tests (see Section 3) and will be referred to as "neutral speech" from now on.

2.2. Speech-smile model

The speech-smile model is created using acoustic model adaptation of the previously obtained main neutral speech model as well. Relying on Picart's results again [18] and considering a speech-smile voice as the target speaker's voice, this method could, with enough data, convert a neutral speech model into speech-smile model.

A speech-smile dataset was collected from the target person (using a rode Podcaster USB microphone at 48 kHz and stored in PCM 16 bits). The subject was instructed to smile while speaking but most importantly to also try to sound "happy". These are opposite of the instructions given in [16] where the author studied the phonetic effect of spread lips while trying to avoid the phonetic effects happiness might add. In fact, evaluations in [21] proved that the perceived degree amusement was better when using Duchenne smiles (smiles containing real enjoyment) data rather than spread lips style of speech data for adaptation.

Eventually, 200 sentences were collected making a total time of 12 minutes and 18 seconds of speech-smile, using a subset of the utterances used to collect the neutral speech database.

2.3. Laughing vowels model

As mentioned earlier, speech-laugh waveforms will be generated by replacing speech or speech-smile vowels with laughing vowels. Therefore, a "laughing vowel" database from the target person was also collected to obtain a laughing vowel HMM model. The subject was instructed to pronounce steady vowels while watching funny videos, breath when necessary, and restart the vowel production as soon as possible afterwards. This triggered laughter during the vowel pronunciation. Recordings of all 16 different vowels of the French language were obtained at 48 kHz and stored in PCM 16 bits. In the end, we used a set of 3 different vowels (the French a, é and i vowels) with a total time of 4 minutes and 35 seconds of laughing vowels sounds (hence 17% from a total of 26 minutes and 41 seconds of actual recording).

Manual transcriptions and segmentations of these signals were made according to the speaker's laughter pattern, and to the annotation scheme proposed in [22] and [11], in which a detailed descriptive structure of laughter episodes is given. See fig.2 for an example of the laughter pattern.

2.4. Speech-laugh model

HMMs were trained for each of the three vowels selected and used to replace some of the vowels contained in the neutral or smiled sentences to be generated. This replacement was made at the level of the context-dependent phonetic transcriptions to be provided as input for synthesis. The speech vowels on which we wished laughter to occur were replaced by their corresponding laughing vowels. The context labels of the speech and laughing vowels full context transcriptions (see the example of full context transcriptions for the English language in [23]) were completely independent. In other words, the context of one of them never contained any information about the other, a consequence of the disjoint nature of the available speech and laughing vowels corpora in terms of left and right phonetic contexts. Thus, the system had to synthesize acoustic feature trajectories like if there was no laughter until it got to the laughing phonemes. Then, it had to synthesize laughter like if there was no speech before or after, until it got again to the speech phonemes. Finally, it had to synthesize the rest of the speech phonemes.

The neutral speech and speech-smile databases on which the adaptations can be made, as well as the laughing vowels database and some examples that were evaluated during the tests (see Section 3) are available on the following address:
`http://tcts.fpms.ac.be/~elhaddad/ICASSP15`
Fig.2 shows an example of a synthesized speech-laugh in which the structure of laughter described in [22] and [11] is well recognized.

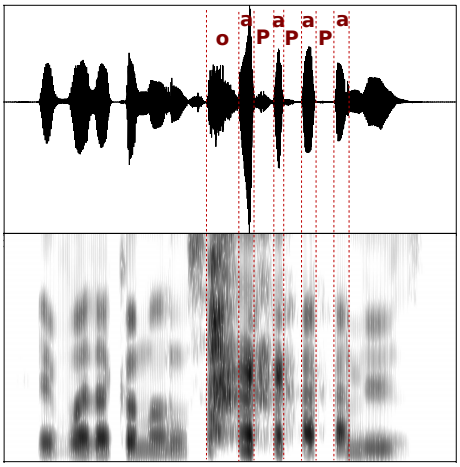

Fig. 2: Spectrogram and waveform pattern of laughing vowel "a" inside a word. o = laughter burst onset [22], a = "a" laughing vowel, P = pulse of air (exhalation).

2.5. Implementation Details

All the recordings of our databases were downsampled to 16 kHz (from 44.1 or 48.0 kHz) before training to have a uniform sampling frequency.

The publicly available HTS (HMM-based Speech Synthesis System) scripts of the adaptation demonstration canvas were used for this work [24]. HTS is a set of speech synthesis tolls delivered as a patch for the HTK (HMM ToolKit) [25]. The HHMs were trained with a left-to-right 5 state configuration using the HTS tools. The filter was modeled using 24 Mel Generalized Cepstral (MGC) coefficients with (α = .42 and γ = 0), together with their dynamic and acceleration coefficients. The state probabilities were estimated through a single Gaussian distribution with diagonal covariance matrix (as usually done in speech synthesis, and related to the fact that the model is single-speaker). The synthesis was eventually made using hts_engine [26] (hts_engine is a software that synthesize speech waveforms from trained HMMs by HTS) [27].

3. AMUSEMENT & NATURALNESS EVALUATION

The participants of this evaluation were asked to grade up to twenty four sentences out of thirty six sentences prepared previously. Among the thirty six sentences, nine were synthesized for each of the four selected styles: SSL (Speech-smile and Laughter), NSL (Neural speech and laughter), neutral speech and speech-smile sentences. Six sentences of each style were randomly selected for each of the twenty eight French speaking test participants. They were asked to grade the proposed sentences on two 5-point scales, covering on one side the degree of amusement, and on the other side the degree of naturalness. The naturalness scale went from 1 labeled unnatural to 5 labeled extremely natural. The amusement scale went from 1 to 5 labeled gradually: serious, neutral, slightly amused, amused and extremely amused. We, thus, counted a total of six hundred and twenty nine ratings for each of the two scales. Results are shown in Figure 3, presenting average ratings as well as standard error over the average estimate.

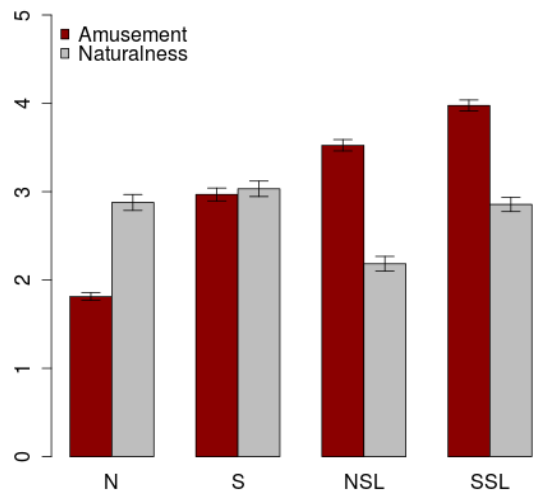

Fig. 3: Mean values barplot and standard error. N=neutral speech, NSL=neural speech and laughter, S=speech-smile, SSL=speech-smile and laughter.

4. RESULTS & DISCUSSION

Several conclusions can be drawn from these results. First, comparing NSL and SSL, we can see that the SSL sentences were better graded on the amusement scale (μ_{SSL} = 3.9748 $>$ μ_{NSL} = 3.5253), as well as on the naturalness scale (μ_{SSL} = 2.8553 $>$ μ_{NSL} = 2.1835). Student's t-tests confirmed that the differences are significant, with p-values $<$ .01 on both scales. Since the SSL communicated amusement better and was perceived as more natural, the co-occurrence of smile and laughter into speech is a better approach to synthesize speech-laughs, compared to the insertion of laughter alone. Student's t-tests were also used to compare the SSL, the speech-smile and the neutral sentences with each other on the naturalness scale. The results showed that the small differences observed in perceived naturalness are not significant. This hence means that the SSL speech-laughs were perceived as natural as the others.

Table 1: Pairwise p-values between the compared methods, on the naturalness scale

	N	S	NSL	SSL
N	1	0.2196	< .01	0.8559
S	0.2196	1	< .01	0.1397
NSL	< .01	< .01	1	< .01
SSL	0.8559	0.1397	< .01	1

Besides, on the amusement scale, the sentences containing laughter were systematically better graded (μ_{NSL} = 3.5253 and μ_{SSL} = 3.9748) than the other sentences (μ_N = 1.8129 and μ_S = 2.9679). This was also true for NSL sentences which had significantly lower scores on the naturalness scale than the neutral and speech-smile sentences. Therefore, whether it sounds natural within the utterance or not, laughter will communicate to the listener the message of amusement it carries. We can also observe that there is an increasing amusement rating when going from neutral speech to speech-smile, then to NSL, and eventually to SSL. Student's t-tests showed a significant difference between each of these approaches (all p-values were $<$ .01).

5. CONCLUSION & FUTURE WORK

In this paper we developed a speech-laugh HMM-based synthesis system considering the co-occurrence of smile and laugh on top of speech signals. To evaluate at the same time our results as a whole and the importance of this co-occurrence, two types of speech-laughs were evaluated: neutral speech and laughter (NSLs) and speech-smile and laughter (SSLs). They were evaluated along with speech-smiles and neutral speech. The evaluation showed that the SSLs were perceived as a more natural type of speech-laugh than the NSL type. It also showed a ranking of the evaluated sentences on the amusement scale with speech-laughs being better perceived, followed by speech-smile and neutral speech.

Inspired by those conclusions, one of our long term objectives will be the development of a real-time amused speech synthesis system with a level control inspired by the MAGE platform [28]. Our shorter term objectives will focus on improving the naturalness of our synthesized sentences. This can be done by using for the excitation signal, the Deterministic plus Stochastic Model (DSM) [29] since it has shown to increase the naturalness for speech, and also gave good results for laughter [9]. The position of laughter bursts inside speech sentences will also be analyzed in order to model and reproduce those. Subjective comments were gathered from the participants after each test. The most common one was: the "weird" feeling given by the pronunciation of what is left of the word or sentence after the laughing vowel occurred. This apparently caused the raters to give lower scores for the naturalness scale. The laughing vowels were indeed inserted in the words arbitrarily. Thus, the context in which speech-laughs are really created and the way laughter is inserted into a sentence is assumed to improve naturalness. This can be done by training HMM models from real speech-laughs, taking into account the context and maybe other features like the intensity of the laugh.

6. ACKNOWLEDGEMENTS

The authors would like to thank N. d'Alessandro, H. Çakmak, A. Moinet and B. Picart for their relevant input in terms of perspectives, and their support in using several software tools.

7. REFERENCES

[1] Jürgen Trouvain, "Phonetic aspects of "speech laughs"," in *Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: L'Harmattan*, 2001, pp. 634–639.

[2] Radosław Niewiadomski, Jennifer Hofmann, Jérôme Urbain, Tracey Platt, Johannes Wagner, Bilal Piot, Huseyin Cakmak, Sathish Pammi, Tobias Baur, Stephane Dupont, et al., "Laugh-aware virtual agent and its impact on user amusement," in *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2013, pp. 619–626.

[3] Klaus J Kohler, ""speech-smile","speech-laugh","laughter" and their sequencing in dialogic interaction," *Phonetica*, vol. 65, no. 1-2, pp. 1–18, 2008.

[4] Menezes Caroline and Yosuke Igarashi, "The speech laugh spectrum," in *Proceedings of the 7th International Seminar on Speech Production (ISSP)*, 2006, pp. 517–524.

[5] S.H. Dumpala, K.V. Sridaran, S.V. Gangashetty, and B. Yegnanarayana, "Analysis of laughter and speech-laugh signals using excitation source information," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 975–979.

[6] Jieun Oh and Ge Wang, "Laughter modulation: from speech to speech-laugh.," in *INTERSPEECH*, 2013, pp. 754–755.

[7] Eva Lasarcyk and Jürgen Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," in *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter*, Saarbrücken, Germany, August 2007, pp. 43–48.

[8] Shiva Sundaram and Shrikanth Narayanan, "Automatic acoustic synthesis of human-like laughter," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 527–535, 2007.

[9] J. Urbain, H. Cakmak, A Charlier, M. Denti, T. Dutoit, and S. Dupont, "Arousal-driven synthesis of laughter," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 2, pp. 273–284, April 2014.

[10] Hüseyin Cakmak, Jérôme Urbain, Joëlle Tilmanne, and Thierry Dutoit, "Evaluation of HMM-based visual laughter synthesis," in *Acoustics Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014.

[11] Willibald Ruch and Paul Ekman, "The expressive pattern of laughter," in *Emotion, qualia and consciousness*, A. Kaszniak, Ed., pp. 426–443. World Scientific Publishers, Tokyo, 2001.

[12] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999.

[13] Vassilios V Digalakis, Dimitry Rtischev, and Leonardo G Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 357–366, 1995.

[14] Junichi Yamagishi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata, and Juri Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained smaplr adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, 2009.

[15] Keiichi Tokuda, Heiga Zen, and Alan W Black, "An HMM-based speech synthesis system applied to english," in *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*. IEEE, 2002, pp. 227–230.

[16] Julie Robson and Beck Janet, "Hearing smiles-perceptual, acoustic and production aspects of labial spreading.," in *XIVth Proceedings of the XIVth International Congress of Phonetic Sciences. Volume 1: 219-222*. International Congress of Phonetic Sciences, 1999, vol. 1, pp. 219–222.

[17] Benjamin Picart, Thomas Drugman, and Thierry Dutoit, "Analysis and HMM-based synthesis of hypo and hyperarticulated speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 687 – 707, 2014.

[18] Benjamin Picart, Thomas Drugman, and Thierry Dutoit, "Continuous control of the degree of articulation in HMM-based speech synthesis.," in *INTERSPEECH*, 2011, pp. 1797–1800.

[19] Junichi Yamagishi and Takao Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[20] Douglas A Reynolds and Richard C Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.

[21] Kevin El Haddad, Stéphane Dupont, Jérôme Urbain, and Thierry Dutoit, "An HMM-based speech-smile synthesis system: An approach for amusement synthesis," in *International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2015)*, in press.

[22] J.Urbain, *Acoustic Laughter Processing*, Ph.D. thesis, University of Mons, 2014.

[23] Heiga Zen, "An example of context-dependent label format for HMM-based speech synthesis in english," *The HTS CMUARCTIC demo*, 2006.

[24] Keiichiro Oura, "Hmm-based speech synthesis system (HTS) [computer program webpage]," http://hts.sp.nitech.ac.jp/, consulted on August, 2014.

[25] Steve J. Young and Sj. Young, "The HTK hidden markov model toolkit: Design and philosophy," in *Entropic Cambridge Research Laboratory, Ltd*, 1994.

[26] Oura Tokuda, "hts_engine [computer program webpage]," *Online: http://hts-engine.sourceforge.net/*, 2011.

[27] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan Black, and Keiichi Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Workshop on Speech Synthesis (SSW-6)*, August 2007.

[28] Maria Astrinaki, Nicolas d'Alessandro, and Thierry Dutoit, "Mage-a platform for tangible speech synthesis," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2012, pp. 353–356.

[29] Thomas Drugman, Geoffrey Wilfart, and Thierry Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis.," in *Interspeech*, 2009, pp. 1779–1782.