Evaluating Emotional Speech Synthesis

Emelie Van De Vreken[1]

[1]University of Edinburgh, United Kingdom

emelie.vandevreken@ed.ac.uk

Abstract

Emotive Text To Speech technology is a very active field of research. Many machine learning approaches are put forward to increase the "expressiveness" of the output (e.g. [1, 2]). Neither the evaluation method nor the (labelled) data are investigated much.

By focussing on these neglected stages of the development process, we can be more confident that recognisable emotion is present in the synthesised speech. Taking a wider perspective also means putting central the end-user, and the intended use case for the technology.

On the practical side of things, there are three areas of focus. First, a practical engineering approach to create synthesised utterances with any emotion and style, called voice puppetry. Second, the type of prompts used in listening tests, and how much content and context is available to the listener. Finally, a look at how much emotion labels vary depending on the amount of context and content available to the annotator.

Index Terms: speech synthesis, expressive speech synthesis, speech synthesis evaluation, human-computer interaction, voice puppetry

1. Introduction

In the field of Text To Speech (TTS) research, there is an increasing focus on models producing more varying prosody to convey a wider array of affective states. Humans vary suprasegmental or prosodic features in their speech, to convey more information than what is present in the words they are producing. For example, it communicates their intention or emotion. As it stands, emotive speech synthesis can not yet produce utterances with any exact affect intended by the user.

The need for emotive TTS is unlikely to decrease, as it is essential for applications such as audiobook production or care robots. The more humans interact with computer and computer speech, the more scenarios there are in which being expressive is required for a successful interaction.

As more machine learning approaches attempt the emotive TTS task ([1, 2, 3, 4] to name a few), how can we be more confident that recognisable emotion is present in synthesised speech? My research aims to highlight the importance of critical user testing, realistic dataset labelling, and keeping the end-user in mind throughout the development process.

2. Contributions

Machine learning approaches generally aim at "expressiveness" without defining which context the TTS system is intended to be used in, or what the intended result is (for example [1] and [2]). In contrast to this, my research focuses on practical, real-world solutions for real-world problems.

On one end of the development process, this translates into dissecting the impact of data and data labelling choices. For example, the amount or type of labels to use, and the amount of context available when assigning labels.

Additionally, the research demonstrates the impact of choosing relevant and critical evaluation methods. This includes the listening test setup, as well as prompt selection and generation.

3. Challenges

Working in the domain of expressiveness and emotion comes with many challenges.

- First and foremost, there is no one accepted definition of emotion. This is in part due to our personal experiences shaping our perception of emotion [5]. Emotions can also be nuanced and complex.
- There also tends to be excessive confidence in the performance of humans. It is true that humans can, for the majority, agree on facial emotions when given some categories [6]. This does not mean that we can unequivocally pinpoint the emotion in speech while disregarding the utterance content [7].
- Finding or collecting useful data is hard (see Section 4).
- Finally, enabling TTS with a wider range of expressions also carries risk. The more aspects of human speech can be mimicked, the more convincing deepfakes could be generated. I.e., maliciously synthesised utterances misrepresenting a person.

4. Analysis of existing datasets in English

A large number of TTS models train on the LJ dataset [8]. For learning more varied prosody, this is not a good starting point, as LJ is known to be quite flat and neutral. A number of expressive TTS models use high-quality expressive datasets that are either proprietary or internal-use-only, e.g. [3, 4]. Two of the largest freely available English datasets labelled for emotion are CMU-MOSEI [9] and MELD [10]. Each comes with its own (dis)advantages.

CMU-MOSEI is based on single-person YouTube videos, and allows for multiple simultaneous emotions, with each their own intensities. The dataset consists of 3,840 videos, but the most common intensity rating is 0.33 on a scale of 0-3. Higher intensities occur exponentially less.

MELD is based on scenes from the Friends TV show. There are 13,708 utterances, each given one emotion label within the context of the whole dialogue. While the audio is high-quality, it is often incorrectly cut into utterances, actors talk over each other, there is audience laughter, and other noise.

In both datasets, joy or happiness are by far the most common emotion, but most utterances are considered neutral. In general, there is a trade-off between natural and high-quality data. Natural data is often noisy, with more subtle or complex emotion. High-quality data is often acted in a sound booth or other artificial setting. It contains more unambiguous, archetypical emotion, but may be less natural.

With the data potentially limiting the success of machine learning models, another way to create emotive TTS utterances is through voice puppetry.

5. Prompts with voice puppetry

Voice puppetry is sometimes also called "voice re-enactment" [11]. It refers to the use of a reference recording at inference time, in order to influence the output of a TTS model [12, 13].

This has been implemented for the FastPitch model [14]. In FastPitch, there are estimators for pitch, energy, and duration. The outputs of these are added to the output of the encoder, before being passed to the decoder to create a mel-spectrogram. At inference time, the output of these estimators is replaced by values extracted and calculated from a reference sound file. In this file, someone is uttering the same phrase as is passed to the model, but in the requested style.

While fine-tuning is ongoing, there is still a degradation of the signal after puppetry.

The benefits of this tool are that it is an engineering approach to add in expressiveness, change emphasis, etc. You can adjust the style of the output in any way you want, without having to put it into words or feature values. If you can act it, you can get it. You have maximum control, in a user-friendly, intuitive method. There is no need to retrain a model, or train a separate model.

Voice puppetry is most useful in cases where a small number of prompts is required that still sound like a TTS voice, but in a particular style. This can either be to update a few utterances in a larger set, e.g. for an audiobook, or when only a small set of prompts is needed, e.g. TTS experiments.

6. How will this come together?

Between suboptimal dataset labels and puppetry inference, there are still some missing pieces to complete the puzzle.

6.1. Evaluation methods

First of all, the voice puppetry method needs to be evaluated. To do this, there will be an experiment using simple Subject-Verb-Object sentences with varying emphasis. The text input to a trained FastPitch model [14] will remain the same, while the puppeteer alternates emphasis on the subject, verb, and object. If this approach is successful, human performance on hearing which part of the sentence is emphasised, will be on par with that of the original puppeteer utterances. Additionally, the experiment includes a test to ensure the identity of the training speaker is recognisable, and a test to inspect the effects of different puppeteer speakers.

Secondly, the puppetry method will be used to create prompts for different listening test setups. The main metric will be emotion recognition, which additional tests for naturalness and appropriateness. The setup will vary in context given to the prompt in question, and will also vary in the content of the prompts, or how useful that content is for the recognition of the intended emotion.

6.1.1. Context

There are varying amounts of context that can inform a listener. Recognising the emotion in an utterance may be easier if we know what sentences or interactions came before it. Seeing facial expressions alongside hearing the utterance can again provide additional evidence. Familiarity with the speaker can also help: is the speaker, and their emotion style, known to the listener?

On the contrary, in the setting of a listening test, this may mean that the context has provided strong evidence for an emotion, while the acoustic signal itself actually carries very little evidence (e.g. emotion recognition from faces is context-sensitive [15]). Visual and text features tend to be more informative than acoustic features in the field of emotion recognition, as seen in e.g. [16, 17].

6.1.2. Content

Generally speaking, listening tests for TTS are performed on listeners who are fluent (if not native) in the language synthesised. A side effect of this is that they have easy access to the content of the utterance: they hear both what is being said, and how it is being said. In order to avoid any emotion guesses based on what is being said, there are multiple methods to make the content as little comprehensible as possible.

1. Use pseudo-language, with same phonotactics as language X, but with made-up (content) words (as in [18]): *[pseudo-English] Your lunce twells aleping.*
2. Use text from a foreign language: *[Dutch] Ik haat het hier.*
3. Use text with contrasting sentiment: *[happy] my dog died last week.*

On top of that, there are also more traditional methods. For example: using neutral, factual sentences, or semantically unpredictable sentences [19].

6.2. Dataset labels

Often, datasets with emotion labels are multi-modal. They are a combination of video (people talking), the corresponding audio, and text (the transcription). Each segment is given a label out of a small set of options, and this label is the same for video, audio, and text alike. For emotion recognition or synthesis based on the auditory dimension only, these labels may be misleading. Parts of a dataset, most likely MELD [10], will be relabelled using audio only. As an alternative to using a predetermined set of labels, there will also be a trial of clustering segments by asking annotators if two utterances have the same or different emotions. Each cluster can then be assigned a label afterwards.

Other than comparing the distributions of labels, there will also be a listening test to measure the impact of the label strategies on the model output. The intention is to train the same FastPitch model [14] twice on the same dataset: once with existing labels, once with newly acquired labels. The listening test will once again revolve around an emotion recognition metric, and incorporate learnings from the first listening test. Depending on time and model performance, these models can then be used for voice puppetry, to highlight differences in puppetry success depending on the trained model's inclusion of emotional data.

7. Acknowledgements

Many thanks to my supervisors Korin Richmond and Catherine Lai, for their ongoing support. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

8. References

[1] M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim, and N. S. Kim, "Expressive text-to-speech using style tag." [Online]. Available: https://arxiv.org/abs/2104.00436

[2] K. Lee, K. Park, and D. Kim, "Styler: Style factor modeling with rapidity and robustness via speech decomposition for expressive and controllable neural text to speech," 2021. [Online]. Available: https://arxiv.org/abs/2103.09474

[3] V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote, "Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6179–6183.

[4] X. Li, C. Song, J. Li, Z. Wu, J. Jia, and H. Meng, "Towards multi-scale style control for expressive speech synthesis," *arXiv preprint arXiv:2104.03521*, 2021.

[5] E. M. Hunter, L. H. Phillips, and S. E. MacPherson, "Effects of age on cross-modal emotion perception." *Psychology and Aging*, vol. 25, no. 4, p. 779, 2010.

[6] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[7] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures," *Journal of Cross-Cultural Psychology*, vol. 32, no. 1, pp. 76–92, Jan. 2001, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/0022022101032001009

[8] K. Ito and L. Johnson, "The lj speech dataset," 2017.

[9] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 2236–2246. [Online]. Available: http://aclweb.org/anthology/P18-1208

[10] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," *arXiv:1810.02508 [cs]*, Jun. 2019, arXiv: 1810.02508. [Online]. Available: http://arxiv.org/abs/1810.02508

[11] F. Bous, L. Benaroya, N. Obin, and A. Roebel, "Voice reenactment with f0 and timing constraints and adversarial learning of conversions," in *30th European Signal Processing Conference (EUSIPCO 2022)*, 2022.

[12] M. P. Aylett, D. A. Braude, C. J. Pidcock, and B. Potard, "Voice puppetry: Exploring dramatic performance to develop speech synthesis," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 117–120.

[13] M. P. Aylett and Y. Vazquez-Alvarez, "Voice puppetry: Speech synthesis adventures in human centred ai," in *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, ser. IUI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 108–109. [Online]. Available: https://doi.org/10.1145/3379336.3381478

[14] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[15] H. Aviezer, R. R. Hassin, J. Ryan, C. Grady, J. Susskind, A. Anderson, M. Moscovitch, and S. Bentin, "Angry, disgusted, or afraid?: Studies on the malleability of emotion perception," *Psychological Science*, vol. 19, no. 7, pp. 724–732, 2008, pMID: 18727789. [Online]. Available: https://doi.org/10.1111/j.1467-9280.2008.02148.x

[16] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing Emotions in Video Using Multimodal DNN Feature Fusion," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 11–19. [Online]. Available: https://www.aclweb.org/anthology/W18-3302

[17] V. Rajan, A. Brutti, and A. Cavallaro, "Is cross-attention preferable to self-attention for multi-modal emotion recognition?" in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4693–4697.

[18] S. Yilmazyildiz, D. Henderickx, B. Vanderborght, W. Verhelst, E. Soetens, and D. Lefeber, "EMOGIB: Emotional Gibberish Speech Database for Affective Human-Robot Interaction," in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, vol. 6975. [Online]. Available: http://link.springer.com/10.1007/978-3-642-24571-8_17

[19] C. Benoît, M. Grice, and V. Hazan, "The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech communication*, vol. 18, no. 4, pp. 381–392, 1996.