An HMM-based Speech-smile Synthesis System: An Approach for Amusement Synthesis

Kevin El Haddad, Stéphane Dupont, Nicolas d'Alessandro, Thierry Dutoit
TCTS lab - University of Mons

Abstract— **This paper presents an HMM-based speech-smile synthesis system. In order to do that, databases of three speech styles were recorded. This system was used to study to what extent synthesized speech-smiles (defined as Duchenne smiles in our work) and spread-lips (speech modulated by spreading the lips) communicate amusement. Our evaluation results showed that the speech-smiles synthesized sentences are perceived as more amused than the spread-lips ones. Acoustic analysis of the pitch and first two formants are also provided.**

Index Terms— **speech-smile, HMM, synthesis, adaptation**

I. INTRODUCTION

Improving speech synthesis and character animation naturalness and expressivity through emotion expressions is currently an important research topic. Laughter and smile are two very important aspects of expressive communication. Both are intuitively related to happiness and/or amusement but also to opposite emotions like fear or stress [1]. The particular focus of this work is on speech-smile. Many studies already cover the phonetic and acoustic properties of speech-smile. Tartter [2] showed in an acoustic and perceptual study that "spreading the lips without attempting to sound happy" increases the first three formants compared to neutral speech. He also showed that this style of speech (which we will refer to as "spread-lips") was acoustically recognized as speech-smile and perceived as happiness. It does, therefore, contribute to transmitting an emotional message during an interaction. In [3], the author compared the lip deformation and acoustic measurements of neutral speech and speech-smiles. Lasarcyk [4] proposed a study of the cues responsible for the audible distinction of smile. For that, an articulatory speech synthesizer was used and the following parameter influence were evaluated: spreading the lips, raising of the larynx and raising of the fundamental frequency (the pitch). The evaluations showed that the pitch was the factor that influences the most the audible distinction of smiling. Nonetheless Emond et al. [5] proposed a study of the cues responsible for the identification of smile sounds by the listeners and also deduced that the pitch height and pitch range were the most discriminant ones and that the rhythm and speech rate were not related to speech-smiles acoustic perception. Furthermore, a study was also made on the abilities of listeners to recognize different kinds of smiles in [6]. This study not only showed that listeners can indeed discriminate between several kinds of smile and non-smile, but it also pointed out the fact that listeners might have preconceptions of "how a smile would sound like". Indeed, the listeners mistakenly related some sounds to smiles when certain acoustical characteristics were changed.

This paper presents a Hidden Markov Model (HMM)-based speech-smile synthesis system. Our work, here, reimplements a part of the system developped in [7], in which "speech-laughs" were synthesized using speech-smiles and "laughing vowels". To the best of our knowledge, no other work has been made on synthesizing amused and/or happy speech building models from small target speech corpora covering those expressions. An acoustic adaptation technique is used. This technique transforms a previously trained acoustic model into an adapted model of the target voice. The adaptation is made making use of speech data where a person speaks while exhibiting a Duchenne smile (which is our definition of speech-smile in the following). A Duchenne smile is a smile expressing true enjoyment. Besides, we also consider adaptation making use of speech data where a person simply speaks while spreading the lips, which is expressing a smile only using the lips, without trying to trigger the enjoyment leading to a Duchenne smile (i.e. without trying to be/sound happy). The comparison of the two styles of speech will allow us to see to what extent do the spread-lips and speech-smile synthesized sentences are perceived as amusement and/or happiness.

In the following, we present our system in section II. Then, our speech databases as well as an acoustic study of these are proposed in sections III and IV respectively. In section V, we provide implementation details. We use a Comparison Mean Opinion Score (CMOS) test in section VI to evaluate the degree of amusement perceived in the synthesised voice. We will also include in both the comparative acoustical and CMOS tests, a neutral speech style which will be used as a reference for our analysis. We conclude in section VII and give our perspectives for future work.

II. SYSTEM OVERVIEW

As mentioned previously, our goal is to create an HMM-based synthesis system for speech-miles, spread-lips and neutral speech styles, and assess comparatively the perceived degree of amusement of utterances synthesized using this approach. Our HMM-based approach can be broken down in three phases: training, adaptation and synthesis.

During the training phase, the system gathers the Mel Generalized Cepstral (MGC) coefficients, excitation and

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.50.01.14.F.

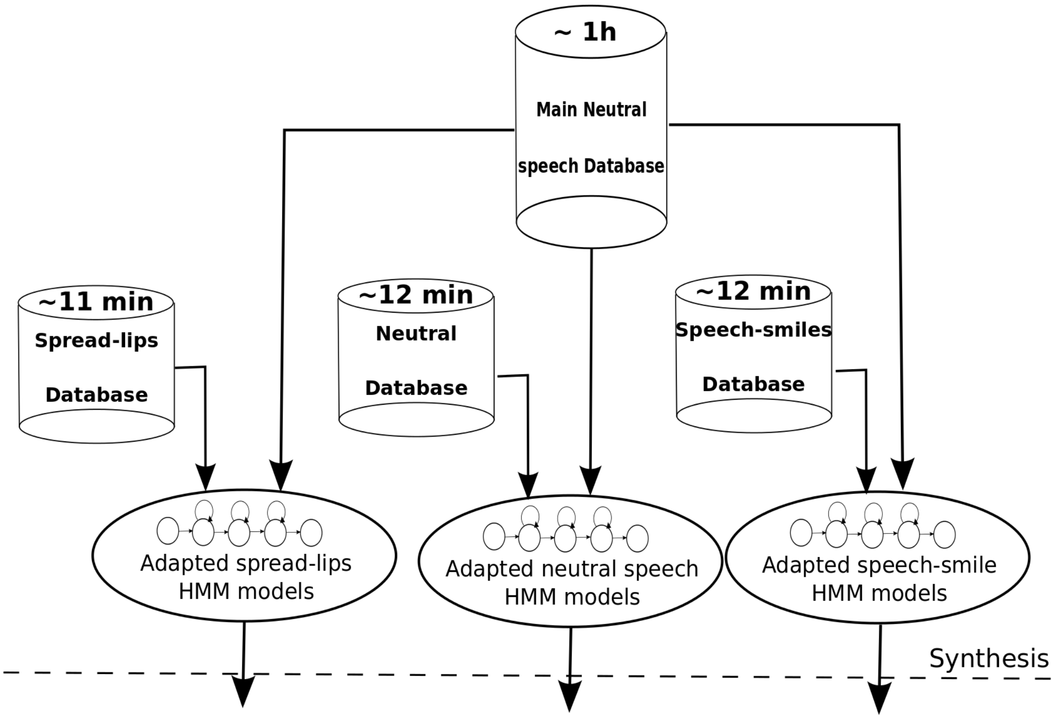

Fig. 1: HMM-based speech-laugh synthesis system pipeline

phoneme duration in a unified HMM model. The probability density function of each HMM state is generally represented by a Gaussian Mixture Model (GMM). We are using here five state left to right HMM topologies and single Gaussian models for each state.

After the training phase comes an adaptation step in which a speaker's source voice is adapted to a target voice using the Constrained Maximum Likelihood Linear Regression (CMLLR) algorithm [8]. This algorithm uses an affine linear function to transform the means and covariances of an initial model (created a speaker A's voice) so as to maximize the likelihood of the target voice (coming from speaker B). This phase allows us to obtain speech-smile, spread lips and neutral speaker B speech acoustic models from limited amounts of data of those speech styles.

Finally, during the synthesis phase, the most likely model representing the desired phonemes are concatenated making a sentence model based on the given labels. This model is then used to synthesize a speech waveform through a parameter generation algorithm and then a speech synthesis filter. The approach is summarised in Fig. 1.

III. Data

A. Source speaker's voice

The data comes from the French database collected in [9]. It was recorded from a native French-speaking person from Belgium. The database contains three different subsets, each corresponding to a different degree of articulation. Here, we only used the normal speech for our purpose. This set was recorded by asking the speaker to read 1359 phonetically balanced sentences (around 55 minutes of speech recordings). The recordings were made at 44.1 kHz and 16 bits PCM.

B. The target voices

Another French-speaking volunteer was asked to read subsets of the same previously mentioned phonetically balanced sentences: 250 sentences with spread-lips (approximately 11 minutes), 250 sentences with neutral speech (approximately 12 minutes), and 200 sentences with speech-smile (approximately 12 minutes), each making approximately 19% of the source speaker's voice database. For the speech-smile sentences, the subject was asked to read the sentences while sounding happy and/or amused. The subject was hence smiling and even close to laughing at some point. For the spread-lips sentences, the subject was asked to read them while spreading the lips as if he was smiling, but with the instruction of not trying to sound happy and/or amused. For the neutral sentences, the subject was asked to read them in a "normal" way without displaying any particular emotion. He was trained in order to have the desired speech style needed for each recording session. The recordings for the three speaking styles were made using a rode Podcaster USB microphone at 48 kHz and 16 bits PCM.

C. Data post-processing

The recordings made in the databases used in this system were all downsampled to 16 kHz in order to have a uniform sampling frequency.

IV. Acoustic studies

As mentioned before, acoustic studies have already been made on the effect lip spreading has on the acoustic of speech. Here, we rather focus on a comparison between the acoustic effects spreading the lips have (spread-lips style), as well as the ones speaking in a “happy way” (speech-smile style). In order to do that, French vowels [a], [e], [i], [o] and [u] were first extracted from each of the three different target voice recordings. The same amount of each vowel instances was extracted from each speech style for comparison by choosing common sentences in all styles. We obtained a total of 455 [a] occurrences, 342 [e] occurrences, 334 [i] occurrences, 237 [o] occurrences and 150 [u] occurrences for each of the three styles. Please note that those are contextual vowels since they were directly extracted from full sentences (this is also in contrast with previous studies on sustained vowels). The mean of the first two formants was calculated for each vowel in each style using the Snack extraction algorithm [10]. Fig. 2 and Fig. 3 are plots of F_1 means with respect to F_2 means with the standard deviation of those two values. Fig. 2 presents the values in the spread-lips and neutral speech styles while Fig. 3 present them in the speech-smile and neutral speech styles. We can now compare the neutral speech style with the spread-lips and speech-smile styles respectively.

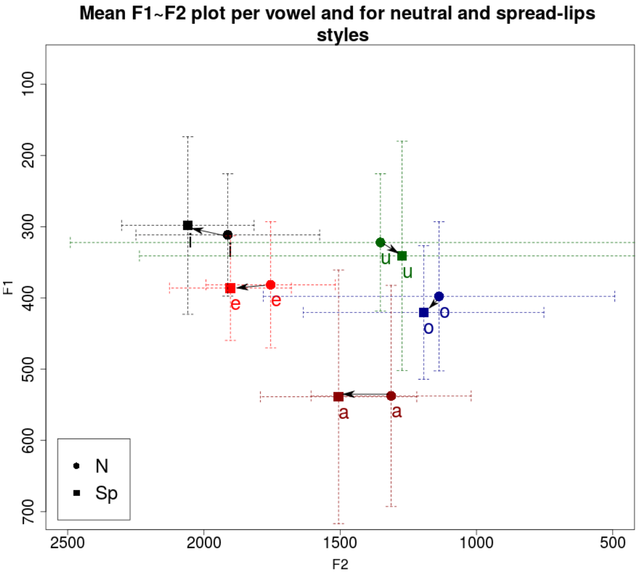

Fig. 2: Mean F_1 values function of mean F_2 values (both in hertz) and their standard deviation of each vowel [a], [e], [i], [o] and [u] in the spread-lips and neutral speech styles

As we can see, the speech-smile style tends to move the formants towards the center of the $F_1 \sim F_2$ vowels’ plane. This is a similar phenomenon to the one observed with the formants of speech-laughs and laughter (cfr. [11]). The more the degree of laughter in speech, the less the speaker keeps control over his/her articulation, therefore bringing novels toward more central positions (in both place and opening).

We hypothesize that this reasoning can also be made here for speech-smiles. In fact, sounding amused and/or happy should end up with the speaker having less control over articulatory motion.

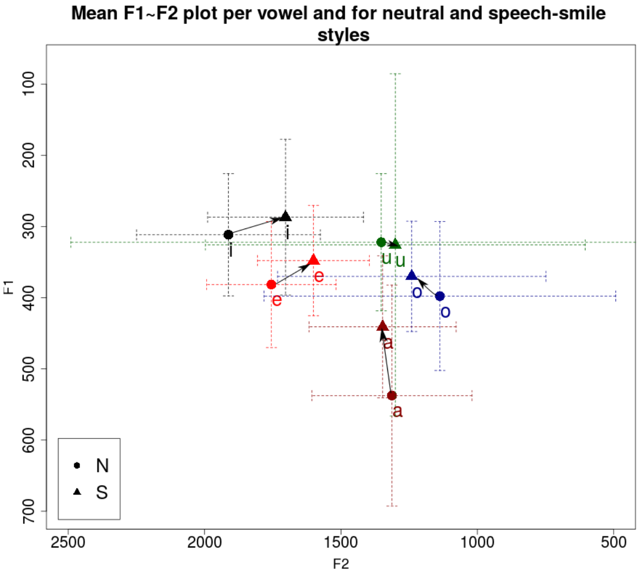

Fig. 3: Mean F_1 values function of mean F_2 values (both in hertz) and their standard deviation of each vowel [a], [e], [i], [o] and [u] in the speech-smile and neutral speech styles

This reasoning implies that there might be a continuum between smile and laughter in some cases. This would align with the smile-laugh continuum theory. Ruch [12] proved that enjoyment smiles were involved in laughter.

The spread-lips and speech-smiles effects on formants can also be partly accounted using a physiological point of view. Spreading the lips induces a shortening of the frontal part of the vocal tract, eventually making the vowel more frontal too. This, on an acoustic level, induces an increase of the second formant. This is what we observe on the spread-lips speech. In the speech-smile case, F_2 doesn’t increase, likely because the effect of the articulation alteration mentioned previously surpasses this purely physiological one. This could be subject to further studies. On the other side, it seems that for the spread-lips, F_2 seems to be more affected than the F_1 since the variation of F_2 due to spreading the lips is more important than F_1’s. In this case, the lip spreading has little effect on the actual vocal tract opening.

Similarly, pitch values were also estimated using the Snack extraction algorithm for each of the vowels in each of the speech styles previously mentioned. A probability density function was estimated representing the pitch value distribution for each speech style, as shown in Fig. 4.

We can observe an overall increase of the pitch for the spread lips style. However, and quite surprisingly, there was only a small increase of the overall speech-smile pitch.

In summary, a first conclusion we can draw is that lip spreading leads to a rise of the second formant values, as well as of the pitch. The pitch rise confirms previously mentioned studies. In the “sounding amused and/or happy” process, spreading the lips, or what it provokes morphologically, must play a big, if not the main part in increasing the pitch

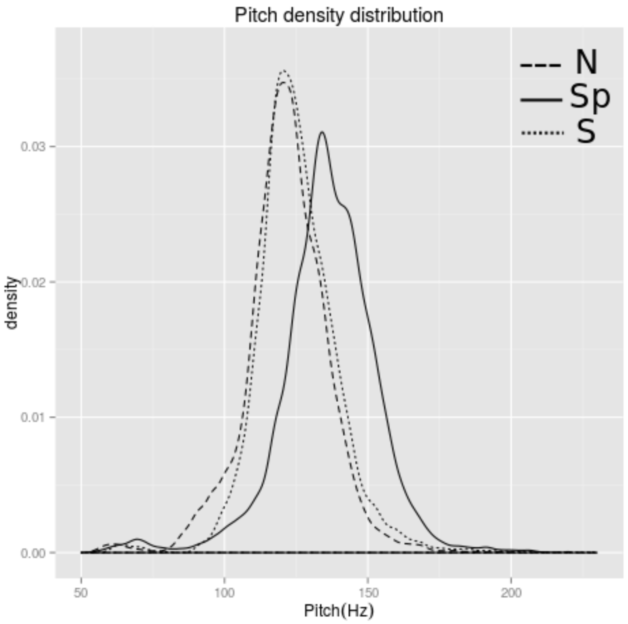

Fig. 4: Probability density distribution of the pitch values for each of the speech styles “neutral” (N), “spread lips”(Sp) and “speech-smile”(S)

value. A second conclusion is that the speech-smiles effect on formant frequencies seem to be similar to the one of speech-laughs [11]. In our view, this is likely due to the speaker attempting to sound amused while talking, as instructed to. These studies would benefit from being extended to a larger corpus including multiple speakers.

V. System implementation

This system was implemented using the publicly available HTS (HMM-based Speech Synthesis System) scripts of the adaptation demonstration canvas. It is a set of speech synthesis tools delivered as a patch for the HTK (HMM ToolKit) [13]. The scripts were run with the parameters α = 0.42, γ = 0 and order of MGC analysis = 24. The synthesis was eventually made using hts_engine [14] (hts_engine is a software that synthesize speech waveforms from trained HMMs by HTS) [15].

VI. Evaluations

An evaluation was made in order to compare the amusement perceived for the three different styles. Thus, 11 random sentences were synthesized for each style. The chosen protocol relies in a Comparison Mean Opinion Score (CMOS) test. Twenty-five French-speaking participants were asked to compare the degree of amusement perceived from two randomly chosen sentences (sentence A and sentence B), each from a different speaking style. They were given the question “which of the following two sentences sound more amused ?” and were asked to score on a unit scale going from -3 to 3. the more negative the grade the more amused sentence A sounded, the more positive it was, the more amused sentence B sounded. A score of 0 would mean that there is no difference between the two sentences. Twenty four evaluations were made by each participants, giving a total of 600 evaluations. The results are shown in Table 1 and the mean and standard error of the results are shown in Fig. 5. N refers to the neutral speech style, Sp to the spread-lips style and S to the speech-smile style. The style disposition in the table is “Sentence A vs Sentence B”.

TABLE I: CMOS test results

	-3	-2	-1	0	1	2	3
S vs N	36	81	61	13	2	5	3
Sp vs S	4	13	16	28	73	47	19
N vs Sp	1	6	9	60	89	31	4

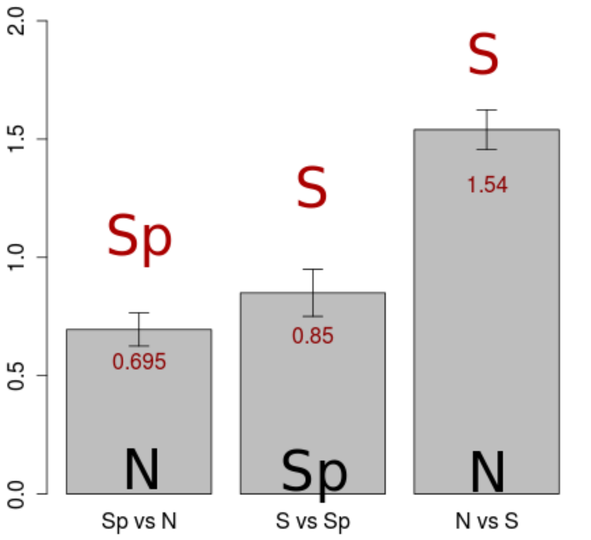

Fig. 5: CMOS test comparing the degree of amusement perceived in each of the speaking styles neutral, spread-lips and speech-smile

After analysing the CMOS scores themselves, we can see that the spread lips sentences were perceived as more amused than the neutral sentences (by .695 points on average), but less amused than the speech-smile sentences (by .85 points on average). The speech-smile style was also perceived more amused than the neutral speech style (by 1.54 on average), making it the style perceived the most amused. Looking deeper into the results, that amusement was better discriminated when comparing speech-smiles and spread lips (in favor of speech-smiles) than when comparing spread-lips and neutral speech styles (in favor of spread lips). This observation might lead to further studies on a possible continuum between spread lips and speech-smile styles on an “amusement degree scale”.

We also noted that some participants perceived neutral sentences as more amused than speech-smile. Taking into account unofficial comments some of the participants made after the evaluation, this is not because the neutral sentences sounded more amused but rather because the speech-smile sentences sounded sad. This confusion in perception of some emotions in their extreme levels was studied by Darwin in [16]. Even earlier, Leonardo Da Vinci in his "Trattato della pittura, talking about facial expressions, mentioned that "Between the expressions of laughter and weeping there is no difference in the motion of the features" [17]. Our results suggest that it may be interesting to study how much this is also true for the vocal expressions of amusement and sadness.

To confirm the above results, a first 95% confidence interval Student's t-test was conducted on each group's obtained scores. This test's null hypothesis was $H0_A$: "The score mean value is null". The test is then run on the scores obtained when comparing each pair of styles. The resulted p-values are given in Table 2.

TABLE II: p-values testing the null hypothesis $H0_A$

	S vs N	Sp vs N	S vs SP
p-value	$< .01$	$< .01$	$< .01$

All the tests conducted here between two different score sets have a p-values lower than .01. Therefore, the null hypothesis $H0_A$ can be rejected. Thus the mean score value of each CMOS score set is not null. So the pesented results of the mean values in Fig. 5 are reliable. Checking weither the differences between the sets' means shown in Fig.4 are significant or not would assure us a reliable comparison.

That is why, we then compared the obtained score mean in favor of the speech-smile style in the "S vs N" group, to the one obtained in favor of the spread-lips style in the "Sp vs N" group. The CMOS test showed a higher mean in the "S vs N" evaluation. the purpose of the test, here, is to confirm this difference, i.e. the mean "S vs N" score is significantly higher than the mean "Sp vs N" score. For that, another 95% confidence interval Student's t-test was run on the above mentioned pair of evaluated speech styles. The null hypothesis of this test was stated as $H0_B$: "There is no significant difference in the means of the two CMOS score sets". The same test was also conducted on the "S vs N" and the "S vs Sp" comparisons with the same null hypothesis $H0_B$. The results are shown in Table 3.

TABLE III: Pairwise p-values testing the null hypothesis $H0_B$

	S vs N	Sp vs N	S vs SP
S vs N	1	$< .01$	$< .01$
Sp vs N	$< .01$	1	-
S vs Sp	$< .01$	-	1

All the tests conducted here have a p-values lower than 0.01 as well. Therefore, the null hypothesis $H0_B$ can be rejected. Thus, there is indeed a significant difference in the means of the respective score sets. Therefore, we can surely state that the amusement discrimination was better perceived when comparing speech-smile and neutral speech style than in any other of the comparative tests.

We can conclude from these tests that with our approach to synthesizing amused speech, the spread lips style can communicate amusement better than a neutral speech style but did it less efficiently than speech-smiles.

VII. Conclusion

In this work, three HMM models for speech synthesis were used, covering the following three speech styles: spread-lips, speech-smile and neutral.

An acoustic analysis of the speech dataset used to build those models was made in order to draw first observations on some of the characteristics of those styles. Taking the neutral speech style as a reference, speech-smile first and second formant behavior was found to be similar to the one of speech-laugh and laugh in [11]. In the spread-lips case, the F_2 average value is increased while the F_1 value remains practically constant on average. Also, we found that the spread-lips speech style contributes to an increase in the pitch value. Pitch increase in the speech-smile data was relatively low compared to the spread-lips one.

Based on those models, tests were conducted to evaluate to what extent the speech-smile and spread-lips speech voices communicate amusement. The results showed that speech smiles communicate amusement better than spread-lips, which came second in the evaluation before neutral speech.

Our near-future perspective is to be able to control the degree of smile/amusement in speech. This is probably achievable using interpolation techniques to cary from on style of speech to another. Another interesting and achievable perspective is to control the degree of amusement in speech by controlling not only the degree of smiles, but also the degree of laughter. For this, the intensity and expressions (smile or laughs) of amusement with respect to the context should be taken into account. A further goal is to have control over the degree of amusement in real time.

References

[1] Laurence Devillers and Laurence Vidrascu, "Positive and negative emotional states behind the laughs in spontaneous spoken dialogs," in *Interdisciplinary Workshop on The Phonetics of Laughter*, 2007, p. 37.

[2] V.C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Perception Psychophysics*, vol. 27, no. 1, pp. 24–27, 1980.

[3] Sascha Fagel, "Effects of smiling on articulation: Lips, larynx and acoustics," in *Development of Multimodal Interfaces: Active Listening and Synchrony*, Anna Esposito, Nick Campbell, Carl Vogel, Amir Hussain, and Anton Nijholt, Eds., vol. 5967 of *Lecture Notes in Computer Science*, pp. 294–303. Springer Berlin Heidelberg, 2010.

[4] Eva Lasarcyk and Jürgen Trouvain, "Spread lips+ raised larynx+ higher f0= Smiled Speech?-An articulatory synthesis approach," *Proceedings of ISSP*, 2008.

[5] Caroline Émond, Lucie Ménard, and Marty Laforest, "Perceived prosodic correlates of smiled speech in spontaneous data.," in *INTERSPEECH*, Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, Franois Pellegrino, and Pascal Perrier, Eds. 2013, pp. 1380–1383, ISCA.

[6] Amy Drahota, Alan Costall, and Vasudevi Reddy, "The vocal communication of different kinds of smile," *Speech Commun.*, vol. 50, no. 4, pp. 278–287, Apr. 2008.

[7] Kevin El Haddad, Stéphane Dupont, Jérôme Urbain, and Thierry Dutoit, "Speech-laughs: An HMM-based Approach for Amused Speech Synthesis," in *Internation Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, in press.

[8] 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.

[9] 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.

[10] Kåre Sjölander, "The Snack Sound Toolkit [computer program webpage]," http://www.speech.kth.se/snack/, consulted on September, 2014.

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

[12] 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.

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

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

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

[16] Charles Darwin, *The expression of the emotions in man and animals*, Oxford University Press, 1872.

[17] Leonardo de VINCI, *Trattato della pittura*, Stamp. de Romanis, 1817.