The AVLaughterCycle Database

Jérôme Urbain*, Elisabetta Bevacqua†, Thierry Dutoit*, Alexis Moinet*
Radoslaw Niewiadomski†, Catherine Pelachaud†, Benjamin Picart*, Joëlle Tilmanne*
Johannes Wagner‡

*Université de Mons, Faculté Polytechnique, TCTS Lab
20, Place du Parc - 7000 Mons, Belgium
jerome.urbain, thierry.dutoit, alexis.moinet, benjamin.picart, joelle.tilmanne@umons.ac.be

†CNRS - LTCI UMR 5141, Institut TELECOM - TELECOM ParisTech
37/39, rue Dareau - 75014 Paris, France
niewiado, bevacqua, pelachaud@telecom-paristech.fr

‡ Institut für Informatik, Universität Augsburg
Universitätsstr. 6a - 86159 Augsburg, Germany
johannes.wagner@informatik.uni-augsburg.de

Abstract

This paper presents the large audiovisual laughter database recorded as part of the AVLaughterCycle project held during the eNTERFACE'09 Workshop in Genova. 24 subjects participated. The freely available database includes audio signal and video recordings as well as facial motion tracking, thanks to markers placed on the subjects' face. Annotations of the recordings, focusing on laughter description, are also provided and exhibited in this paper. In total, the corpus contains more than 1000 spontaneous laughs and 27 acted laughs. The laughter utterances are highly variable: the laughter duration ranges from 250ms to 82s and the sounds cover voiced vowels, breath-like expirations, hum-, hiccup- or grunt-like sounds, etc. However, as the subjects had no one to interact with, the database contains very few speech-laughs. Acted laughs tend to be longer than spontaneous ones and are more often composed of voiced vowels. The database can be useful for automatic laughter processing or cognitive science works. For the AVLaughterCycle project, it has served to animate a laughing virtual agent with an output laugh linked to the conversational partner's input laugh.

1. Introduction

Laughter is an essential signal in human communications. It conveys information about our feelings and helps to cheer up our mood. Moreover, it is communicative, eases social contacts and has the potential to elicit emotions to its listeners. It is thus a very important signal to detect for applications dealing with the users' moods, as well as a crucial signal to synthesize if one wants to design a machine with human expressive capabilities.

In consequence, automatic laughter processing has gained in popularity during the last decades. If a few systems able to distinguish between laughter and speech have recently been built on the recognition side (e.g. (Truong and van Leeuwen, 2007; Knox and Mirghafori, 2007; Petridis and Pantic, 2009)), automatic laughter synthesis is still inefficient. Interesting approaches have been explored to generate human-like laughs (e.g. (Lasarcyk and Trouvain, 2007; Sundaram and Narayanan, 2007)), but perceptive tests have shown that the resulting laughs do not sound natural. They miss an important characteristic of human laughs: variability.

The AVLaughterCycle project, launched during the eNTERFACE'09 Workshop held in Genova, aims at developing an audiovisual laughing machine, capable of recording the laughter of a user and to respond to it with a virtual agent's laughter linked with the input laughter. This goal implies three tasks: laughter detection, laughter analysis/classification (to link the output laugh with the input) and audiovisual laughter (copy-)synthesis. To perform theses tasks, an audiovisual laughter database has been recorded. The aim of this database is to provide a broad corpus for studying the acoustics of laughter, the facial movements involved, and the synchronization between these two signals. During the Workshop, the laughter database has been used to drive the facial movements of a 3D humanoid virtual character, Greta (Niewiadomski et al., 2009), simultaneously with the audio laughter signal. This paper presents the database itself.

2. Goal and organisation of the paper

The database, recorded as part of the AVLaughterCycle project, is meant to be useful for many researches about laughter. To our knowledge, it is the first database of laughter combining both the acoustic signal and facial motion tracking. The paper aims at deeply presenting the AVLaughterCycle Database in order to provide all the information required by researchers willing to use it. The paper is organized as follows. Section 3 gives information about the database participants. Section 4 presents the stimuli used to elicit laughter. The database recording protocol is detailed in Section 5. Section 6 focuses on the devices used for face motion tracking. The corpus annotation is explained in Section 7. Section 8 gives an overview of the database contents. Advantages and limitations of the database are discussed in Section 9. Finally, potential applications are presented in Section 10, before the conclusion (Section 11).

3. Participants

24 subjects participated in the database recordings: 8 (3 females, 5 males) with the ZignTrack (Zign Creations, 2009) setting and 16 (6 females, 10 males) with the OptiTrack (Natural Point, Inc., 2009) setting (see Section 6). They came from various countries: Belgium, France, Italy, UK, Greece, Turkey, Kazakhstan, India, Canada, USA and South Korea. The female, male and overall average ages were respectively 30 (standard deviation: 7.8), 28 (sd: 7.1) and 29 (sd: 7.3). All the participants gave written consent to use their data for research purposes.

4. Stimuli

It is strongly suspected that there is a difference between the expressions of real and acted emotions (e.g. (Wilting et al., 2009; Douglas-Cowie et al., 2003)). To collect a corpus representative of humans' natural behaviours, one should capture the data in a natural environment, the subjects being unaware of the database collection until the end of the recording. Laughter being an emotional signal, it is affected by the same phenomenon: one cannot expect natural laughter utterances by simply asking subjects to laugh.

To find spontaneous laughter utterances, it is popular to take the laughs recorded while collecting data for another purpose. For example, (Truong and van Leeuwen, 2007), (Knox and Mirghafori, 2007) and (Kennedy and Ellis, 2004) use the ICSI Meeting Corpus (Janin et al., 2003), recorded for studying speech in general by placing microphones in meeting rooms. This corpus contains a significant number of laughs, which are assumed spontaneous since they occur in regular conversations (even though the participants knew there were microphones). When for some reason natural data cannot be used, it is common to try to induce laughter - and not tell beforehand that laughter is the object of the study - rather than instructing to laugh. One way to achieve it is to display a funny movie (Trouvain, 2003).

In our case, both audio recording and accurate facial motion tracking were desired. To our knowledge, there existed no laughter database providing these 2 signals. Due to the markers required for facial motion tracking, a natural laughter recording was impossible. To push the participants towards spontaneous laughter, a 13-minutes funny movie was created by the concatenation of short videos found on the Internet.

5. Database recording protocol

Participants were invited to sit in front of a computer screen, used to display the funny movie. They wore a headset microphone for audio recording and stimuli listening. A webcam was placed on top of the screen, recording 25 frames per second (FPS) with a 640x480 resolution, stored in RGB 24 bits. The audio sampling frequency was set to 16kHz, stored in PCM 16 bits. The material for facial motion capture will be presented in Section 6..

The database was recorded through University of Augsburg's Smart Sensor Integration (SSI) (Wagner et al., 2009). This software enables the synchronization between the different input signals (here, microphone and webcam), handles the stimuli display and can process the signals to segment and label interesting parts. SSI was also used for the database annotation (Section 7).

Participants were asked to relax, watch the video and react freely to it, with two limitations: they should try to 1) keep their head towards the screen, and 2) not put anything between their head and the webcam (e.g hands), otherwise the facial tracking would fail. All the instructions were displayed on the screen before the experiment. Once the protocol was clear, participants were left alone in the experiment room and started the stimuli playing. For synchronisation and data saving reasons, the protocol had to be slightly modified when using OptiTrack (see Section 6.2).

At the end of the movie, subjects were instructed to perform one acted laughter, pretending they had just seen something hilarious. The objective of these acted laughs is to provide some material to analyse the differences between spontaneous and acted laughs, to determine whether the subjects, when acting, tend to mimic the spontaneous laughs they had just performed, etc.

6. Facial motion capture

Since markerless facial motion tracking is nowadays not reliable enough to capture the small variations of facial expression during laughter, we turned towards techniques using markers placed on the subject's face. Two systems have been successively used, ZignTrack and OptiTrack.

6.1. ZignTrack

ZignTrack (Zign Creations, 2009) uses one single camera to realize the 3D tracking, which is an extrapolation from a 2D image, using a fixed face template. Facial features are marked with simple stickers or make-up (Figure 1). ZignTrack presents the advantages of being cheap and requiring few material, but has several drawbacks: the extrapolation from 2D causes head distortions, the tracking fails when there are rapid movements and is unable to recover after an erroneous frame. To obtain the accurate facial motion, a lot of manual corrections are then needed (several hours per recording). For these reasons, we turned towards a more professional device, OptiTrack, after the first 8 recordings.

6.2. OptiTrack

OptiTrack (Natural Point, Inc., 2009) uses 7 synchronized infrared cameras: 6 placed in a semi-circular way for facial motion capture and an additional one for scene audiovisual recording. Each camera contains a grayscale CMOS imager capturing up to $100FPS$. Infrared reflectors need to be stuck on the skin (Figure 2). For the recordings performed with the OptiTrack device, the infrared cameras were added to the previous setting (Figure 3, with the OptiTrack cameras highlighted by circles). Participants were asked to clap their hands in order to synchronize the facial motion tracking with the audio and webcam signals. OptiTrack provided high quality tracking with few manual corrections required. However, the data acquisition sometimes stopped after around 5 minutes. To make sure the data of the whole experiment would be usable, it was then decided to shorten the stimuli video to 10 minutes and to split it into 3 parts slightly longer than 3 minutes. At the beginning of each

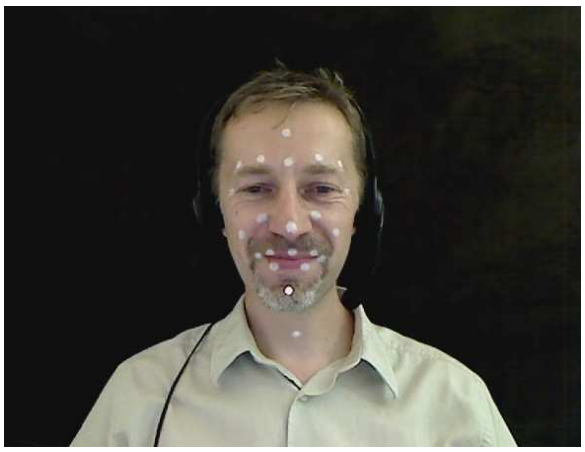

Figure 1: Markers drawn for facial motion tracking using ZignTrack

Figure 2: Infrared markers placed for facial motion tracking using OptiTrack

session, the instructor started the face motion acquisition system, left the room and the subject clapped for synchronisation with the other signals. At the end of each session, the subject was again instructed to clap, so that the instructor would enter the room and stop the face motion tracking. The microphone and webcam recorded the experiment from the beginning of the first session to the end of the third session, without interruption.

7. Database annotation

The recorded data have been labeled by one annotator, using SSI. A hierarchical annotation protocol was designed: segments receive the label of one main class (laughter, breath, verbal, clap or trash; silence being the default class) and "sublabels" can be concatenated to give further details about the segment. Laughter sublabels characterize both:

- The laughter temporal structure, following the 3 segmentation levels presented in (Trouvain, 2003). These sublabels indicate whether the *episode* (i.e. the full laughter utterance) contains:
  - several *bouts* (i.e. parts separated by inhalations), it is then annotated with the sublabel "episode";
  - only one syllable, labeled as "monosyllabic';
  - several syllables but only one bout - default category (no particular sublabel).

  These sublabels are mutually exclusive.
- The laughter acoustic contents, through labels describing the type of sound: vowel, breathy (oral exhalation), nasal exhalation, grunt-like, hum-like, "hiccup-like", speech-laughs or laughter that are mostly visual (quasi-silent). These sublabels can be combined to reflect content changes during the laughter episode.

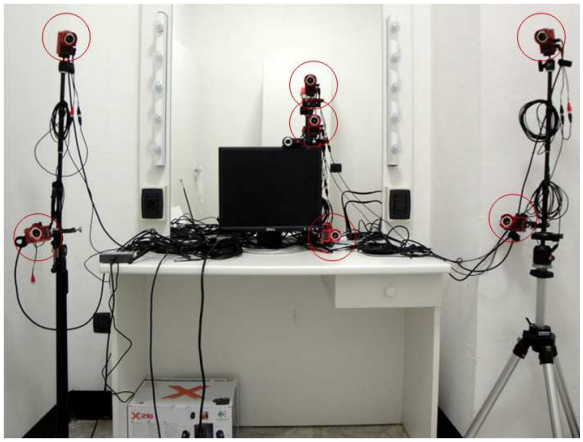

Figure 3: Desktop setup for database recording. Optitrack cameras are highlighted by circles.

For example, a laughter episode composed of several bouts, starting by grunt-like sounds and followed by hiccup-like sounds is annotated: $laugh\_episode\_grunt\_hiccup$.
To cope with exceptional conflicts that might influence the classes models when training a classifier (for example when there is a strong noise in the middle of a laughter episode), a "discard" main class has been added.
The annotation primarily relies on the audio, but the video is also looked at, to find possible neutral facial expressions at the episode boundaries or annotate silent laughs. In addition, laughs are often concluded by an audible inspiration, sometimes several seconds after the laughter main part. When such an inhalation, obviously due to the preceding laughter, can be found after the main audible part, it is included in the laughter segment.

8. Database contents

The number of occurrences of the main classes over the full recordings or only inside the stimuli sessions are presented in Table 1. Subjects spend, in average, 21.8% of the stimuli sessions laughing, which is a huge proportion. The number of laughter episodes per participant stands around 42, with extreme values of 4 and 82, for a total of 1021 episodes inside the dedicated stimuli sessions. The database contains 27 acted laughs, uttered by 22 subjects (2 subjects did not produce an acted laughter).

Main class	Occurrences	
	Full database	Stimuli sessions
Laughter	1066	1021
Trash	267	207
Verbal	186	64
Clap	93	1
Breath	41	31
Discard	31	23

Table 1: Occurrences of the main classes

8.1. Laughter kinds

Table 2 presents the occurrences of the laughter sublabels, for the 1021 laughs elicited by the stimuli sessions, considered as spontaneous, as well as the 27 acted laughs. It is important to remember that the acoustic content sublabels can be combined to specify different contents in an episode. This explains why the total number of laughter sublabels is larger than the number of occurrences in the laughter class. On a structural level, it appears that most laughs contain several syllables forming one single bout. Monosyllabic utterances are relatively frequent (17.5%) when subjects laugh spontaneously, but no subject produced a monosyllabic laugh when asked to pretend he had witnessed something hilarious. Episodes with several bouts separated by inhalations occurr from time to time spontaneously and with a larger proportion when acting.

Regarding the acoustic contents, it can be seen that the spontaneous laughs cover a broad variety of sounds: the labels are spread over all the laughter kinds. One third of the annotations reflect a vowel-like content, and the vowel 'a' is the most frequent one. Nasal exhalations represent 20% of the annotations. Other categories like breathy (oral exhalation), hum-like, hiccup-like or even silent laughing are also well represented. However, the database contains only 20 speech-laughs [1]. This can be explained by the fact that the subjects were left alone and had nobody to interact with: there is few speech in the stimuli sections, hence few speech-laughs.

The acoustic content sublabel occurrences are different when considering acted laugh. There, voiced vowel are clearly the most frequent annotations. This might be due to the stereotypical image of laughter ("hahaha").

8.2. Laughter duration

Excluding laughs involving speech, the average duration of a spontaneous laughter episode is 3.5s (standard deviation: 5.3s; median: 2.2s). A histogram of the spontaneous laughter duration and its cumulative distribution function are presented in Figure 4. The large majority (83%) of the laughter episodes lasts less than 5s, but longer episodes should not be neglected as they represent 51.4% of the total laughter duration and are the most striking ones. The longest giggle in the database lasts 82s.

Acted laughs tend to be longer. Their mean duration is 7.7s (median: 5.26; std: 5.94). A t-test assuming the 2 samples come from normal distribution with unknown, unequal variances, shows the difference between the mean duration of spontaneous and acted laughs is highly significant ($p = 0.0012$). Using a t-test might seem daring since the duration distribution is clearly not Gaussian and the number (27) of acted laughs is not sufficient to use the Central Limit Theorem with full confidence, but the outcome of the t-test is strengthened by a Kolmogorov-Smirnov test (measuring whether the 2 samples are likely to belong to the same population, without any assumption on their distribution), which states the spontaneous and acted laughs belong to different distributions with high significance ($p = 1.1 \cdot 10^{-8}$).

[1] We should even state that in 9 out of the 20 cases, speech and laugh do not overlap but follow each other so closely that it is impossible to separate them.

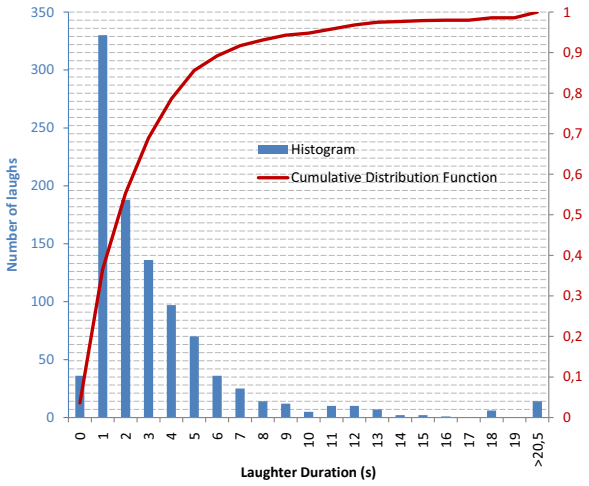

Figure 4: Histogram and cumulative distribution function of the laughter duration

9. Limitations and benefits

The biggest limitation of the AVLaughterCycle database might reside in the absence of active communication provided by the subjects. Unlike popular databases used in laughter processing like the ICSI Meeting Corpus (Janin et al., 2003) or the AMI Meeting Corpus (Carletta, 2007), participants had no one to interact with. It has been shown that the conversational partners influence the way we laugh (Campbell, 2007). The laughs from the AVLaughterCycle database, obtained without conversational partners, might be considered as the "base" laughs from our participants, when they are alone watching a movie, and we have no guarantee these people would laugh the same way when they interact with other people. The most dramatic consequence of the absence of interaction is the very small number of speech-laughs, much less than in human conversations (speech-laughs are as numerous as breath-laughs in the ICSI Meeting Corpus). In addition, people knew they were being recorded, which is the case of many databases. The protocol (stimulus induction, etc.) was meant to elicit as natural reactions as possible given the constraint of wearing markers on the face. The main benefits of the database are: the number of laughs (over 1000), their variety both in duration and acoustic contents, the presence of visual information (the AMI Meeting Corpus includes video but not face motion tracking) and the annotation focusing on

Category	Laughter sublabel	Occurrences	
		Spontaneous	Acted
	TOTAL UTTERANCES	1021	27
Temporal structure	Monosyllabic	179	0
	One bout (several syllables)	677	14
	Several bouts	165	13
	TOTAL	1021	27
Acoustic content	Vowel: a	277	18
	Vowel: e	101	5
	Vowel: i	37	1
	Vowel: o	26	0
	Vowel: u	5	2
	Nasal exhalation	277	1
	Breathy (oral exhalation)	237	2
	Hum-like	169	2
	Hiccup-like	95	5
	Grunt-like	18	1
	Speech-laugh	20	0
	Silent	94	1
	TOTAL	1359	38

Table 2: Occurrences of the laughter sublabels for the spontaneous and acted laughs

laughter. Finally, the database contains some acted laughs, recorded by the same subjects at the end of the experiment.

10. Applications

In the AVLaughterCycle project, the database has been used to endow a virtual agent, Greta, with the capability of laughing. In a nutshell, the application consists in acoustically detecting incoming laughter, compare the laugh with the utterances of the AVLaughterCycle Database to select an appropriate answering laughter (so far, spectral features are computed and the closest vector in the database is selected) and animate Greta with the facial movement of that selected laughter, synchronously with the audio playing. More details can be found in the eNTERFACE Project report (Urbain et al., 2009).

Following this application, the AVLaughterCycle will be used to improve acoustic laughter detection/recognition and classification/clustering as well as the animation of a laughing avatar. Other signal processing potential uses include laughter synthesis, studies on the synchronisation between audio and facial movements while laughing, etc.

The database could also interest people working on laughter or humour from cognitive science points of view: studying how different persons react to humour, how laughs follow each other, whether there are periods where the subject is particularly prone to laugh, comparing spontaneous and acted laughs, etc.

11. Conclusion

A large laughter database had been recorded and presented in this paper. 24 subjects participated and produced more than 1000 spontaneous laughs, elicited by a funny stimulus. The main drawback of the database is the absence of speech-laughs. On the other hand, the presence of facial motion data in addition to the acoustic signal makes this database unique. The corpus covers a broad variety of laughter kinds. Regarding the laughter duration and structure, laughs composed of one single bout but several syllables are the most frequent, but the corpus contains numerous examples of monosyllabic or longer episodes. Regarding the acoustic contents, the annotations are spread over all the categories, led by voiced vowels, nasal and oral breath-like sounds. The database can be used for various research purposes: audio and/or visual laughter recognition or synthesis, etc. The corpus is freely available from http://www.tcts.fpms.ac.be/∼urbain/.

Acknowledgments

The authors would like to thank all the database participants.

This project was partly funded by the European IP 6 project CALLAS and by the Ministry of Région Wallonne under the Numediart research program (grant $N^0716631$). J.Tilmanne receives a PhD grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

12. References

N. Campbell. 2007. Whom we laugh with affects how we laugh. In *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, pages 61–65, Saarbrucken, Germany, August.

J. Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41(2):181–190.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40:33–60.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus.

In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong-Kong, April.

L. Kennedy and D. Ellis. 2004. Laughter detection in meetings. In *NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, May.

M. T. Knox and N. Mirghafori. 2007. Automatic laughter detection using neural networks. In *Proceedings of Interspeech 2007*, pages 2973–2976, Antwerp, Belgium, August.

E. Lasarcyk and J. Trouvain. 2007. Imitating conversational laughter with an articulatory speech synthesis. In *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, pages 43–48, Saarbrucken, Germany, August.

Natural Point, Inc. 2009. Optitrack - optical motion tracking solutions. http://www.naturalpoint.com/optitrack/, Consulted on October 20,.

Zign Creations. 2009. Zign track - the affordable facial motion capture solution. http://www.zigncreations.com/zigntrack.html, Consulted on October 20,.

Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. 2009. Greta: an interactive expressive ECA system. In Carles Sierra, Cristiano Castelfranchi, Keith S. Decker, and Jaime Simão Sichman, editors, *8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary, May 10-15, 2009, Volume 2*, pages 1399–1400. IFAAMAS.

Stavros Petridis and Maja Pantic. 2009. Is this joke really funny? judging the mirth by audiovisual laughter analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1444–1447, New York, USA, June.

S. Sundaram and S. Narayanan. 2007. Automatic acoustic synthesis of human-like laughter. *Journal of the Acoustical Society of America*, 121(1):527–535, January.

Jurgen Trouvain. 2003. Segmenting phonetic units in laughter. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 2793–2796, Barcelona, Spain, August.

K. P. Truong and D. A. van Leeuwen. 2007. Automatic discrimination between laughter and speech. *Speech Communication*, 49:144–158.

Jérôme Urbain, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Radoslaw Niewiadomski, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner. 2009. AVLaughterCycle: An audiovisual laughing machine. In *Proceedings of eNTERFACE'09 (To appear)*.

Johannes Wagner, Elisabeth André, and Frank Jung. 2009. Smart sensor integration: A framework for multimodal emotion recognition in real-time. In *Affective Computing and Intelligent Interaction (ACII 2009)*.

Janneke Wilting, Emiel Krahmer, and Marc Swerts. 2009. Real vs. acted emotional speech. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 ICSLP)*, pages 805–808, Pittsburgh, USA, September.