BREATH AND REPEAT: AN ATTEMPT AT ENHANCING SPEECH-LAUGH SYNTHESIS QUALITY

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

University of Mons (UMONS) - TCTS Lab

ABSTRACT

In this work, we present a study dedicated to improve the speech-laugh synthesis quality. The impact of two factors is evaluated. The first factor is the addition of breath intake sounds after laughter bursts in speech. The second is the repetition of the word interrupted by laughs in the speech-laugh sentences. Several configurations are evaluated through subjective perceptual tests. We report an improvement of the synthesized speech-laugh naturalness when the breath intake sounds are added. We were unable, though, to make a conclusion concerning a possible positive impact of the repetition of the interrupted words on the speech-laugh synthesis quality.

Index Terms— HMM-based, laughter, synthesis, speech-laugh

1. INTRODUCTION

Amusement is a common feature of our daily emotional states and social interactions. If added to the machine's dialog repertoire, amusement will contribute to make the interaction more natural and therefore more comfortable to the users. Laughter being one of the most common ways of expressing amusement,it has been the subject of many research studies in different fields. In particular, some studies focused on the co-occurrence of laughter and speech, i.e. speech-laughs. Speech-laughs have been found to be very common in conversations [1, 2]. Studies were also made comparing speech-laughs and isolated laughter. Dumpala proposed a feature extraction and comparison successfully discriminating between laughter and speech-laughs [3]. In [4], Menezes exposed an acoustic comparison (formant frequency values and pitch) between neutral speech, speech-laughs and laughter. Further phonetic and acoustic studies on speech-laughs can also be found in [1], [2] and [5]. However, to the best of our knowledge, only very few work was made concerning speech-laugh synthesis.

Oh [6] proposed to modulate speech in order to create speech-laughs via a control of acoustic parameters such as pitch, duration, tempo, etc. Also, a Hidden Markov Model (HMM)-based synthesizer system for speech-laughs was developed in our previous work [7]. In this system, laughter bursts were inserted into speech-smile sentences (i.e. speech modulated by the effect of smiling, see also [8]). The work presented in this paper is an attempt to improve the degree of naturalness of the speech-laughs synthesized through that approach. Firstly, relying on results in [9] emphasizing the importance of breath in laughter synthesis, we attempted here to add synthesized breath intake sounds to synthesized speech-laughs utterances. Secondly, we also studied the possibility of generating synthesized sentences where the words interrupted by laughter are repeated. Evaluations are made to determine whether those approaches are efficient to improve the degree of naturalness perceived by listeners.

The article is organized as follows. In section II, we first give a brief description of our proposed baseline system [7]. Our proposed updates are also exposed there. In order to study the impact of adding the breath intake sounds and/or repeating the interrupted words, five different configurations were elaborated. Those serve further in the evaluations and are described in details in section III. The evaluation protocol is described in section IV. The results are exposed and analyzed in section V. Finally section VI gives our conclusion as well as perspectives for future work.

2. SPEECH-LAUGH SYSTEM IMPROVEMENT ATTEMPT

Our HMM-based speech-laugh synthesis system workflow is given in Fig.1. Please note that this is only a summary. For more details please refer to [7] and [8]. This system is trained on french speech data. An acoustic model trained on a database of approximately one hour of neutral speech was adapted using a small speech-smile database, in order to obtain an acoustic model enabling the synthesis of speech-smile sentences.The adaptation was made using the Constraint Maximum Likelihood Linear Regression (CMLLR) algorithm [10]. An acoustic model of laughter bursts is also available enabling to insert such events when synthesizing new amused sentences. Those laughter burst models are trained on a database of so-called laughing vowels (cf section 2.1). More details are given in the following.

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.

H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

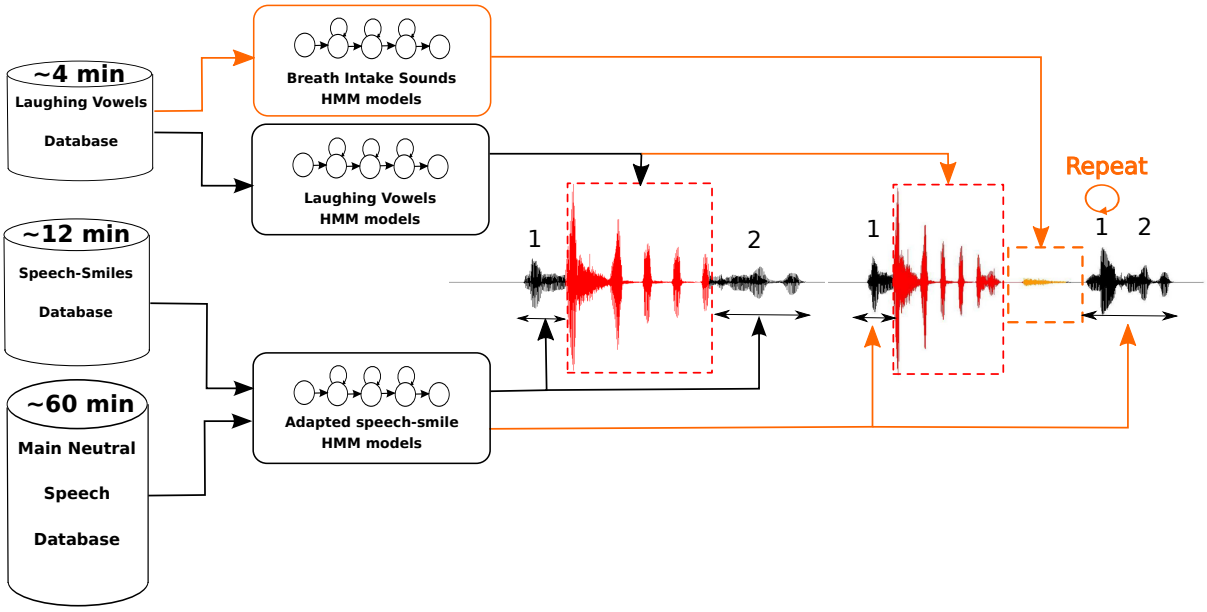

Fig. 1: HMM-based speech-laugh synthesis system pipeline and updates made on it for the purpose of this work (orange).

2.1. Baseline Databases and Acoustic Models

The neutral database was taken from [11] where audio recordings were made of hypo-articulated, hyper-articulated and neutral speech styles. This neutral database contains about 1 hour of french sentences read by a French speaking actor from Belgium. The data was sampled at 44.1 kHz and stored in 16 bits PCM. Speech-smiles were recorded from another french speaking naive who was asked to read a subset of the same sentences, giving approximately 10 minutes of material. That same person was also asked to produce sustained French vowels while watching funny videos. Laughter sometimes occurred in the middle of the vowels pronunciation. We called those events "laughing vowels". They enabled us to train acoustic models of laughter bursts, appropriate for synthesizing speech-laughs. The laughing vowels pattern is the same as the laughter one. It usually contains fricative and vowel sounds. In most cases, this pattern consists of a succession of fricatives and vowels, but in others, it varies. To make an optimal use of our laughing vowels database, HMM models are created for each vowel and for the fricatives. In this way, these HMM models can be concatenated to form any laughter pattern desired. We thus have control over the laugh pattern during synthesis.

2.2. Improvements over the Baseline

Breath intake sounds can also be found in the laughing vowels database. Nine similar sounding breath intakes were used to create a left-to-right 5 state HMM model. In fact, on the contrary of the laughing vowels variable pattern, the breath intake one is consistent. This is due to the similarity in the breath intake sounds. This is why, a relatively small number of breath intake instances was enough to successfully create an HMM model. This model was then used to insert breath intake sounds in the synthesized speech-smile sentences, with an approach similar to the one used for inserting laughter bursts (see Fig. 1), i.e. by altering the transcription passed to the synthesis engine.

3. METHODS

As introduced earlier, the aim of this work was to study the impact of two factors on the perceived degree of naturalness of the synthesized speech-laugh sentences:

1. adding breath intake sounds,
2. repeating words interrupted by laughter.

To do so, five different configurations are proposed and described here.

3.1. Restriction Rules

The parameters of laughter bursts (e.g. position, intensity, duration etc.) are very variable as they interrupt, intermingle with and alter speech to form speech-laughs. Indeed, in speech-laugh sentences, laughter bursts can probably occur anywhere, and their position, intensity and duration depend on the situational or social context. The relation between those parameters and the context is however out of scope of this paper. Hence, we focused on a restricted set of configurations, as described here.

The speech-laughs synthesis rules in this work are at two levels. At the sentence level:

- the same laugh pattern will be synthesized for all the configurations and for each sentence to be evaluated.
- only speech-laughs occurring in the middle of sentences are considered.

These will allow us to reduce the previously mentioned parameter variability.

At the word level:

- only words made of more than one syllable are considered.
- laughter will not be inserted in the last syllable of the words.

In fact, other studies would be needed to determine whether it would be appropriate to repeat the words in case laughter is interrupting monosyllabic words, or multisyllabic words at their borders (first or last syllables).

3.2. Configurations

Table 1 describes five configurations of the speech-smile sentences, breath intake sounds and repetition of interrupted word by laughter bursts.

- **L** indicates whether laughter was inserted at all or not,
- **BI** refers to added breath intake,
- **R** indicates whether the interrupted word is repeated or not.

The $+$ in the table refers to the fact that **L**, **BI** or **R** is used and $-$ refers to the fact that it was not used for the considered configuration. Thus, **C1** correspond to speech-smiles while all other configurations correspond to speech-laughs. The difference between the later configurations is whether or not **BI** and/or **R** are used to synthesize them. **C5** is thus the configuration where both **BI** and **R** are included.

Configurations	**C1**	**C2**	**C3**	**C4**	**C5**
L	$-$	$+$	$+$	$+$	$+$
BI	$-$	$-$	$+$	$-$	$+$
R	$-$	$-$	$-$	$+$	$+$

Table 1: Configurations

4. EVALUATIONS

Evaluations are done using 10 different randomly chosen sentences. These sentences did not have any funny content in particular. A common laughter pattern to be inserted in all the synthesized sentences was chosen. This was done to avoid the effect the laughter pattern could have on the sentences perception. The study of this effect is beyond the scope of this work. The common laugh pattern chosen consists of a succession of four fricative-vowel pattern.

As described previously, each sentence is generated in one of the 5 possible configurations exposed in Table 1. A total of 21 French speaking subjects participated in the tests. Each voted for 20 randomly selected sentences among the 50 available. They were asked to grade their perceived degree of naturalness.

5. RESULTS

The mean and standard errors of the obtained ratings for each configuration are shown in Fig.2.

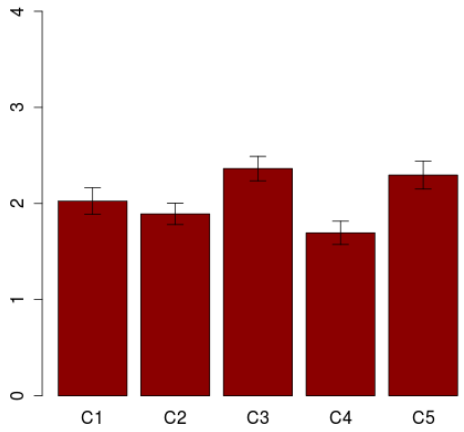

Fig. 2: Mean scores and standard errors for the five configurations

A 99% confidence interval Tukey's HSD (Honest Significant Difference) test was conducted to compare the 5 configurations. The p-values obtained are reported in Table 2.

Config.	1	2	3	4	5
1	-	0.797	0.385	0.318	0.668
2	0.797	-	**0.038**	0.939	0.115
3	0.385	**0.038**	-	**0.003**	0.992
4	0.318	0.939	**0.003**	-	**0.014**
5	0.668	0.115	0.992	**0.014**	-

Table 2: Pairwise p-values between the synthesis configurations. p-values showing significant differences are in bold.

According to Table 2 we can conclude that adding **BI**s improves the quality of synthesized speech-laughs by improving the degree of naturalness perceived. The configurations C_3 and C_5 are indeed significantly better than C_4 and better than C_2 (significantly better when comparing C_3 to C_2). We were not able to make any conclusion concerning the impact of repeating the interrupted word. Further studies should be made

in order to advance the understanding of that aspect, and possibly find configurations that would alter the perceived naturalness positively.

6. CONCLUSION AND PERSPECTIVES

This study successfully proved the usability of our HMM-based speech-laugh synthesis system as a tool to investigate the study of the interaction between speech and laughter in a sentence. It also helped improving the perceived synthesized speech-laughs quality. Further studies will focus on studying the position of laughter in speech-laugh sentences, through statistical analysis of naturalistic data. The investigation of the intensity of laughter as a function of the context in which it occurs will also be considered. The effect inserting laughs coming from another speaker has on the quality perceived will also be studied in future work.

REFERENCES

[1] Eva E. Nwokah, Hui-Chin Hsu, Patricia Davies, and Alan Fogel, "The integration of laughter and speech in vocal communicationa dynamic systems perspective," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 4, pp. 880–894, 1999.

[2] 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.

[3] 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.

[4] Caroline Menezes and Yosuke Igarashi, "The speech laugh spectrum," *Proc. Speech Production, Brazil*, 2006.

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

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

[7] Kevin El Haddad, Stéphane Dupont, Jérôme Urbain, and Thierry 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.

[8] Kevin El Haddad, Stéphane Dupont, Nicolas d'Alessandro, 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'15)*, May 4-8 2015.

[9] J. Urbain, H. Çakmak, and T. Dutoit, "Automatic phonetic transcription of laughter and its application to laughter synthesis," in *Proceedings of the Fifth biannual Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013.

[10] 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.

[11] 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.