The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech

Ingmar Steiner[1–3], Marc Schröder[2], and Annette Klepp[2,3]

[1]*Cluster of Excellence "Multimodal Computing and Interaction", Saarland University*
[2]*Language Technology Lab, DFKI GmbH, Saarbrücken*
[3]*Computational Linguistics & Phonetics, Saarland University*
steiner@coli.uni-saarland.de

Abstract

The nature of expressive and emotional speech has garnered a mounting body of research over the past decade (Scherer, 2003; Schröder, 2009; Schuller et al., 2011, among many others); a number of research projects have been, or are being, conducted in order to investigate phonetic parameters of expressive speech and to implement their findings in technological applications. Independent scientists in phonetics and related disciplines may however share an interest in this field and the research questions it entails, open or answered but unreplicated. A significant obstacle however is the requirement for speech corpora of appropriate size and content, especially those extensively annotated with linguistic metadata; especially for German, not many such resources are available (cf. however Burkhardt et al., 2005).

This paper presents a corpus of read speech from a single male speaker of German,[1] which contains five distinct speaking styles, viz. *neutral*, *cheerful*, *depressed*, *aggressive*, and a "cool, laid-back" *poker* style. The corpus comprises 3 000 sentences, optimized for phonetic coverage; 400 of these sentences, as well as 150 domain-specific utterances, were recorded in each of the expressive styles. Phone-level segmentation is available for all of the recorded utterances, and the labels were manually checked and corrected where needed.

The corpus has been used for voice conversion (Türk and Schröder, 2010) and to create voices for expressive text-to-speech synthesis (Gebhard et al., 2008; Steiner et al., 2010), which in turn have found use in a number of studies (e.g. Scheffler et al., 2012; Székely et al., 2013). However, the data itself was never made available to the public, and so its use as a resource for the analysis of expressive speech, or as an asset for novel technological applications, was hitherto restricted. With this paper, we announce the availability of the full corpus, free of charge, under

[1] Stefan Röttig, http://www.stefan-roettig.de/

a much more permissive license, in the belief that the scientific community will regard it as a valuable resource for phonetic research and other applications. In the spirit of Rosenberg (2012), we use distributed version control (Torvalds, n.d.) and peer-to-peer data mirroring (Hess, n.d.) to manage the phonetic annotations and speech data, respectively, allowing the corpus to be easily maintained and enhanced, and integrated into other projects as a submodule.

References

Burkhardt, F., A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss (2005). "A database of German emotional speech". In: *Interspeech*. Lisbon, Portugal, pp. 1517–1520. URL: http://www.isca-speech.org/archive/interspeech_2005/i05_1517.html.

Gebhard, P., M. Schröder, M. Charfuelan, C. Endres, M. Kipp, S. Pammi, M. Rumpler, and O. Türk (2008). "IDEAS4Games: building expressive virtual characters for computer games". In: *8th International Conference on Intelligent Virtual Agents (IVA)*. Tokyo, Japan, pp. 426–440. DOI: 10.1007/978-3-540-85483-8_43.

Hess, J. *git-annex*. URL: http://git-annex.branchable.com/.

Rosenberg, A. (2012). "Rethinking the corpus: moving towards dynamic linguistic resources". In: *Interspeech*. Portland, OR, USA, pp. 1392–1395. URL: http://www.isca-speech.org/archive/interspeech_2012/i12_1392.html.

Scheffler, T., R. Roller, F. Kretzschmar, S. Moeller, and N. Reithinger (2012). "Natural vs. synthesized speech in spoken dialog systems research ?? Comparing the performance of recognition results". In: *10th ITG Conference on Speech Communication*. Braunschweig, Germany, pp. 26–28. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6309595.

Scherer, K. (2003). "Vocal communication of emotion: a review of research paradigms". In: *Speech Communication* 40.1-2, pp. 227–256. DOI: 10.1016/S0167-6393(02)00084-5.

Schröder, M. (2009). "Expressive speech synthesis: past, present, and possible futures". In: *Affective Information Processing*. Ed. by J. Tao and T. Tan. Springer. Chap. 7, pp. 111–126. DOI: 10.1007/978-1-84800-306-4_7.

Schuller, B., A. Batliner, S. Steidl, and D. Seppi (2011). "Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge". In: *Speech Communication* 53.9-10. DOI: 10.1016/j.specom.2011.01.011.

Steiner, I., M. Schröder, M. Charfuelan, and A. Klepp (2010). "Symbolic vs. acoustics-based style control for expressive unit selection". In: *7th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW)*. Kyoto, Japan, pp. 114–119. URL: http://www.isca-speech.org/archive/ssw7/ssw7_114.html.

Székely, É., I. Steiner, Z. Ahmed, and J. Carson-Berndsen (2013). "Facial expression-based affective speech translation". In: *Journal on Multimodal User Interfaces*, in press. DOI: 10.1007/s12193-013-0128-x.

Torvalds, L. *Git*. URL: http://git-scm.com/.

Türk, O. and M. Schröder (2010). "Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.5, pp. 965–973. DOI: 10.1109/TASL.2010.2041113.