An HMM Approach for Synthesizing Amused Speech with a Controllable Intensity of Smile

Kevin El Haddad, Hüseyin Çakmak, Alexis Moinet, Stéphane Dupont, Thierry Dutoit

TCTS lab - University of Mons, Belgium

{kevin.elhaddad},{huseyin.cakmak},{alexis.moinet},
{stephane.dupont},{thierry.dutoit}@umons.ac.be

Abstract

Smile is not only a visual expression. When it occurs together with speech, it also alters its acoustic realization. Being able to synthesize speech altered by the expression of smile can hence be an important contributor for adding naturalness and expressiveness in interactive systems. In this work, we present a first attempt to develop a Hidden Markov Model (HMM)-based synthesis system allowing to control the degree of smile in speech. It relies on a model interpolation technique, enabling speech-smile sentences with various smiling intensities to be generated. Sentences synthesized using this approach have been evaluated through a perceptual test. Encouraging results are reported here.
Index Terms: speech-smile, computational paralinguistics, Hidden Markov Models (HMM), interpolation

1. Introduction

Smiles, when occurring with speech, alters it in such a way that it is perceived as smiled speech [1]. Speech-smiles can appear in our daily social verbal exchanges to express not only positive states (empathy, welcomeness, joy, amusement etc.) but also sometimes negative ones (embarrassment, sarcasm etc.). Some of these states might be culturally and/or social-contextually dependent but others, like amusement, can be found in every culture. In fact, humans begin smiling when they are still babies. Because of their importance in our social expressions, smiles and speech-smiles were the subject of several studies and researches.

In [2], Robson studied the effects of labial spreading on speech. The authors showed that spreading the lips while talking was acoustically perceived as speech-smiles and also presented acoustical. Fagel studied the anatomical effects and acoustic consequences of the smiling articulation on speech [3]. Studies of the acoustical factors responsible for the production of speech-smiles are conducted in [4]. The pitch was shown to be the contributing factor. Nonetheless, Émont et al. proposed in [5], a comparative study of the acoustical parameters responsible for the discrimination of smiles. The authors also found the pitch to be the most important cue for speech-smile discrimination. The authors in [6], proved that listeners can discriminated acoustically between different types of speech-smiles. They also suggested that the listeners linked the speech-smile sounds to prototypical ideas in order to judge whether the sentence is altered by smiles or not.

Our main objective is to create a controllable amused speech-synthesis system with different level of amusement. Some previous work were made in the framework of this subject, Urbain et al. proposed in [7] an HMM-based laughter synthesis system on different degrees of arousal. We previously developed a Hidden Markov Model (HMM)-based speech-smile synthesis systems [8]. We used this system to synthesize speech-smile created from Duchenne smiles (smiles containing real amusement and/or enjoyment), neutral speech altered by labial spreading and neutral speech. We showed that the speech-smiles were perceived as more amused than the other two speech styles. This speech-smile synthesis system was also used in [9] to create an HMM-based speech-laugh synthesis system.

We present in this work an HMM-based speech-smile synthesis system with controllable levels of smiles. In fact, we do not know of any previous work dedicated to the control of the degree of smile for synthesized speech. An interpolation technique was leveraged to create this system using the MAGE library [10]. Indeed, HMM-models were created from the recorded neutral and smiled speech of a Belgian French native actor. A weighted interpolation was then used to vary the level of smile in the synthesized speech. This system is a first step towards the creation of a controllable amused synthesis system which will include also speech-laughs and isolated laughter and maybe other forms vocal expression of amusement. A perceptual Comparative Mean Opinion Score (CMOS) test was conducted in order to test on one side the efficiency of our system in communicating amusement, and on the other side, the degree of naturalness with which it is perceived by listeners.

In the remainder of this article, section 2 will contain the description of our system. It will describe the HMM models created for neutral and smiled speech as well and explain the interpolation technique utilized. We will then expose in section 3 the evaluations that we conducted. Section 4 will discuss the results of those evaluations. At last, we will conclude and present our perspectives for future work in section 5.

2. System Description

As described in Fig. 1, our HMM-based speech-smile synthesis system for controllable smile level can be broken down in three steps:

1. Creating an HMM neutral speech model
2. Creating an HMM speech-smile model
3. Using a weighted sum interpolation of the two previously created HMM models to create a speech-smile model at a level of smile corresponding to the weights used.

In order to create the HMM models several databases were used.

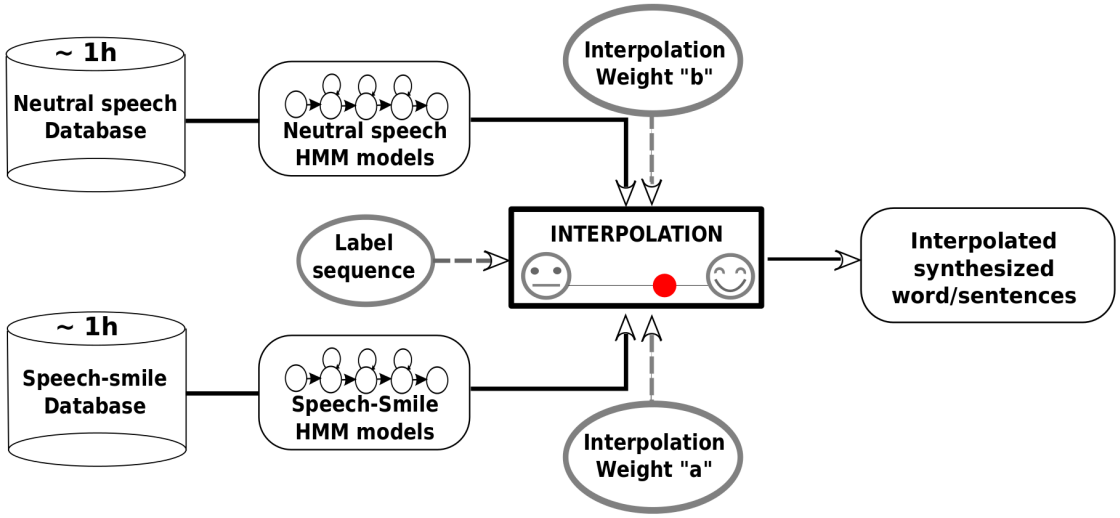

Figure 1: HMM-bases speech-smile synthesis system controllable on different levels of smile

2.1. Neutral speech database

The neutral speech database collected in [11] was used to create the neutral speech HMM model. This audio database contains approximately 1 hour of French sentences read by a Belgian French native actor and was recorded along with hypo and hyper-articulated database. The recordings were made at 44.1 kHz and stored PCM 16 bits using a high quality microphone.

2.2. Speech-smile database

The speech-smile database was collected from the same actor from which the previously mentioned neutral speech database was recorded. The database was recorded in a sound-proof booth using a high quality microphone. The recordings were made at 48 kHz and stored PCM 16 bits. The actor was asked to read the same sentences as the ones used for the neutral speech database. He was also instructed to read them while smiling and sounding happy but not to laugh. During the data processing step, some sentences had to be removed from the database because they either contained laughter bursts or one or more vowels in the sentences were altered by some kind of tremolo/vibrato as also mentioned in [12]. The total of the remaining speech-smile database contained approximately 50 minutes of data.

2.3. Neutral and Speech-smile HMM models

We wanted to have models trained with the same data to simplify the interpretation of the results later on in the study. Therefore, in order to interpolate between the two models, we first used the same sentences from each database to train the models. In order to do so, some sentences were removed from the neutral speech database so that the sentences in it corresponded to the ones in the speech-smile database. Then HMM models were trained for the neutral speech (HMM-N) and for the speech-smiles (HMM-S).

2.4. Interpolating between the models

To obtain several levels of smile, an interpolation technique was used. Each state of HMM-N and HMM-S were represented by a multivariate single Gaussian distribution summarized by a mean vector μ and a covariance diagonal matrix Σ. The interpolation is made by a weighted sum of the HMM-N μ's and Σ's with their corresponding μ's and Σ's of the HMM-S models as shown in the following equations [13]:

$$\mu = a\mu_S + b\mu_N \tag{1}$$

$$\Sigma = a^2\Sigma_S + b^2\Sigma_N \tag{2}$$

Where μ and Σ correspond to the interpolated mean vector and diagonal covariance matrix respectively, a to the weight related to HMM-S and b to the weight related to HMM-N such that $a + b = 1$. Thus, the higher a is, the more the interpolated model will tend towards the speech-smile model. For $a = 1$ the interpolated model would be equal to HMM-S and for $a = 0$, it will be equal to HMM-N. Speech-smile sentences on different levels of smile could then be synthesized by changing the value of a.

2.5. Implementation details

The HMM models were trained using the publicly available HMM-based Speech Synthesis System (HTS) demo scripts [14, 15]. HTS is a patch code to Hidden Markov Model Toolkit (HTK) [16]. The data used were all downsampled to 16 kHz to obtain a uniform sampling frequency on one side and for compatibility with future related projects on the other side. To model the filter, 24 Mel Generalized Cepstral coefficients (MGC) and their dynamic and acceleration coefficients were used with a wrapping frequency of α=.42 (this value was chosen because it approximated best the auditory frequency scale considering a sampling frequency of 16 kHz [17]).
The interpolation was made using MAGE which is a library for reactive speech synthesis.

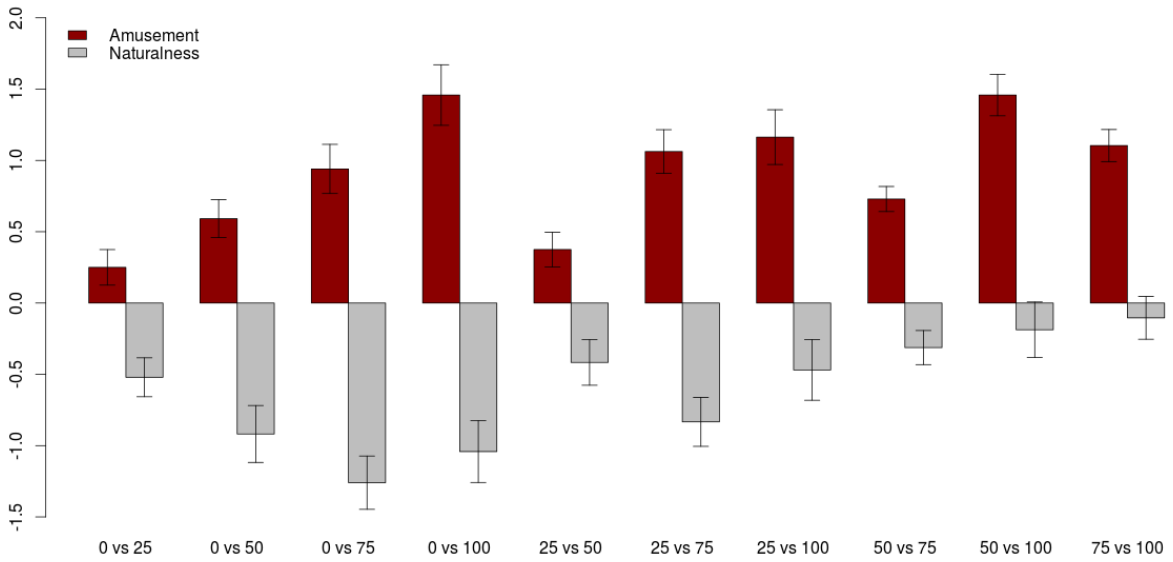

Figure 2: Mean and confidence interval of the scores obtained from the pairwise comparison of the interpolated sentences. The values on the x axis represent the weight related to the HMM-S in percent.

3. Evaluation of the System

In order to evaluate the efficiency of our system, we synthesized 10 randomly chosen sentences, interpolated using 5 different weight values α for each sentence: 0, 0.25, 0.5, 0.75 and 1. The 50 obtained sentences were then evaluated by 24 French-speaking participants via a CMOS test. The participants were given 20 pairs of sentences, chosen randomly among the 50, each pair formed by the same sentence with a different level α of smile. A pair of sentences, was presented to the participants as "sentence A" and "sentence B". The participants were asked to make the comparison on two different axes: the first concerned the degree of amusement perceived and the second the degree of naturalness of the sentences.
They were asked to choose between labels on those two axes which will later correspond to scores (shown between parentheses in the following) for the statistical studies. Thus it was requested to say whether sentence A was *much more* (3), *more* (2), *slightly more* (1), *slightly less* (-1), *less* (-2), *much less* (-3) amused or natural (depending on the axis on which he/she was grading) than, or *identical* (0) to sentence B.

4. Results and Discussion

The pairwise comparison of the 5 different interpolation weights for each sentence are given in Fig 2. This barplot represents the mean and confidence interval of the scores obtained when comparing each pair of sentences on the amusement scale and on the naturalness scale. The labels on the abscissa axis represent the weights in percent related to the HMM-S. For example "25 vs 50" refers to the comparison of a sentence created by interpolating 25% of HMM-S and 75% of HMM-N, with a sentence created by interpolating 50% of HMM-S and 50% of HMM-N.
The labels are presented in a pattern "**a** vs **b**" where **a** received negative scores and **b** the positive ones. This means that if the bar has a negative value, **a** was better graded than **b** and if the bar has a positive value **b** was better graded than **a**. For example in the "0 vs 100" comparison case, the sentences synthesized with 100% of HMM-S were better graded that the ones synthesized with 0% of HMM-S on the amusement axis. On the naturalness axis, The sentences with 0% of HMM-S were better graded than the ones with 100% of HMM-S.
Students t-tests were applied on each of the set of scores obtained for the pairwise comparisons. The null hypothesis being that the mean values of the scores obtained in each case is null. On the amusement scale, we obtained p-values smaller than 0.05 in all the cases. This means that we can reject the null hypothesis in favor of the alternative hypothesis. Therefore the mean values obtained on the amusement scale are significantly different from zero. These results, thus, validate and confirm the ones obtained from the evaluations. On the naturalness scale, the p-values were greater than .05 in the "50 vs 100" an "75 vs 100" cases, we can thus not reject the null hypothesis. They were smaller than 0.05 in all other cases, in this case the null hypothesis can also be rejected. These results also explain the presence of the value 0 inside the confidence interval on the naturalness scale for the "50 vs 100" and "75 vs 100" cases and outside of it for all other cases.
For a clearer and more detailed presentation of the data collected, these were placed in Table 1 and Table 2. They both show the sentences compared as described previously and the total choice made by the participants in each case of the pairwise comparison. In the tables "**a vs b**" refers the pair of sentences, **Pref. a** and **Pref. b** show the total number of times the participants chose **a** and **b** respectively and **Ties** shows the number of times the participants rated the sentences as identical. Table 1 shows these results for the amusement scale while Table 2 shows them with regard to the naturalness scale.

a vs b	Pref. a	Ties	Pref. b
0 vs 25	5	25	18
0 vs 50	3	22	23
0 vs 75	4	10	34
0 vs 100	5	3	40
25 vs 50	4	22	22
25 vs 75	3	9	36
25 vs 100	5	8	35
50 vs 75	1	14	33
50 vs 100	2	6	40
75 vs 100	0	10	38

Table 1: Preferences on the amusement scale

a vs b	Pref. a	Ties	Pref. b
0 vs 25	23	22	3
0 vs 50	30	11	7
0 vs 75	37	5	6
0 vs 100	32	9	7
25 vs 50	19	20	9
25 vs 75	31	12	5
25 vs 100	26	9	13
50 vs 75	18	26	4
50 vs 100	21	11	16
75 vs 100	12	27	9

Table 2: Preferences on the naturalness scale

The first conclusion we can draw from these results is that concerning the amusement perceived any sentence with a given HMM-S weight was better graded than the same sentence synthesized with a lower HMM-S weight. This shows that we have successfully been able to control the level of amusement by controlling the level of smile in a sentence. We also notice that the naturalness perceived is lower with higher HMM-S weights. This might be due to a preconception from the participants of what an amused sentence sounds like as suggested in [6]. It might also be due to a poorer synthesis quality from the HMM-S compared to the one obtained from HMM-N. The amelioration of the perceived naturalness will be one of the goals of our further studies.

5. Conclusion and Perspectives

In this work we have presented an HMM-based speech-smile synthesis system with controllable levels of smile. This system is a first step towards the creation of a synthesis system with controllable amusement levels. After evaluations, our system proved to successfully express amusement through speech-smiles on different levels. A higher amusement degree was perceived when the synthesized sentence contained a higher level of smile. The results were validated by a Student's t-test.
Future work include the amelioration of the naturalness of our system. The factors causing the degradation of the naturalness degree with the increase of the HMM-S weight will be investigated to find the most suited solution. The amelioration of the naturalness perceived from our system could also probably be done by training the HMM models on a lager amount of data (for each of the HMM-N and HMM-S).
We will also focus on developing an amused synthesis system with controllable levels of amusement by integrating work previously made on this subject like [9]. Another objective would be to control this system in real-time. This can be done by using MAGE, the same library used for the interpolation in this work, as it allows reactive, real-time speech synthesis.

6. References

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

[2] J. Robson and B. 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.*, vol. 1. International Congress of Phonetic Sciences, 1999, pp. 219–222.

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

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

[5] C. Émond, L. Ménard, and M. Laforest, "Perceived prosodic correlates of smiled speech in spontaneous data." in *INTERSPEECH*, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, Eds. ISCA, 2013, pp. 1380–1383.

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

[7] 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.

[8] K. El Haddad, S. Dupont, N. d'Alessandro, and T. 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'15)*, May 4-8 2015.

[9] K. El Haddad, S. Dupont, J. Urbain, and T. Dutoit, "Speech-laughs: An HMM-based Approach for Amused Speech Synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, April 19-24 2015, pp. 4939–4943.

[10] M. Astrinaki, N. d'Alessandro, and T. Dutoit, "MAGE-a platform for tangible speech synthesis," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2012, pp. 353–356.

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

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

[13] M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "HMM-based speech synthesis with various speaking styles using model interpolation," in *Proc. Speech Prosody*, 2004.

[14] K. Oura, "HMM-based Speech Synthesis System (HTS) [computer program webpage]," http://hts.sp.nitech.ac.jp/, consulted on August, 2014.

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

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

[17] T. Masuko, "Hmm-based speech synthesis and its applications," 2002.