Roles of the Average Voice in Speaker-adaptive HMM-based Speech Synthesis

Junichi Yamagishi[1], *Oliver Watts*[1], *Simon King*[1], *Bela Usabaev*[2]

[1]The Centre for Speech Technology Research,
University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom
[2]Universität Tübingen, Wilhelmstr. 7 72074 Tübingen, Germany
jyamagis@inf.ed.ac.uk

Abstract

In speaker-adaptive HMM-based speech synthesis, there are typically a few speakers for which the output synthetic speech sounds worse than that of other speakers, despite having the same amount of adaptation data from within the same corpus. This paper investigates these fluctuations in quality and concludes that as mel-cepstral distance from the average voice becomes larger, the MOS naturalness scores generally become worse. Although this negative correlation is not that strong, it suggests a way to improve the training and adaptation strategies. We also draw comparisons between our findings and the work of other researchers regarding "vocal attractiveness."

Index Terms: speech synthesis, HMM, average voice, speaker adaptation

1. Introduction

Until recently, developing a text-to-speech synthesis system for a particular target speaker required a large amount of speech data read from a carefully prepared script. However, with the advent of HMM-based speech synthesis [1], statistical acoustic models for spectral, excitation, and duration features can now be precisely adapted from an average voice model (derived from other speakers) or a background model (derived from one speaker) using only a very small amount of speech data from the target speaker.

Recent experiments with speaker-adaptive HMM-based speech synthesis have also demonstrated its robustness to non-ideal speech data that have been recorded under varying conditions and with varying microphones, that are not perfectly clean, and/or that lack phonetic balance [2]. In fact, we have demonstrated that we can create thousands of TTS voices from non-TTS corpora such as ASR corpora [3, 4]. This technique opens up new applications in various domains. For example, medical voice banking or voice reconstruction for patients who have, or are threatened, by throat cancer, or the creation of alternative communication aids for patients with conditions such as Parkinson's disease, whereby the patient's original voice characteristics can be preserved [5].

The many TTS voices we have built so fat are available via an interactive online TTS demonstration system with a geographical interface[1]. The voices in this demonstration were built using pre-defined training recipes for each corpus. Importantly, this demonstration provides an opportunity to compare the quality of synthetic speech for many different speakers at the same time.

Careful listening reveals that 1) the quality of synthetic speech varies according to which corpus is used to train the average voice model, or according to the amount of adaptation data used and 2) there are a few speakers whose synthetic speech sounds worse than that of other speakers, even though they have the same amount of adaptation data and from within the same corpus.

With regard to the first issue, our previous analysis has already shown that the minimum amount of adaptation data required for reproducing speaker similarity to a certain level varies by target speaker (and acoustic features) and ranges from three minutes to six minutes [6] and also that the naturalness of the synthetic speech generated from the adapted models is closely correlated with the amount of data used for training the average voice model [7]. We also know that gender-dependent average voice models provide better speaker adaptation performance than gender-independent average voice models for TTS [7]. This directly explains the relatively low quality of voices built from a small corpus (such as the RM corpus) since the small corpus has neither a sufficient total amount of data to train a good average voice model, or sufficient data per speaker to perform high-quality adaptation.

The second phenomenon – those few speakers for whom synthetic speech quality is much worse – is more interesting; it is analogous to the familiar situation in ASR, where WER varies widely across some speakers and is especially high for a small number of speakers [8]. In this paper we investigate this phenomenon from the point of view of TTS.

Initially we suspected the negative effects of recording condition mismatch, because we have found that acoustic differences due to inconsistent recording conditions can be greater than differences between speakers [3, 4]. During the analysis of the recording conditions and sites, however, we stumbled upon a correlation between the naturalness of synthetic speech and the distance between the adapted speaker model and the average voice model.

2. HMM-based Speech Synthesis Systems and Experimental Conditions

A speaker-adaptive HMM-based speech synthesis system comprises four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

In the speech analysis part, three kinds of parameters for the STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram [9]) mel-cepstral vocoder with mixed excitation (i.e., the mel-cepstrum, $\log F_0$ and a set of band-limited aperiodicity measures) are extracted as feature vectors for the HMMs. In the average voice training part, context-dependent multi-stream left-to-right tied-state multi-space distribution hidden semi-Markov models are

[1]http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/map.html

trained on multi-speaker databases in order to simultaneously model the acoustic features and duration. A set of model parameters (mean vectors and diagonal covariance matrices of Gaussian pdfs) for the speaker-independent MSD-HSMMs is estimated using the EM algorithm. All EM re-estimation processes utilize speaker-adaptive training based on constrained maximum likelihood linear regression [10].

In the speaker adaptation part, the speaker-independent MSD-HSMMs are transformed by using constrained structural maximum *a posteriori* linear regression [7]. In the speech generation part, acoustic feature parameters are generated from the adapted MSD-HSMMs using a parameter generation algorithm that considers both the global variance of the trajectory to be generated and trajectory likelihood [11]. Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add. This signal is used to excite a mel-logarithmic spectrum approximation filter corresponding to the STRAIGHT mel-cepstral coefficients, to generate the speech waveform.

Using the framework above, we built gender-dependent average voice models from short term, long term (excluding the speakers from very long term), development, and evaluation subsets of the WSJ0 corpus [12]. The number of training sentences was 10847 and 12151 sentences (21.1 hours and 24.6 hours of speech) respectively.

3. Visualization using multidimensional scaling

A useful way to investigate the distribution of these 120 voices is to visualize them in a low dimensional space derived from the properties of the speech. There are several conventional approaches for visualizing speakers or speaking styles based on acoustic models or acoustic features [13, 14]. A similar visualization can be straightforwardly achieved using multidimensional scaling (MDS) [15].

Although we already gave parts of this result in [3], the low dimensional space is very important for the analysis of the listening tests presented later in the current paper, so we reproduce the visualisation results here using more voices and a three-dimensional space.

Using all test sentences from the Blizzard Challenge 2008, we generated a set of speech samples from the gender-dependent average voice models and 120 HTS voices, each of which was based on 100 adaptation sentences. We then calculated the average mel-cepstral distance between the speech for all pairs of voices, placing the values in mel-cepstral distance tables. For simplicity, the unadapted duration models of the average voice model were used so that the number of frames of synthetic speech for each speaker was the same. Then we applied a classic multidimensional scaling technique to the mel-cepstral distance table and examined the resulting three-dimensional space, which is shown in Figure 1.

The axes of this space do not have any *pre-defined* meaning, but MDS attempts to preserve the pairwise distances between speakers given in the mel-cepstral distance table. In other words, similar speakers will be close to one another in this space. On examining the figure in detail, we noticed that all three-characters codes (corresponding to the names of speakers) distributed in the bottom part start with 0 and the codes for speakers distributed in top part start with 4. The first character of the names represents recording site for these speakers (0: MIT, 4:SRI, and 2:TI) [12].

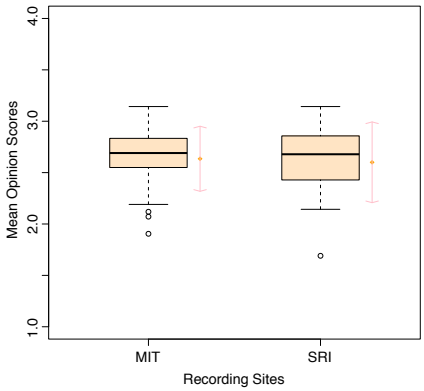

Figure 2: Standard box-plots are presented for evaluation scores of each site where the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. In addition mean scores and their standard deviation are shown using arrows next to the box-plots.

It is apparent that recording conditions were not consistent among the recording sites even though the same microphone was utilised. Acoustic differences due to the inconsistent recording conditions are greater than acoustic differences between speakers.

4. Subjective evaluations of 59 adapted voices and an average voice

A natural next step is to perform listening tests and to evaluate whether the acoustic differences due to the inconsistent recording conditions cause fluctuation of the quality of synthetic speech generated from speaker-adapted models based on the same average voice model and using the same amount of adaptation data.

We used the same adapted voices and the same average voice used for MDS in the previous section and evaluated their naturalness using a MOS test in which four test sentences were randomly chosen from all the test sentences used for MDS above. The number of listeners was 40.

The score distributions for each site are shown in Figure 2; we cannot see any clear differences between the results for each site. In fact, the Pearson product-moment correlation coefficient between the mean MOS scores obtained in the evaluation and the first axis of MDS (which corresponds to recording site) is just -0.13. In summary: the MOS naturalness scores are not correlated with recording site and the associated recording condition differences. Interestingly, the second axis of the MDS figure had somewhat stronger correlation with mean MOS (-0.38) than the first axis.

Therefore we decided we should examine other possible distances and focus on mel-cepstral distance between average voice and each voice, which can be viewed as a transformed distance of the voice. This correlation was stronger and it was -0.48. The fluctuation of the quality of synthetic speech was somewhat correlated inversely with mel-cepstral distance from the average voice. Its 95% confidence intervals are from -0.20 to -0.68.

Figure 3 illustrates the relationship between naturalness judgements in the listening test and the mel-cepstral distance between the adapted voices and the average voice. We can see that as the mel-cepstral distance from the average voice be-

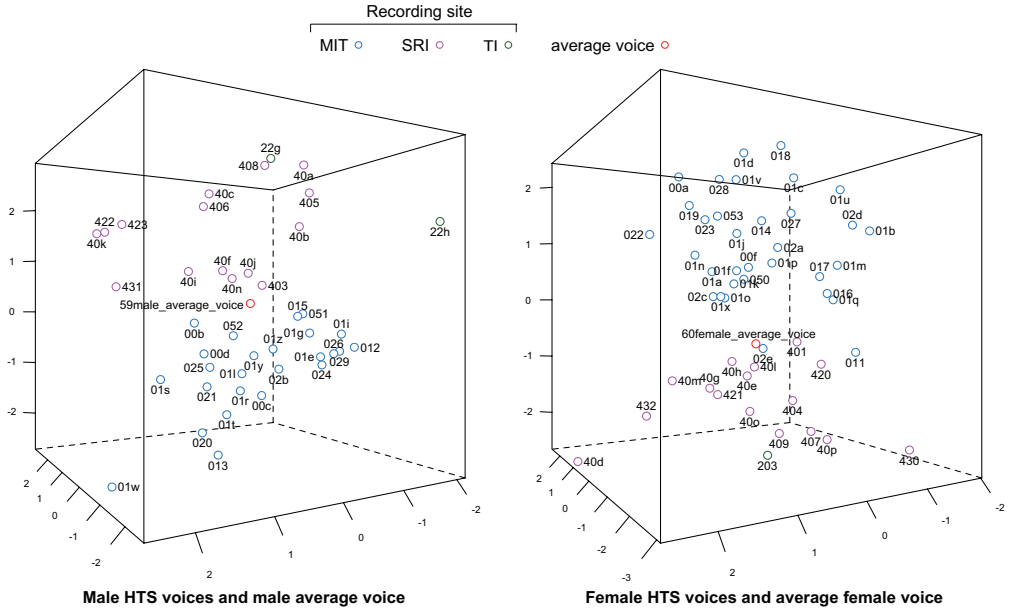

Figure 1: Multidimensional scaling of 120 HTS voices trained on the WSJ0 corpus. The three characters at each point correspond to the name of each speaker in the database. The left plot shows the the male adapted voices and male average voice and the right plot shows the female adapted voices and female average voice.

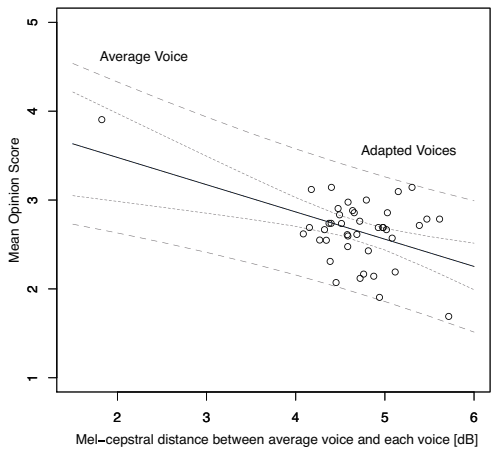

Figure 3: Scatter plot of mean MOS naturalness vs. mel-cepstral distance from the average voice for 59 male speaker-adapted voices and a male average voice model. Each point represents either a male voice or the male average voice. A linear regression and 95% confidence and prediction intervals are also shown. For computation of the mel-cepstral distance between the average voice and itself, a random-sampling-based parameter generation algorithm [16] was used.

comes larger, the MOS score generally becomes worse. A similar correlation between transform distance and quality reduction of output speech has been observed in voice conversion [17]. It may come as a surprise to see that the average voice is rated as the most natural by listeners (mean MOS score of 3.9.).

The correlation found is modest: it explains only 23% of the behaviour of the adapted voices (the cause of the remaining 77% of the variation is still unknown). However, it is still an important factor to take into consideration when training average voice models from many speakers. The finding is also consistent with the previous finding that gender-dependent average voice models provide better speaker adaptation performance than either gender-independent average voice models or speaker-dependent models for TTS. In addition, for achieving better quality synthetic speech, it also implies that we could use multiple gender-dependent average voice models and choose the nearest one as the basis for the adapted voice for particular target speaker (assuming sufficient data are available to construct multiple average voice models). Note that the amount of data for the average voice model is the dominant factor for the quality of the resulting synthetic speech.

5. Average voice sounds more attractive than individuals?

In addition to the transform distance mentioned in previous section, we hypothesize that there is a psychological reason why listeners prefer adapted voices which are closer to the average voice.

In a well-known study, published in their paper "Attractive Faces are Only Average" [18], Langlois and Roggman showed that averaged faces are judged to be more attractive than individuals. In a similar fashion, a possible psychological explanation for the higher naturalness score of the average voices in our study is that *attractive voices are also average*. This is an intriguing possibility with further implications for the statistical parametric approach to speech synthesis, since the statistical averaging effect, which is an acknowledged weakness of current HMM-based speech synthesisers, might in fact have the potential to produce voices that sound more attractive than individuals.

A very recent psychoacoustic study by Belin and colleagues [19], involving many speakers' vowels and their averaged vowels, supports this hypothesis (for natural speech, not synthetic speech). They found that their listening test scores for attractiveness are correlated with distance to the average vowel, as shown in Figure 4.

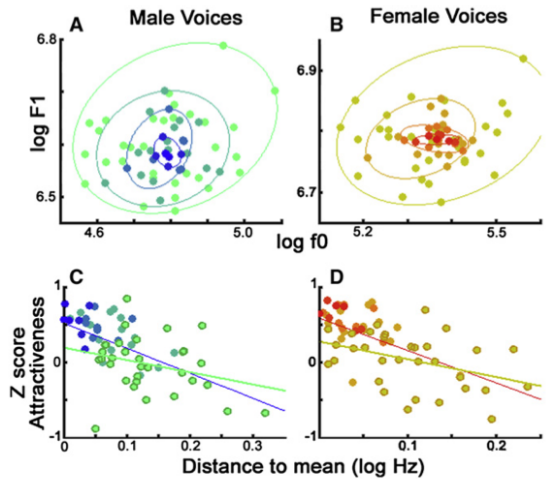

Figure 4: "In the logf0-logF1 space, Euclidean distance to mean was negatively correlated to vocal attractiveness rating (r=-0.59, adjusted R^2=0.34, p<0.001)." This figure is taken from [19].

There are some differences between their experiments and ours:

- They used only vowels, whereas we used complete sentences.
- We had only two average voices whereas they evaluated various combinations of speakers for constructing several average voices.
- They considered the Z score of attractiveness, rather than MOS of naturalness.
- Log F_0/F_1 space was used instead of mel-cepstral space.
- There is a larger gap between the average voice and adapted voices in our experiments. This may be explained by the recording condition inconsistency of our data. Our average voice models are located at the centre of recording conditions rather than the centre of the speakers due to the inconsistent recording conditions observed in Fig. 1.

The striking similarity between our study and that of Belin's group, leads us to consider if there is a possibility that our listeners were judging both vocal naturalness and attractiveness. Whilst we cannot answer this question yet, there is already no doubt that averaging across multiple speakers has a positive effect on the speech produced by the statistical parametric approach to speech synthesis.

6. Conclusions

In speaker-adaptive HMM-based speech synthesis, there are typically a few speakers whose synthetic speech sounds worse than that of other speakers who have the same amount of adaptation data and are from the same corpus. In this paper, we presented an investigation into this fluctuation in quality which found that, as mel-cepstral distance from the average voice becomes larger, the MOS naturalness score generally becomes worse. Although the negative correlation found is not that strong, we believe it gives a sufficient basis for developing improved training and adaptation strategies for average voice models. Furthermore, we have drawn comparisons with work on "vocal attractiveness" and have identified an area worthy of further investigation: the attractiveness of average voice-based synthetic speech.

Acknowledgements The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project http://www.emime.org).

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[2] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[3] J. Yamagishi *et al.*, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech 2009*, Brighton, U.K., Sep. 2009, pp. 420–423.

[4] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Speech, Audio & Language Process.*, 2010, (in press).

[5] S. Creer, P. Green, S. Cunningham, and J. Yamagishi, "Building personalised synthesised voices for individuals with dysarthria using the HTS toolkit," in *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, J. W. Mullennix and S. E. Stern, Eds. IGI Global, Jan. 2010.

[6] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[7] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 1, pp. 66–83, 1 2009.

[8] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki, "1993 benchmark tests for the ARPA spoken language program," in *HLT '94: Proceedings of the workshop on Human Language Technology*. Morristown, NJ, USA: Association for Computational Linguistics, 1994, pp. 49–74.

[9] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[10] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[11] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[12] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, Harriman, New York, 1992, pp. 357–362.

[13] M. Shozakai and G. Nagino, "Analysis of speaking styles by two-dimensional visualization of aggregate of acoustic models," in *Proc. ICSLP 2004*, Jeju Island, Korea, Oct. 2004, pp. 717–720.

[14] A. Maier, M. Schuster, U. Eysholdt, T. Haderlein, T. Cincarek, S. Steidl, A. Batliner, S. Wenhardt, and E. Nöth, "QMOS – a robust visualization method for speaker dependencies with different microphones," *Journal of Pattern Recognition Research*, vol. 1, pp. 32 – 51, 2009.

[15] T. Cox and M. Cox, *Multidimensional Scaling*. Chapman and Hall, 2001.

[16] K. Tokuda, H. Zen, and T. Kitamura, "Reformulating the HMM as a trajectory model," *IEICE technical report. Natural language understanding and models of communication*, vol. 104, no. 538, pp. 43–48, Dec. 2004.

[17] D. Erro, "Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models," Ph.D. dissertation, Universitat Politecnica de Catalunya, 2008.

[18] J. H. Langlois and L. A. Roggman, "Attractive faces are only average," *Psychological Science*, vol. 1, no. 2, pp. 115–121, 1990.

[19] L. Bruckert, P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G. A. Rousselet, H. Kawahara, and P. Belin, "Vocal attractiveness increases by averaging." *Current Biology*, vol. 20, no. 2, pp. 116–120, 2010.