Edinburgh Research Explorer

Combining conversational speech with read speech to improve prosody in Text-to-Speech synthesis

Combining conversational speech with read speech to improve prosody in Text-to-Speech synthesis

Johannah O'Mahony, Catherine Lai, Simon King

The Centre for Speech Technology Research, University of Edinburgh, UK

`johannah.o'mahony@ed.ac.uk`

Abstract

For isolated utterances, speech synthesis quality has improved immensely thanks to the use of sequence-to-sequence models. However, these models are generally trained on read speech and fail to generalise to unseen speaking styles. Recently, more research is focused on the synthesis of expressive and conversational speech. Conversational speech contains many prosodic phenomena that are not present in read speech. We would like to learn these prosodic patterns from data, but unfortunately, many large conversational corpora are unsuitable for speech synthesis due to low audio quality. We investigate whether a data mixing strategy can improve conversational prosody for a target voice based on monologue data from audiobooks by adding real conversational data from podcasts. We filter the podcast data to create a set of 26k question and answer pairs. We evaluate two FastPitch models: one trained on 20 hours of monologue speech from a single speaker, and another trained on 5 hours of monologue speech from that speaker plus 15 hours of questions and answers spoken by nearly 15k speakers. Results from three listening tests show that the second model generates more preferred question prosody.

Index Terms: conversational speech synthesis, speech synthesis, expressive speech synthesis

1. Introduction

End-to-end speech synthesis models have led to significant quality improvements for isolated read speech utterances. But these models don't generalise well to unseen material [1] such as when training on read speech and synthesising conversational speech [2]. Spontaneous conversational speech exhibits prosodic characteristics that distinguish it from read speech [3], e.g. syllabic reduction, decreased prosodic range [4] and differing stress placement [5]. Further, conversational turns are highly context-dependent, exhibiting changes in emotion and intent, topic/focus structure, and subtle prosodic changes due to phenomena such as (dis)entrainment. These are features that are usually absent in read data. Due to this, synthetic voices used in dialogue systems trained on read speech do not sound conversational and struggle with certain prosodic contours found in conversation, e.g. different question types. To generate more natural conversational speech we need to look to new sources of data containing these phenomena [2].

In order to generate more conversationally-appropriate speech, different methods have been proposed. For example, by recording conversational data from chatbot scripts [6]. Though this approach should lead to an increase in prosodic coverage and conversational style, the speech is not truly spontaneous because it is still read speech spoken by a single speaker [2]. Transcriptions of spontaneous data read aloud differ significantly from actual spontaneous data [7] in stress placement, number of pauses, etc [5]. Further, [6] recorded isolated conversational utterances, thus losing prosodic phenomena which arise from interaction with another speaker, and other important contextual information coming from prior turns. Because of this, these data are not suitable for training context-aware models [7]. To overcome this [7] used a more natural yet still controlled approach, recording 45 conversations between two female speakers with semi-scripted interactive scenarios. The speakers were allowed to deviate from the script, permitting phenomena such as false starts. The conversations were then manually transcribed and used to train a context-aware speech synthesis model.

The approaches mentioned above involve recording new data; this is time-consuming and expensive. In contrast, [2] looked to podcasts as a source of truly spontaneous speech. They compared synthetic speech generated from models trained on read speech, lab-recorded spontaneous speech, and found spontaneous speech (from one speaker in a two-person podcast series) of which the last was judged to be more acceptable for casual conversations and spontaneous monologues.

The work in [2] shows the potential of podcasts as a source of truly spontaneous speech. However, in common with all the other approaches described so far, it still involved using a new speaker. This means that the resulting synthetic voices have different speaker identities; real use cases may demand a single speaker identity for both read and conversational speaking styles. So, we follow [2] by using podcast data but, in contrast, we use this naturally-produced spontaneous speech to *enrich* the prosodic repertoire of a target speaker based on read speech.

Spontaneous speech is inherently heterogeneous due to the wide range of contexts and functions in which it is used [3]. This is major obstacle to doing research, which we mitigate by narrowing our focus to the question-answer adjacency pair [8]. We chose this particular pair due to its ubiquity in speech synthesis applications such as dialogue systems. Further, because our ultimate research goal beyond this paper is to synthesise context-dependent speech, we chose question-answers due to their use in eliciting different prosodic renditions of the same text, e.g. [9]. Finally, though it is only a subset of conversation, the prosodic realisation of both questions and answers is highly dependent on the interactional context in which they are found, so they are a good test case before tackling the wider properties of conversation in general.

The goal of the current work is to:

- create a corpus of question-answer pairs from found two-party spontaneous podcast data.

- create a multi-speaker model using read monologue data combined with the above data.

- evaluate whether adding spontaneous data improves prosody of questions and answers.

2. Data

2.1. Spontaneous Speech Data

The Spotify 100,000 Podcast dataset [10] contains 2 TB of data from a selection of 100k podcasts which have been automatically transcribed, punctuated and speaker diarised using Google Cloud Services. The recordings contain some overlapping speech, background music, and laughter and the automatic transcriptions contain errors in word recognition and diarisation. Filled pauses and hesitations such as *uhm* are not transcribed. Since we do not have the resources to manually correct or even to quantify the above errors, we applied several stages of filtering to discard suspect data.

2.1.1. Data Filtering

We split the data into subsets according to the number of speakers detected by speaker diarisation and retained only podcasts containing exactly two speakers, to obtain ~74k podcasts. Based on punctuation and diarisation we split the transcripts into utterances. Errors in diarisation or punctuation led to some such utterances being attributed to two speakers. We retained only utterances attributed to a single speaker and whose transcript was a complete sentence (according to the automatic punctuation).

To extract question and answer pairs, we located utterances which ended in a question mark, though this might exclude questions without typical question syntax, e.g. declarative questions. To extract the answer we simply took the following turn if attributed to the other speaker. The speech corresponding to the extracted question-answer pairs forms our corpus. This resulted in 123,943 question-answer pairs. We removed question-answer pairs containing symbols or numbers to avoid text normalisation issues, as well as recordings under 500 ms or over 15 s in duration. The resulting set at this stage contains 92,478 question-answer pairs.

2.1.2. Audio Filtering

To ensure that each question and its answer were actually spoken by different speakers (recall that the provided diarisation is imperfect), we extracted speaker embeddings using Speech Brain [11] ECAPA-TDNN [12] for both and performed speaker verification. We removed pairs for which the model deemed the speakers were the same. The audio data is single channel and therefore does not offer the possibility to separate speakers by channel. So we used Pyannote audio [13] to detect overlapping speech and removed pairs in which any overlap was found. We also used a laughter detector [14] and removed any pairs in which laughter was found. The final set comprises 26,876 question-answer pairs amounting to ~18 hours of questions and ~20 hours of answers.

2.2. Read Speech Data

Our target speaker dataset is LJ Speech which consists of 13,100 utterances from audiobooks read by a female speaker of American English.

3. Method

Our method involves training speech synthesis models on a combination of spontaneous and read speech.

3.1. Data Selection

For the current work, we randomly selected an even number of hours of questions and answers from the question-answer dataset (henceforth simply 'spontaneous speech') described in the previous section. All selected utterances had a duration for 1 s to 10 s and a podcast country label of UK, US, Canada or general English. For a baseline read speech model we randomly selected 20 hours of data from LJ Speech in which maximum utterance duration is 10 s (henceforth 'read speech').

In early experiments, we compared models trained on data comprising 0%, 25%, 50%, or 75% spontaneous speech with 100%, 75%, 50%, or 25% read speech respectively. Informal listening showed that the model trained with 75% spontaneous speech + 25% read speech did not suffer significantly in quality compared to the using 100% read speech, and that larger prosodic improvements were observed than with the 25%+75% or 50%+50% models. Table 1 summarises the data used in subsequent experiments.

Table 1: *Approximate training data for each model*

| model | read speech | spontaneous speech | | total |
		questions	answers	
baseline	20 hours			20 hours
datamix	5 hours	7.5 hours	7.5 hours	20 hours

3.2. Model

We used FastPitch 1.1 [15] which is a multi-speaker non-autoregressive model with a transformer encoder-decoder architecture. It employs three variance adapters which predict values for F_0, intensity and duration. We trained two models. The first model is the **baseline** trained only on read speech (from LJ Speech). For consistency, **baseline** was trained using a speaker embedding table of the same size as in our second model, with only one entry being used. Our second model, **datamix**, combines read speech (from LJ Speech) and spontaneous speech (selected from Spotify podcasts using the procedure in Section 2) in the ratio specified in Table 1. The speaker embedding table has 14849 entries: 1 for the single LJ Speech speaker and the remainder for the speakers in the spontaneous speech data; all speaker codes are used during training. We trained each model for 1k epochs with a batch size of 20 on 3 GPUs.

3.3. Evaluation

We hypothesised that **datamix** would generate more natural- and conversational-sounding speech. This was tested in two preference tests using identical stimuli and differing only in the instructions to listeners. In the first test, listeners were presented with 100 pairs of stimuli (each pair comprising the output of **datamix** and **baseline** for the same text) and asked *Which of the following sounds the most conversational?* Stimulus order was randomised within and across pairs, differently for each listener. In a post-test questionnaire we asked them what they understood by the term 'conversational'. The second listening test was identical, except that it asked a new set of participants *Which of the following do you prefer?*

To gauge the overall quality of both models we also performed a Mean Opinion Score (MOS) test which presented another new set of listeners with 50 synthesised questions and

50 synthesised answers (counterbalanced across two listener groups so that no participant heard the same text spoken by both systems). Participants were asked to rate each stimulus on a scale labelled *1–bad*, *2–poor*, *3–fair*, *4–good* and *5–excellent*.

3.3.1. Stimuli

We started from 100 questions and 100 answers randomly selected from the question-answer dataset and not used for model training. Based only on the natural speech and its automatic transcription, we manually removed utterances containing fewer than 2 words, more than 15 words, profanity, controversial topics, gross grammatical errors, nonsensical content, false starts, or acronyms (to avoid text normalisation errors). From the remaining utterances, we randomly selected 50 questions and 50 answers for use in all listening tests.

3.3.2. Listeners

We recruited ∼30 listeners for each of the 3 listening tests through Prolific[1] who were US residents, native English speakers with no reported hearing impairments, and balanced for sex. Listeners were removed if they did not complete the test, did not use headphones or had issues playing the audio samples. No listener was permitted to participate more than once.

3.3.3. Statistical Analysis

We used mixed-effects regression models to account for lack of independence in the data due to repeated measures for listeners and stimuli. These sources of variance have been found to be quite significant in speech synthesis evaluation studies and pose problems in evaluating TTS output [16]. For the analysis of preference test results we use a binomial mixed-effects model using the logit-link function with random intercepts for listener and stimulus to account for variance between listeners and stimuli. We used the lme4 package [17]. In this model, we used no predictors and are therefore testing whether the intercept coefficient is different from the null hypothesis, which is that both models have an equal probability of being chosen. This form of testing is roughly equivalent to the exact binomial test, but now we are able to account for random variation due to listeners and stimuli. For the MOS analysis, we used a cumulative link mixed-model (CLMM) using the ordinal package [18].[2] These models have been shown to be more suitable for analysis of ratings, as they account for the ordinal nature of the response, and have already been used in Natural Language Generation evaluation [19]. Again, the inclusion of random intercepts allows us to account for variance caused by listeners and stimuli, e.g. listeners using different levels of the ordinal scale. Equations for each model are given in the next section.

4. Results

4.1. Preference Tests

31 listeners completed the first preference test, which asked *Which of the following sounds more conversational?*. We removed 1 listener for not using headphones, 1 for having issues with a number of audio files and 1 for completing the test in less time than the total duration of the audio. This left 28 listeners. For both preference tests we used a binomial mixed-effects

model with the logit function using the following formula which tests whether the distribution of model preferences differs from chance:

```
choice ∼ 1 + (1|listener) + (1|stimulus)
```

Results are summarised in figure 1. We found a significant intercept for questions (β=0.47 (0.61 prob), CI=(0.56,0.67), p < 0.01) which means that **datamix** was chosen significantly more times than **baseline**. For answers, we found no significant difference between **datamix** and **baseline** (β=-0.17 (0.46 prob), CI=(0.41,0.51), p=0.09).
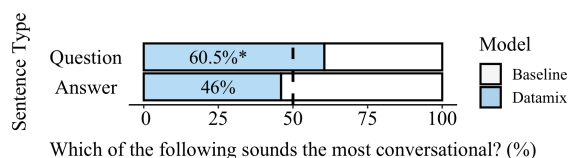

Figure 1: *Results for 'Which sounds the most conversational?'*

The second preference simply asked *Which of the following do you prefer?* and 32 listeners took part, of which 1 was removed for not using headphones, and 1 for having issues playing some audio files. The results are summarised in figure 2. Again **datamix** was chosen significantly more times over **baseline** for questions (β=0.44 (0.61 prob), CI=(0.54,0.67), p < 0.01), but not for answers (β=-0.21 (prob=0.45), CI=(0.39,0.50), p=0.06).
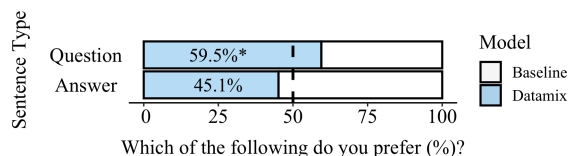

Figure 2: *Results for 'Which of the following do you prefer?'*

4.2. MOS Test

A total of 62 listeners (30 and 32 per listener group) completed this test, of which 2 were removed for failing to use headphones. The mean MOS for **baseline** when synthesising answers was M=3.19 SD=1.21 and for questions M=2.83 SD=1.20. For **datamix**, the mean MOS for answers was M=3.07 SD=1.24 and for questions M=3.12 SD=1.22. The MOS results are summarised in figures 3 and 4.
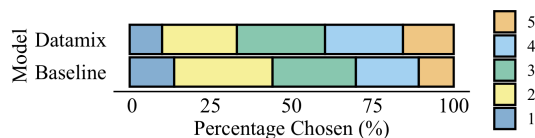

Figure 3: *MOS results for questions.*

To test whether the models were rated significantly differently, we fitted an ordinal mixed-effects model predicting the effect of each `model` and `sentence type` on the log odds of receiving a particular MOS. We specified a random effects structure to account for repeated measures of both stimulus and listener which accounts for random variation of both,
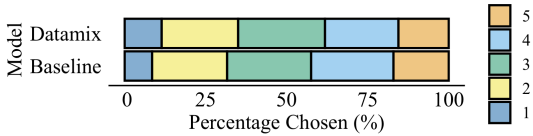
Figure 4: *MOS results for answers*

i.e. some listeners will use the scale differently and some stimuli will show random variation, and a random slope for listeners to account for baseline preference of one model over another. We used the following formula:

```
MOS score ~ model * sentence type +
        (model|listener) + (1|stimulus)
```

To test the significance of the fixed effect, we compared models using a log-likelihood test between models with and without each factor of interest. We found that there was no main effect of `model` (β=-0.25, SE=0.25, $G^2(1)$=0.82, p=0.37) or of `sentence type` (β=-0.74, SE=0.25, $G^2(1)$=3.15, p=0.08). This means that, taken *independently*, there is no significant difference between ratings of questions and answers, or between the models. We did however find a significant *interaction* between `model` and `sentence type` using log-likelihood ratio test between the full CLMM and the CLMM without an interaction factor (β=0.83, SE=0.35, $G^2(1)$=5.47, p=0.02). To examine this interaction we calculated the predicted probability of each MOS rating per model and sentence type (see 5). As we can see, questions in **baseline** have a higher probability than answers of being scored as a 1 or a 2, and consequently a lower probability of getting a higher rating. As in the preference tests, this shows us that **datamix** performs better for questions, but performs roughly equally well for answers.
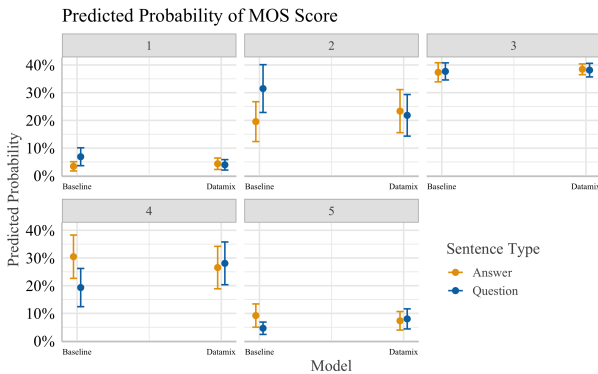

Figure 5: *CLMM predicted probabilities of MOS scores for models and sentence types.*

5. Discussion

The preference test results show that enriching training data with spontaneous speech (in this case from multiple speakers) leads to an increase in listener preference and ratings. The difference between **datamix** and **baseline** is significant when synthesising questions. There may be a trend towards a lower preference for **datamix** answers, but this is not significant. The MOS test paints a similar picture: questions were rated significantly higher for **datamix**; for answers, both models performed similarly. Comparing the results of the *Which sounds*

the most conversational? preference test with those of *Which do you prefer?*, we conclude that listeners did not find our models significantly more *conversational* as the results of both preference tests are quite similar. This is probably due to the multi-speaker set-up which has learned specific prosodic features for each speaker and thus still uses more prosodic information from the read data than the spontaneous data. A possible improvement to our approach might be to remove speaker conditioning from the variance adapters of FastPitch thus allowing them to be speaker independent. This prosodic information could then be shared across speakers, similar to [1] who pre-trained the duration variance adapter using ASR data.

Next to adapting how the spontaneous data is incorporated during model training, our future work will also focus on data development. Improvement can likely be achieved using more data filtering and selection techniques e.g. matching speaker gender to our target speaker and by adding more information about emotion or pragmatic intent of the questions and answers to condition the variance adapters during training. To achieve this our future work aims to further filter the corpus of question-answer pairs using clustering of audio and textual features. Choosing speakers with more similar characteristics to our model speaker may also improve the quality of the model.

As mentioned, our ultimate goal is to synthesise more appropriate questions and answers based on prior context. When dealing with spontaneous data we are likely to have richer and more contextually-dependent prosodic realisations in both questions and answers. There is likely significant variation in question prosody depending on dialogue function, for example whether a question is asking for clarification or opening a new topic [20]. In this study we used questions taken out of their context and trained a context-unaware model, which might revert to generating an average prosodic representation of questions, losing the variation in the data. It is therefore likely that using context will lead to an increase in quality of question intonation as we include information to account for different prosodic realisations. This work is therefore a stepping stone to future work tackling these issues.

6. Conclusions

We have shown that enhancing training data with speech from real spontaneous conversations leads to improvements in the prosody of synthetic speech for a target speaker for whom we only have read speech. The introduction of speech from several thousand speakers did not lead to a reduction in quality for the target speaker, and did improve listener ratings of question prosody. We acknowledge that there other methods we could use to inject this prosodic knowledge into the model, for example pre-training the FastPitch variance adapters. Though our ultimate goal is to use this data for context-aware question and answer generation, here we have already shown that a simple use of this data leads to improvements in prosody. We will conduct further analysis on the output of **datamix** to investigate where it fails for answer prosody, and why questions are rated more highly. Future work will also focus on selecting cleaner speech samples from our question-answer dataset, and use context to predict the prosodic realisation of questions and answers.

7. Acknowledgements

8. References

[1] Z. Li, Y. Zhang, M. Nie, M. Yan, M. He, R. Zhang, and C. Gong, "Improving Prosody for Unseen Texts in Speech Synthesis by Utilizing Linguistic Information and Noisy Data," in *ArXiv*, 11 2021. [Online]. Available: http://arxiv.org/abs/2111.07549

[2] E. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 4435–4439, 2019.

[3] M. E. Beckman, "A Typology of Spontaneous Speech," in *Computing Prosody*, Y. Sagisaka, N. Campbell, and N. Higuchi, Eds. New York: Springer, 1997, pp. 7–26.

[4] V. Hazan and R. Baker, "Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style?" in *Proceedings of DiSS-LPSS Joint Workshop*, Tokyo, Japan, 9 2010, pp. 7–10.

[5] P. Howell and K. Kadi-Hanifi, "Comparison of prosodic properties between read spontaneous speech material," *Speech Communication*, vol. 10, pp. 163–169, 1991.

[6] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2106.08468

[7] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational End-to-End TTS for Voice Agents," *2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings*, vol. 2, pp. 403–409, 2021.

[8] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974. [Online]. Available: http://www.jstor.org/stable/412243

[9] S. Latif, I. Kim, and L. Besacier, "Controlling Prosody in End-to-End TTS : A Case Study on Contrastive Focus Generation," in *25th Conference on Computational Natural Language Learning (CoNLL)*, 2021, pp. 544–551.

[10] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette, and R. Jones, "100,000 Podcasts: A Spoken English Document Corpus," in *Proceedings ofthe 28th International Conference on Computational Linguistics*, Barcelona, 2021, pp. 5903–5917.

[11] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A General-Purpose Speech Toolkit," 2021.

[12] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[13] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, 5 2020.

[14] J. Gillick, W. Deng, K. Ryokai, and D. Bamman, "Robust Laughter Detection in Noisy Environments," in *Interspeech*, 2021, pp. 2481–2485.

[15] A. Lancucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction," 2021.

[16] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-August. International Speech Communication Association, 2017, pp. 3976–3980.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[18] R. H. B. Christensen, "ordinal—Regression Models for Ordinal Data ," 2019.

[19] D. M. Howcroft and V. Rieser, "What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 11 2021, pp. 8932–8939. [Online]. Available: https://aclanthology.org/2021.emnlp-main.703

[20] S. Strömbergsson, J. Edlund, and D. House, "Prosodic measurements and question types in the Spontal corpus of Swedish dialogues," *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, vol. 1, pp. 838–841, 2012.