Emotional Speech Datasets for English Speech Synthesis Purpose : A Review

Noé Tits, Kevin El Haddad and Thierry Dutoit

University of Mons,
Numediart Institute,
Mons, Belgium 7000
{noe.tits, kevin.elhaddad, thierry.dutoit}@umons.ac.be

Abstract. In this paper, we review the datasets of emotional speech publicly available and their usability for state of the art speech synthesis. This is conditioned by several characteristics of these datasets: the quality of the recordings, the quantity of the data and the emotional content captured contained in the data. We then present a dataset that was recorded based on the observation of the needs in this area. It contains data for male and female actors in English and a male actor in French. The database covers 5 emotion classes so it could be suitable to build synthesis and voice transformation systems with the potential to control the emotional dimension.

Keywords: Speech Synthesis, Emotional Speech, Deep Learning

1 Introduction

One of the major components of human-agent interaction systems is the speech synthesis module. The state-of-the-art speech synthesis systems such as wavenet and tacotron [1–3] are giving impressive results. They can produce, intelligible, expressive, even human-like speech. But, they cannot yet be used to control the emotional dimensionality in speech which is a crucial parameter in order to obtain human-like controllable speech synthesis system.

Although still being relatively neglected by the affective computing community, the interest for emotional speech synthesis systems has been growing for the past two decades. After the improvement parametric systems brought to this field [4, 5], deep learning-based systems were also employed for this task.

One of the problems in the emotional speech synthesis research community is the lack of publicly available data and the difficulty to collect them. In fact, to the best of our knowledge, no emotional speech database for synthesis purpose and suitable for deep learning systems is publicly available. In this paper, we try to tackle this problem.

In what follows we will present a review of emotional speech datasets in Section 2. We will then describe the motivations for collecting a new database in

Section 3 and detail the content of a newly released database [1] that fulfill these motivations in Section 4.

2 Review

Emotions can be represented in different ways. A first representation, is Ekman's six basic emotion model [6] which identify anger, disgust, fear, happiness, sadness and surprise as six basic emotions from which the other emotions may be derived. Emotions can also be represented in a multidimensional continuous space like in the Russels circomplex model [7] (valence and arousal being the currently most famous dimensions used). A more recent way of representing emotions is based on ranking which prefer a relative preference method to annotate emotions rather than labeling them with absolute values [8].

Several open-source databases can be found but to the best of our knowledge, none is really suitable for emotional speech synthesis purpose. In this section we will explain why and mention some examples.

The RAVDESS database emotional data for 24 different actors [9]. The actors were asked to read 2 different sentences in a spoken and sung way in North American English. The spoken style was recorded in 8 different emotional styles: neutral, calm, happy, sad, angry, fearful, disgust, surprise. Each utterance was expressed at 2 different intensities each (except for the neutral emotion) and 2 times thus giving a total of 1440 files. A perception test was then undertook to validate the database on the emotional categories, intensity and genuineness.

The CREMA-D database [10] is similar to the RAVDESS. For this database, 12 different sentences were recorded by 91 different actors, for the 6 basic emotions:happy, sad, anger, fear, disgust, and neutral. Only one of the 12 sentences was expressed in 3 different intensities, for the other 11, the intensity was not specified. The authors report 7442 files in total. This database was also validated through perception tests and helped validate the emotion category and intensity.

Also similar to the previous ones, the GEMEP database [11] is a collection of 10 French-speaking actors, recorded uttering 15 different emotional expressions at three levels of intensity, in three different ways: improvised sentences, pseudo-speech, and nonverbal affect bursts. This database counts a total of 1260 audio files. It was also validated through perception tests.

The Berlin Emotional Speech Dataset [12] contains the recording of 10 different utterances by 10 different actors in 7 different emotions (neutral, anger, fear, joy, sadness, disgust and boredom) in German, making it a total of 800 utterances (counting some second version of some of the sentences). This database was, like the previous ones, validated using perception experiments.

These databases are not suitable for current state of the art speech synthesis purpose because of the limited amount of sentences recorded.

Moreover, the six basic emotions do not really occur in daily conversations. Indeed, in Ekman's model, on which the choice of emotions was based for these

[1] https://github.com/numediart/EmoV-DB

datasets, the basic emotions are the ones from which other emotions derive. But that does not necessarily mean that they are frequently expressed in speech in our daily interactions.

The IMPROV [13] and IEMOCAP [14] databases both contain a large amount of diverse sentences of emotional data. IEMOCAP contains audio-visual recordings of 5 sessions of dyadic conversations between a male and a female subjects. In total it contains 10 speakers and 12 hours of data. IMPROV contains 6 sessions from 12 actors resulting in 9 hours of audiovisual data. Both databases were evaluated in terms of category of emotions [6] and emotional dimensions [7] by several subjects. However they are not suitable for synthesis purpose either because although the data is well recorded and post-processed it contains overlapping speech due to the data recording setup (dyadic conversation) and some external noise.

The CMU Arctic Speech Database [15] and the SIWIS French Speech Synthesis Database [16] are collections of read utterances of phonetically balanced sentences in English and French respectively. The CMU-Arctic database contains approximately 1150 sentences recorded from each of 4 different speakers while SIWIS contains a total of 9750 utterances from a single speaker. These are database suitable for speech synthesis purpose as there is a large amount of different sentences recorded from a single speaker in noiseless environment. However the sentences are neutral and do not express any emotions.

The AmuS database contain audio data dedicated to amused speech synthesis [17]. We showed in previous work [18–20] that this database was well suited for amused speech synthesis. But AmuS contains data only for amused speech and not other emotions.

3 Motivations

This database's [2] primary purpose is to build models that could not only produce emotional speech but also control the emotional dimension in speech [21, 22]. The techniques to allow this are either text-to-speech like systems where the system would map a given text sentence to a speech audio signal or voice transformation systems where a source voice would be converted to a specific target emotional voice. Considering this, it is obvious that a lot of data is required. One of the primary difficulties of building emotional speech-based generation systems is the collection of data. Indeed not only must the recording be of good quality and noise free, but the task of expression emotional sentences in a large enough amount is challenging. Also it is often preferable concerning these types of systems, that a certain category of emotion contains data that are similar on the acoustic level.

The database presented here was built with these requirements in mind. The aim was also for it to fit with other currently open-source databases to maximize the quantity of data available. As mentioned previously, the CMU-Arctic database (English) and the SIWIS (French) databases are two datasets

[2] https://github.com/numediart/EmoV-DB

of neutral speech. Each of them contain a relatively large amount of data that can be used as source voices for a voice conversion system or as pre-training data for a system. They are also transcribed which makes the transcription also available for our database. The transcribed utterances as well as annotations at phonetic level are available. A subset of these were used to build our database. The phonetic annotations are not time-aligned with our data yet, but methods can be used such as forced alignment systems [23].

We chose five different emotions: amusement, anger, sleepiness, disgust and neutral. We chose emotions that are more likely to be expressed in daily conversations than Ekman's basic emotions. These emotions were chosen because of the ease to produce them by actors and in order to cover a diverse space in the Russel Circumplex to allow experimenting with interpolation techniques to obtain intermediate emotions.

4 Database Content

The data was recorded in 2 different languages English (North American) and French (Belgian). English natives (2 females and 2 males) and a single male French native were asked to read sentences while expressing one of the above mentioned emotions. The English sentences were taken from the CMU-arctic database. The French ones from the SIWIS database. Both databases contain freely available open-source phonetically balanced sentences.

The recordings for the English data were carried on in two different anechoic chambers of the Northeastern University campus. The ones for the French data were made in an anechoic room at the University of Mons.

The utterances were recorded in several sessions of about 30 minutes recordings followed by a 5 to 15 minutes break and the data collection was spread across several days depending on the availability of the actors. The actors were asked to repeat sentences that were mispronounced.

The actors were asked to record each emotion class separately in different sessions. At the moment of redaction of this article, the sentences were segmented manually for some of the speakers (annotation and segmentation is still ongoing). By segmentation we mean determining the intervals of start and end of each sentence. The total number of utterances obtained is summarized in Table 1.

Table 1. Gender and language of recorded sentences from each speaker and amount of utterances segmented per speaker and per emotion.

Speaker	Gender	Language	Neutral	Amused	Angry	Sleepy	Disgust
Spk-Je	Female	English	417	222	523	466	189
Spk-Bea	Female	English	373	309	317	520	347
Spk-Sa	Male	English	493	501	468	495	497
Spk-Jsh	Male	English	302	298	-	263	-
Spk-No	Male	French	317	-	273	-	-

Amused speech can contain chuckling sounds which overlap and/or intermingle with speech called speech-laughs [24] or can be only amused smiled speech [5]. So, for the amused data in our database, in order to collect as much data as possible and considering the relatively limited time the actors provided us, we focused on amused speech with speech-laughs. This choice was motivated by our previous study showing that this type of amused speech was perceived as more amused than amused smiled speech (without speech-laugh). Also in another study, we show that including laughter in synthesized speech is always perceived as amused no matter the style of speech it is inserted in (neutral or smiled) [20]. Based on the previous studies made on amusement, the actors were encouraged, while simulating the other emotions, to use nonverbal expressions [25] before and even while uttering the sentences if they felt the need to (e.g. yawning for sleepiness, affect bursts for anger and disgust).

Acknowledgments

Noé Tits is funded through a PhD grant from the Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture (FRIA), Belgium.

References

1. Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," in *SSW*, 2016.
2. Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *INTERSPEECH*, 2017.
3. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017.
4. Hiromichi Kawanami, Yohei Iwami, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Gmm-based voice conversion applied to emotional speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003.
5. 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)*, Ljubljana, Slovenia, 4-8 May 2015.
6. Paul Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, Tim Dalgleish and M. J. Powers, Eds., pp. 4–5. Wiley, 1999.
7. Jonathan Posner, James A Russell, and Bradley S. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology.," *Development and psychopathology*, vol. 17 3, pp. 715–34, 2005.
8. G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, Oct. 2017, vol. 00, pp. 248–255.
9. Steven R. Livingstone and Frank A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018.
10. Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.
11. Tanja Bänziger, Marcello Mortillaro, and Klaus R Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception.," *Emotion*, vol. 12, no. 5, pp. 1161, 2012.
12. Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.
13. Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

14. Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.
15. John Kominek and Alan W Black, "The cmu arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.
16. Pierre-Edouard Honnet, Alexandros Lazaridis, Philip N. Garner, and Junichi Yamagishi, "The siwis french speech synthesis database ? design and recording of a high quality french database for speech synthesis," *Online Database*, 2017.
17. Kevin El Haddad, Ilaria Torre, Emer Gilmartin, Hüseyin Çakmak, Stéphane Dupont, Thierry Dutoit, and Nick Campbell, "Introducing amus: The amused speech database," in *Statistical Language and Speech Processing*, Nathalie Camelin, Yannick Estève, and Carlos Martín-Vide, Eds., Cham, 2017, pp. 229–240, Springer International Publishing.
18. Kevin El Haddad, Hüseyin Cakmak, Stéphane Dupont, and Thierry Dutoit, "An HMM Approach for Synthesizing Amused Speech with a Controllable Intensity of Smile," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Abu Dhabi, UAE, 7-10 December 2015.
19. Kevin El Haddad, Hüseyin Cakmak, Stéphane Dupont, , and Thierry Dutoit, "Breath and repeat: An attempt at enhancing speech-laugh synthesis quality," in *European Signal Processing Conference (EUSIPCO 2015)*, Nice, France, 31 August-4 September 2015.
20. 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)*, Brisbane, Australia, 19-24 April 2015, pp. 4939–4943.
21. Noé Tits, Kevin El Haddad, and Thierry Dutoit, "Exploring transfer learning for low resource emotional tts," *arXiv preprint arXiv:1901.04276*, 2019.
22. Noé Tits, Kevin El Haddad, and Thierry Dutoit, "Asr-based features for emotion recognition: A transfer learning approach," *arXiv preprint arXiv:1805.09197*, 2018.
23. Sandrine Brognaux, Sophie Roekhaut, Thomas Drugman, and Richard Beaufort, "Train&Align: A new online tool for automatic phonetic alignment," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 416–421.
24. Jürgen Trouvain, "Phonetic aspects of speech-laughs," in *Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: LHarmattan*, 2001, pp. 634–639.
25. Kevin El Haddad, No Tits, and Thierry Dutoit, "Annotating nonverbal conversation expressions in interaction datasets," in *Proceedings of Laughter Workshop 2018*, 09 2018.