Semi-Parametric Synthesis of Speaker-Like Laughter

Grégory Beller

beller@ircam.fr

IRCAM

1. place Igor Stravinsky

75004 Paris, France

Abstract. This paper describes a semi-parametric speaker-like laughter synthesis method. A large corpus of spontaneous laughter is presented. An attempt to use traditional automatic segmentation on the data is discussed. Significant results from the statistical analysis of the corpus are then presented, with concern to the static and dynamic acoustic characterizations of bouts and syllables. Interestingly, laughter prosody seems to be guided by the same physiological constraints as verbal speech. After this analysis part, a method for synthesizing laughter from any neutral utterance using information from the previous results is described. A TTS algorithm selects some phones that are duplicated to create a homotype series. Finally, speech processings modify the prosody of this series, providing a realistic high quality speaker-like bout of laughter.

1 Introduction

A database management system for speech is being constructed to allow the manipulation of large corpora for various artistic objectives [17]. One of our objectives is high quality expressive Text-To-Speech synthesis. This objective is divided into two parts: high quality neutral TTS synthesis and high quality expressive speech transformation. The latter part requires statistical context-dependent analysis of prosodic parameters according to expressivity [4] [2]. This is achieved by para-linguistic speech manipulations such as articulation degree modification [3]. This modification can then be applied to either synthesized or spoken speech. It is currently being used by film and theater directors and also in dubbing studios. It is within this outline that the study is conducted.

The analysis of a very large corpus of naturally-occurring conversational speech [6] reveals that approximately one in ten utterances contain laughter. Therefore laughter is a powerful means of emotion expression which is beginning to be analyzed and used in speech synthesis [12]. Acoustic studies on spontaneous [6], semi-spontaneous [1] and simulated corpora exhibit interesting acoustic features, tendencies, and variabilities. Despite the youth of this new research topic, the need of a common terminology for laughter description has already been adressed and partly solved [16]. In order to compare this study to others, we refer to the terminology of Trouvain et al. for the definition of the terms used in this paper.

Although the acoustic of laughter is highly variable, some regularities can be observed with regard to its temporal structure. Laughter bouts are typically initiated by one or two singular elements (i.e. non-repeated, with large variability in acoustic parameters). These are often followed by a succession of *syllables* with predictable similarity, i.e. a homotype series [9]. The overall temporal behavior can be captured by a parametric model based on the equations that govern the simple harmonic motion of a mass-spring system [13].

Our main goal is to apply a desired expressivity to a spoken or synthesised neutral utterance. In the case of happiness, adding speaker-specific laughter as a para-verbal burst to the transformed utterance makes the result more likely to be perceived as the intended expressivity. Unfortunately, no laughter is present in the neutral utterance and one must synthesize it taking in to account only the verbal content of the utterance. This paper explains a semi parametric method that is able to provide speaker-like laughter from a neutral utterance using a corpus based analysis of the dynamics of laughter.

2 General Overview

First, a large corpus of spontaneous laughter [6] is presented. An attempt to use traditional automatic segmentation of the data is discussed. Significant results from the statistical analysis of the corpus are then presented, with concern to the static and dynamic acoustic characterizations

of bouts and syllables. After this analysis part, a method for synthesizing laughter from a new neutral utterance using information from the previous result is presented. Then, a rule-based selection algorithm picks up some phones that are duplicated to create a homotype series. Finally, speech processing modify the prosody of this series, providing a realistic high quality bout of laughter.

===

Please insert here Figure 1. Overview of the semi-parametric synthesis method.

===

3 Corpus

The data came from a large corpus of spontaneous Japanese conversational speech [7]. Two sets of laughter bouts have been extracted: one of a male speaker JMA and one of a female speaker JFA. Corpora consist of 1150 bouts of JMA and 953 bouts of JFA recorded with a head-mounted Sennheiser HMD-410 close-talking dynamic microphone and a DAT (digital audio tape) at a sampling rate of 48kHz.

4 Analysis part

4.1 Automatic Segmentation

The bouts were automatically labeled by an ”unsupervised” automatic HMM-based segmentation system [11] trained on a neutral multi-speaker database [10]. A first attempt presented bad segmentation results. This was because of the lack of breathing in the training corpus, a lot of devoiced vowels and breathing had been tagged as voiced fricatives or liquids, as shown by figure 2. In order to circumvent this problem, we removed the corresponding models of voiced consonants leading to a supervised automatic HMM-based segmentation. Although breathing was ill-marked as occlusive, vowels seemed have a better response to the automatic segmentation. Results of supervised automatic segmentation led to 4273 (JMA) and 5487 (JFA) voiced

segments used in following segment analysis. JFA's vowel distribution confirms theoretical prediction that laughter is mainly based on central vowels [15].

===

Please insert here Figure 2. Pies of phonetic distributions issued of automatic segmentation.

===

4.2 Acoustic features

The observed variability highlights the need for large sample sizes when studying laughter [1]. Therefore it was important to use statistical analysis over computed continuous acoustic features. For each bout and segment, three types of acoustic features are computed:

- continuous features: energy, loudness, voicing coefficient, pitch (f0), formant frequencies and Rd [8] are data evolving during segment time span.
- static features: mean and standard deviation of previous continuous features are computed. Duration
- dynamic features: 2-order polynomials of Legendre, model temporal evolution of continuous feature trajectories by slope and curve values.

4.3 Segment analysis

Segment analysis exhibits interesting values reported and commented in table 7. laughter syllables are predominantly based on central vowels (/e/ →/a/). They show higher formant frequencies than normal speech vowels because of extreme positions adopted by the vocal tract during laughter in combination with physiological constraints accompanying production of a pressed voice [15].

===

Please insert here Table 1. Segment acoustic analysis of JMA and JFA vowels.

===

4.4 Bout analysis

Acoustic features computed on bouts show several tendencies summed up in the figure 3. Interestingly, some common aspects of verbal speech prosody seem to be present in laughter prosody like negative pitch slope, negative loudness slope, correlation between f0 mean and loudness mean, and positive vowel duration slope. This last aspect relative to final lengthening has to be further confronted to other models that don't take it into account [13]. Mean number of vowels per bout is 5 (5) for JMA and 6 (6) for JFA which correspond to the mean number of syllables that compose verbal prosodic groups. The positive 1^{st} formant frequency slope mean is interpretable as a progressive jaw opening during laughter [14] [15]. Not only variations (social functions) of laughter are brought up by its type [7], but also by its prosody which seems to be guided by the same physiological constraints as verbal speech.

===

Please insert here Figure 3. Bout acoustic analysis of JMA laughter.

===

5 Synthesis part

In order to provide a speaker-like laughter from only one utterance, a unit selection method is used. However no laughter is present in the utterance and one must combine a parametric method to generate realistic laughter from a few units. Observations made in the analysis part are all taken into account for synthesizing laughter. The first phase is composed of automatic segmentation, symbolic analysis and acoustic analysis of the neutral utterance to add laughter. The second phase is a selection algorithm that extracts three segmented phones of the utterance. The third phase designs bout attack and syllables from these phones. The fourth stage is bout prosody synthesis using a parametric model guided by previous bout analysis. Finally, the last phase is prosodic modification of a duplicated signal by a speech processing algorithm. The

overall synthesis process is exemplified in figure 4 and is parameterized by values explained in table 7.

===

Please insert here Table 2. Parameters default values that drive synthesis process.

===

===

Please insert here Figure 4. Example of a laughter synthesis from a neutral utterance.

===

5.1 Phones selection

Once the neutral utterance is phonetically segmented, a unit selection algorithm identifies three phones. The first phone is arbitrarily the occlusive that possesses the maximum positive loudness slope for starting the bout (attack). The second selected phone is also an occlusive (because of automatic segmentation analysis results) but with the minimum absolute loudness slope, which emulates the breathing part of syllables. The third selected phone is a vowel, preferably /a/, then central vowel, then nasal vowel, which satisfies following acoustic constraints: minimum f0 slope, minimum voicing coefficient and minimum Rd mean. The three phones are balanced regarding attack, breath and vowel relative loudness parameters (P1,P2,P3).

5.2 Signal duplication

The vowel is truncated if its duration is longer than the maximum number of periods parameter (P4), using pitch and duration. The first syllable, called attack, is made of the concatenation of the first occlusive and of the vowel. Other syllables are made of the concatenation of the breathing (second occlusive) and of the vowel. Before duplication of this syllable in a number of syllables (P5), the syllable is energy-windowed by a Tukey's window that eliminates occlusive attack and fades out the vowel.

5.3 Bout prosody generation

Bout prosody generation comes from an empiric parameterizable mathematical model (as [13]) that is inspired by the bout analysis part and that is used for providing every transformations factors. A normalized linear function takes decreasing values over the number of syllables and gives the overall movement of the laughter. A triangularly windowed random signal is added to the linear function, in order to simulate laughter variability [7].

5.4 Signal transformation

The same generated abstract prosodic function is used to provide transformation factors used by phase vocoder technology [5] to transform the duplicated signal. Every syllable is time-stretched, transposed, gained and frequency warped to modify respectively rhythm, intonation, loudness and articulation degree [3] (jaw opening in this case) of the laughter bout. The same function is used to generate all factors to respect the natural and physiological correlation of speech production that seems to be as relevant as in verbal speech (see part 4.4). Furthermore, parameters range (P6, P7, P8, P9) can be automatically estimated on the neutral utterance.

5.5 Results

The randomness of the selected syllables and of the prosodic parameters make the results highly variable as demonstrated by some examples that can be heard in attached sound files or at the following address: `http://www.ircam.fr/anasyn/beller`. The quality of the resulting synthesis has not yet been evaluated by perceptive tests, but informal characterization of provided laughter bouts encourages the method. Even if the adequacy of the synthesised laughter in the original statement always remain a difficulty [12], the proposed method resolves partially the problem. The use of segment of the neutral utterance and the limitation of the bout prosody by the physiological constraints measured on the neutral utterance reduces the perceptual distance between the neutral utterance and the speaker-like synthesised laughter. These conditions

are necessary but not sufficient because we believe that a part of the function of a laughter lies in the interaction of its prosody in that of the sentence to which it is attached.

6 Conclusion

In this paper, we presented our motivation for artistic laughter synthesis. Segmental and prosodic analyses were conducted on a laughter corpus of two Japanese speakers. Statistical acoustic feature analysis of the dynamics of laughter emphasize some natural tendencies reliable to the physiological constraints that prevail in verbal speech prosody. The results then lead to the design of a laughter bout prosodic prototype that encounters randomness to simulate variability of laughter. A semi-parametric method to synthesize speaker-like laughter from one neutral utterance was presented. Future works will now focus on the modification of the voice quality as laughter segments are significantly uttered with pressed voice. The presented method allows laughter-speech synthesis that is another part of our future directions.

7 Acknowledgments

The author would like to kindly thank N. Campbell (ATR) for providing the laughter corpus. This work was partially funded by the French RIAM network project VIVOS.

References

1. J.-A. Bachorowski, M. J. Smoski, and M. J. Owren. The acoustic features of human laughter . *Acoustical Society of America Journal*, 110:1581–1597, September 2001.
2. Grégory Beller. Context dependent transformation of expressivity in speech using a bayesian network. In *ParaLing*, Germany, Saarbrücken, August 2007.
3. Grégory Beller, Nicolas Obin, and Xavier Rodet. Articulation degree as a prosodic dimension of expressive speech. In *submitted to Speech Prosody*, Campinas, May 2008.
4. Grégory Beller, Diemo Schwarz, Thomas Hueber, and Xavier Rodet. Speech rates in french expressive speech. In *Speech Prosody*, Dresden, may 2006. SproSig, ISCA.

5. Niels Bogaards, Axel Roebel, and Xavier Rodet. Sound analysis and processing with audiosculpt 2. In *International Computer Music Conference (ICMC)*, Miami, USA, Novembre 2004.
6. N. Campbell, H. Kashioka, and R. Ohara. No laughing matter. In *Interspeech*, pages 465–468., 2005.
7. Nick Campbell. Whom we laugh with affects how we laugh. In *Interdisciplinary Workshop on The Phonetics of Laughter*, 2007.
8. Gunar Fant. The voice source in connected speech,. *Speech Communication*, 22:125–139, 1997.
9. Silke Kipper and Dietmar Todt. Series of similar vocal elements as a crucial acoustic structure in human laughter. In *Interdisciplinary Workshop on The Phonetics of Laughter*, Saarbrucken, August 2007.
10. L. Lamel, J. Gauvain, and M. Eskenazi. Bref, a large vocabulary spoken corpus for french.
11. Andrew Morris. Automatic segmentation. Technical report, IRCAM, 2006.
12. Mark Schrder and Jrgen Trouvain. How (not) to add laughter to synthetic speech. In *Workshop on Affective Dialogue Systems Kloster Irsee*, 2004.
13. Shiva Sundaram and Shrikanth Narayanan. Automatic acoustic synthesis of human-like laughter. *The Journal of the Acoustical Society of America*, 121:527–535, January 2007.
14. J. Sundberg and J. Skoog. Jaw opening, vowel and pitch. In *STL-QPSR*, volume 36, pages 043–050, 1995.
15. D.P. Szameitat, C.J. Darwin, A.J. Szameitat, D. Wildgruber, A. Sterr, S. Dietrich, and K. Alter. Formant characteristics of human laughter. In *Interdisciplinary Workshop on The Phonetics of Laughter*, Saarbrucken, 2007.
16. Jürgen Trouvain. Segmenting phonetic units in laughter. In *ICPhS*, Barcelona, 2003.
17. Christophe Veaux, Grégory Beller, and Xavier Rodet. Ircamcorpustools: an extensible platform for speech corpora exploitation. In *submitted to LREC*, 2008.

acoustic feature	JMA value (std)	JFA value (std)	comment
voicing coef. mean	0.18 (0.14)	0.16 (0.15)	weakly voiced
Rd mean	1.24 (0.35)	1.37 (0.31)	pressed voice quality
f0 mean	154 (49)	307 (104)	normal register
f0 slope mean	-4.8 (13.1)	-4.7 (42.3)	not significant for JFA
duration mean	0.08 (0.03)	0.09 (0.05)	sexe independent
mean number of periods	11.8 (6.5)	27.8 (21.5)	mean(duration) * mean(f0)
1^{st} formant freq. mean	305 (193)	351 (114)	central
2^{nd} formant freq. mean	1486 (356)	1524 (570)	vowel
3^{rd} formant freq. mean	2588 (435)	2622 (605)	/e/

Table 1. Segment acoustic analysis of JMA and JFA vowels.

parameter ID	name	default value	unity
P1	attack relative loudness	0.5	normalized
P2	breath relative loudness	0.1	normalized
P3	vowel relative loudness	1.0	normalized
P4	maximum number of periods during vowel	15	integer
P5	number of syllables	5	integer
P6	time stretch start and end values	$0.8 \rightarrow 1.2$	slower if > 1
P7	transposition start and end values	$1.5 \rightarrow 0.8$	higher if > 1
P8	gain start and end values	$1 \rightarrow 0$	louder if > 1
P9	formant warping function start and end values	$400 \rightarrow 500$	[Hz] displaced frequency zone

Table 2. Parameters default values that drive synthesis process.

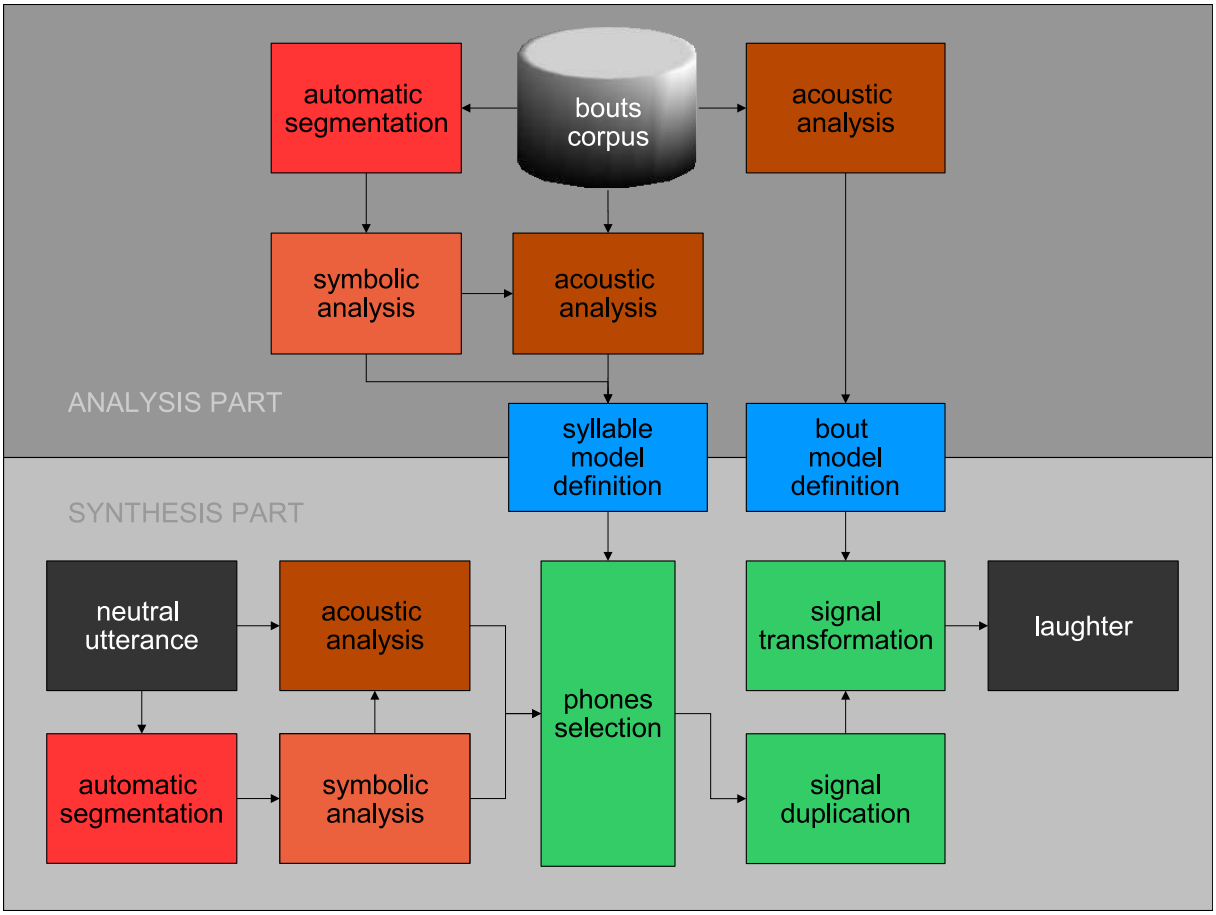

Fig. 1. Overview of the semi-parametric synthesis method.

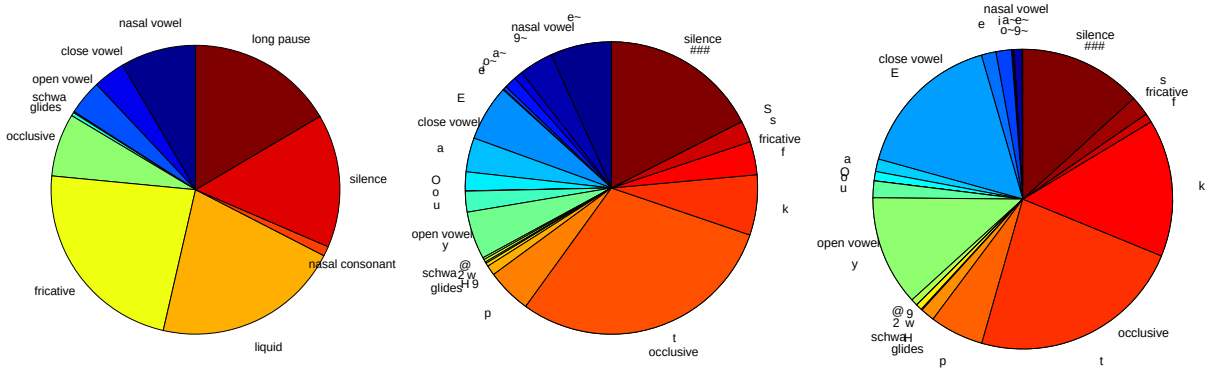

Fig. 2. Pies of phonetic distributions issued of automatic segmentation. Left: Unsupervised segmentation of JMA. Center: Supervised segmentation of JMA. Right: Supervised segmentation of JFA.

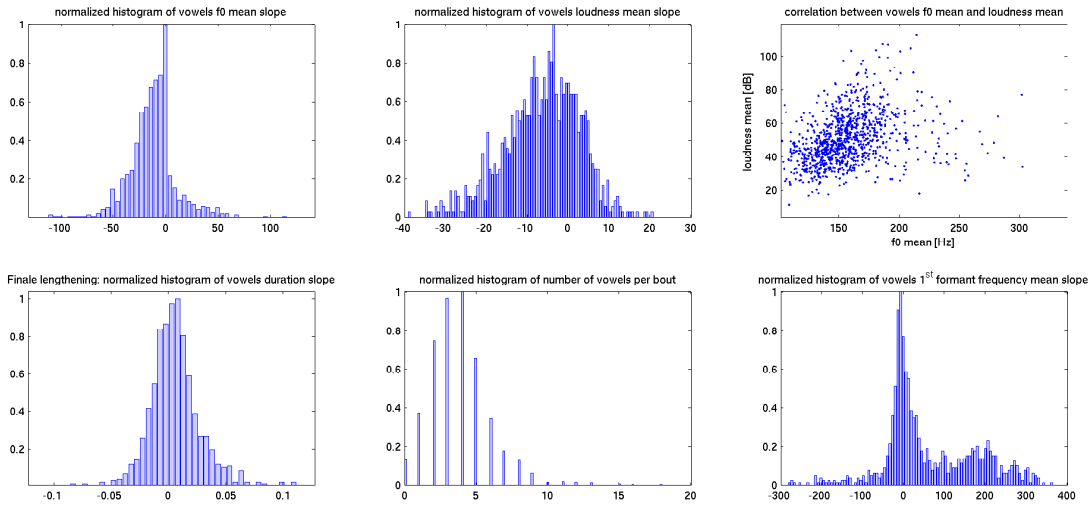

Fig. 3. Bout acoustic analysis of JMA laughter.

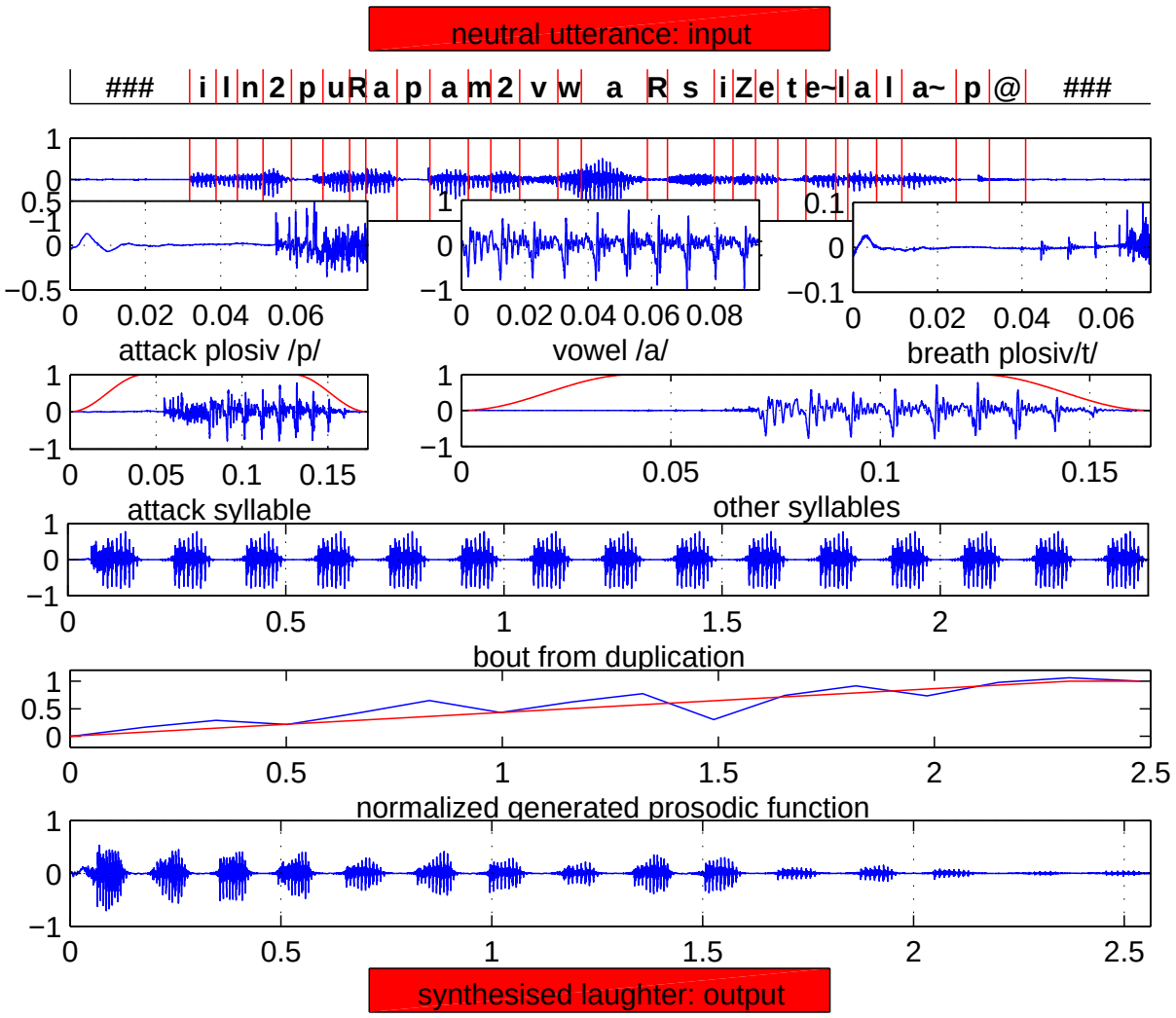

Fig. 4. Example of a laughter synthesis from a neutral utterance. The number of syllables is 15 to supply a better visualization of the normalized generated prosodic function. Neutral utterance, bout from duplication and synthesised laughter can be listened to in attached sound files.