Analysis of speaker clustering strategies for HMM-based speech synthesis

Rasmus Dall, Christophe Veaux, Junichi Yamagishi, Simon King

The Centre for Speech Technology Research, The University of Edinburgh, U.K.

R.Dall@sms.ed.ac.uk, cveaux@inf.ed.ac.uk, jyamagis@inf.ed.ac.uk, simon.king@ed.ac.uk

Abstract

This paper describes a method for speaker clustering, with the application of building average voice models for speaker-adaptive HMM-based speech synthesis that are a good basis for adapting to specific target speakers. Our main hypothesis is that using perceptually similar speakers to build the average voice model will be better than use unselected speakers, even if the amount of data available from perceptually similar speakers is smaller. We measure the perceived similarities among a group of 30 female speakers in a listening test and then apply multiple linear regression to automatically predict these listener judgements of speaker similarity and thus to identify similar speakers automatically. We then compare a variety of average voice models trained on either speakers who were perceptually judged to be similar to the target speaker, or speakers selected by the multiple linear regression, or a large global set of unselected speakers. We find that the average voice model trained on perceptually similar speakers provides better performance than the global model, even though the latter is trained on more data, confirming our main hypothesis. However, the average voice model using speakers selected automatically by the multiple linear regression does not reach the same level of performance.

Index Terms: Statistical parametric speech synthesis, hidden Markov models, speaker adaptation

1. Introduction

One of the advantages of HMM-based speech synthesis [1] over unit selection is the ability to perform speaker adaptation, which allows text-to-speech synthesizers to be built for a target voice by starting from a well-trained average voice model and then using relatively small amounts of data from the target speaker [2]. Our recent analyses of speaker adaptation performance have found that the quality/naturalness of synthetic speech of adapted voices is moderately correlated with how "far" the transform has had to move away from the average voice model; transforming the average voice model "further" tends to degrade quality [3]. The distance measures we have used to quantify how "far" the transform has moved include mel-cepstral distance [3] and logF_0-F_1 distance [4].

These results suggest that, for best performance, the average voice model should be designed with a particular target speaker in mind. In other words, multiple average voice models are required. We therefore hypothesize that applying speaker clustering to select speakers from which to train the average voice model, and choosing speakers who sound similar to the intended target speaker, will effectively reduce the transform distance and thus improve the performance of speaker adaptation in HMM-based speech synthesis. The main hypothesis of the work we present here is that this form of speaker selection will improve quality *even if it reduces the amount of data on which the model is trained*: "better data" beats "more data".

One approach is to use the same acoustic features for speaker clustering that will be used by the synthesis HMMs and to maximize the likelihood of the clustering and the HMMs simultaneously. An alternative is to use perceptual information about speaker similarity and to identify clusters of speakers prior to training the synthesis HMMs. The latter approach has the advantage that it offers the possibility of incorporating additional information, gained from perceptual experiments, so this is what we choose to do. Our experimental questions include: how to obtain human perceptual judgements of speaker similarity, what criteria should be used for automatic speaker selection to approximate these human judgements, and whether there is a trade-off between the specificity of the speaker cluster vs. the number of speakers (and therefore amount of data) per cluster.

Our experiments are restricted to a single gender because a) gender-dependent average voice models are a better starting point for speaker adaptation [2] and b) Murry & Singh found that different perceptual strategies are used by listeners to distinguish speakers across gender [5]. To test our main hypothesis and examine the above questions, we used speech from a set of female speakers and carried out perceptual tests to investigate similarity among these speakers. We also sought objective measures (i.e., not based on perceptual data) that correlate with listeners' judgements of speaker similarity. Generally the effect on naturalness and overall quality of the synthetic speech is not directly evaluated, however closer similarity to a natural target provides implicit support. Section 2 explains our perceptual study of speaker similarity. Section 3 reports prediction results of the objective measures. Section 4 provides the results from the speech synthesis experiment and discusses some outstanding issues that need further work.

2. Perceptual Study of Speaker Similarity

2.1. Speaker Database

For the perceptual study of speaker similarity, we used speech from 30 female speakers who each said the same sentence "People look, but no one ever finds it." These utterances were recorded using an identical recording setup: an omni-directional microphone, 96kHz sampling frequency at 24 bits and in a hemi-anechoic chamber. All recordings were normalized to -26 dBov based on ITU-T P.56 and were manually end-pointed.

These speakers were selected so as to have a large spread in age (from 19 years old to 64 years old, mean = 36.8, sd = 12.3) and to include various accents (Scottish English, Irish English, Other UK and North American – SE, IE, OU and NA, respectively.). Because all speakers were recruited and recorded in Edinburgh, the sample is inevitably biased: over half the speakers (16 of 30) were Scottish English speakers.

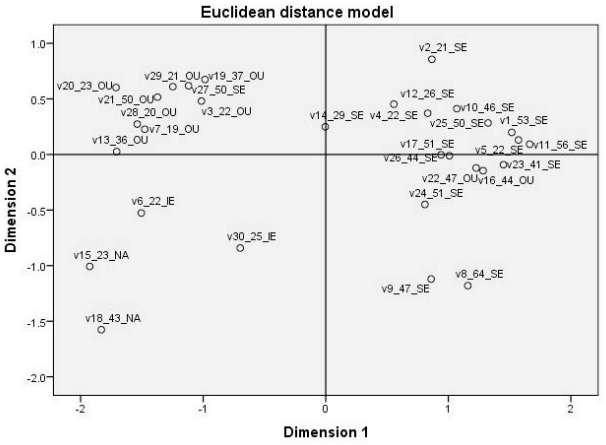

Figure 1: Results of a two-dimensional multi-dimensional scaling (MDS) analysis. Each voice is labeled with id_age_accent. The accents are SE - Scottish English, IE - Irish English, OU - Other UK, and NA - North American.

2.2. Listening Test of Speaker Similarity

For the listening test to measure similarity, all 900 possible speaker pairs (30x30) were included and each pair was presented four times. Identical pairs were included also to provide a means of consistency-checking. 20 listeners participated in the test, which took place in sound-proof listening booths using good quality audio presentation equipment and headphones. Each participant was presented with three rounds of 60 pairs, for a total of 180 pairs. The pairs were ordered in 5 groups such that no participant listened to the same pair twice, except for reverse order pairs, and presentation order was randomised such that no participant listened to all pairs in the same order.

Participants were asked to listen to each pair of sentences and rate their similarity on a 4-point scale in which the points were labelled 1: Very dissimilar, 2: Dissimilar, 3: Similar, and 4: Very similar. They were told that they could listen to each pair more than once and were encouraged to take breaks if needed; between each round, participants could take a longer break. Participants were not given any further instruction as to what kind of similarity to rate; if asked, the experimenter gave a deliberately vague answer that they should rate according to what they found similar. This was in order to avoid biasing participants towards rating any specific form of similarity.

2.3. MDS Analysis

The resulting scores were averaged across listeners and presentations for each speaker pair and a two-dimensional multi-dimensional scaling (MDS) analysis was performed in order to visualize the results. The MDS space is shown in Figure 1: there do appear to be clusters of similar speakers.

In addition to their speech, we have additional meta-data about the speakers in our set, such as self-reported accent, geographical information, age, etc. Inspection of the metadata in connection with the MDS space reveals a general division according to "accent" – each accent forms a distinct cluster. Only three speakers deviate from this pattern: one SE speaker in the OU cluster and two OU speakers in the SE cluster. Another tendency is that speaker age is lower on the left and increases towards the right – this might be a consequence of the SE speakers being generally older than OU speakers, although the two OU speakers in the SE cluster are both relatively old, which suggests otherwise.

3. Prediction of Speaker Similarity Scores

Since perceptual data are expensive to obtain, and because the size of the perceptual test scales badly with the number of speakers being compared, it is desirable to find an objective measure which can perform the same job.

3.1. Stepwise Multiple Linear Regression

To explore which acoustic and meta-data factors are related to perceptual judgements of speaker similarity and thus to predict such speaker similarity scores automatically, a stepwise multiple linear regression analysis was performed. We used SPSS Statistics version 19. The stepwise multiple linear regression uses a variation of the forward algorithm, in which the significance of the change in the F-score is used as a criterion to add explanatory variables to the multiple linear regression. As the variables for multiple regression, we considered the following factors and used Euclidean distances of the individual factors as the actual explanatory variables (except age and accent). Acoustic variables are calculated from audio downsampled to 48kHz at 16 bit depth.

- **Accent**: A binary accent decision was made in which a score of 1 was given to a pair if they were of the same accent group (as used in the MDS) and 0 if the accent differed.
- **Age**: The difference in age (in years) between the speakers.
- **Duration**: Two measures of duration were used, one for the whole sentence excluding pauses and phone duration.
- **Mel-cepstral coefficients**: Six different subsets of the Mel-cepstral co-efficients were used: low-range (dimensions 1-20), mid-range (dimensions 21-40), high-range (dimensions 41-59), mid-low (1-40), mid-high (21-59) and c0 alone. Each of these were used as the average over the whole sentence and as the average for each phone. The Mel-cepstral coefficients were extracted from the STRAIGHT spectrum [6].
- **Aperiodic Component (AC)**: The full band and two sub-bands (0-4 kHz and above 4-24 kHz) of AC [7] were included, both as an average value for the whole sentence and the average per phone.
- **Discrete Cosine Transformation (DCT) of log F_0**: A DCT was applied to the log F_0 values of voiced segments of phrases and the whole sentence [8]; the first three DCT coefficients were used. The 0th DCT coefficient is the mean of the logF0 of the segments, and the 1st and 2nd DCT coefficients are expected to capture F0 tilt and local prosody to some extent, respectively. We calculated the mean of the three DCT coefficients over the whole sentence (where F0 exists).
- **Jitter and Shimmer**: Jitter and shimmer are both measures of perturbations in the vibrations of the vocal folds. This is speaker-dependent and contributes to perceived voice quality. Jitter quantifies perturbations in periodicity and shimmer quantifies perturbations in intensity of vibration. They were included because they have been used to describe voice quality of the elderly [9] and between-gender differences [10], amongst other things, and they also seem to relate to perceived qualities such as hoarseness and roughness [11].
- **Harmonic-to-Noise Ratio (HNR)**: While similar to AC; HMN is a ratio of the amount of non-periodic energy in the speech where AC is the component itself. HMN reflects the

Table 1: Stepwise Multiple Linear Regression Results. The numbers represent the weight (beta-coefficient) of the explanatory variables automatically chosen using the forward algorithm based on the F-score. + represents mel-cepstral distance using low-mid dimensions (1-40). Adjusted R^2 values are shown in the last column.

	Explanatory Variables														
			Duration		Mcep		Aperiodicity			DCT F_0					
Model	Accent	Age	Phone	Sent	Phone	Sent	All freq	0-4kHz	4kHz–	0th	1st	2nd	Jitter	Shimmer	HNR	SpecTilt	Adjusted R^2
Full	-0.62		-0.27				-0.14				-0.14		-0.08	0.14	-0.12		0.63
Meta	-0.67	-0.13															0.49
Acoustic			-0.30		-0.16 +				-0.11		-0.22		-0.28	0.12			0.29

ratio between harmonic (periodic) vs noise (non-periodic) energy in the voice; it is represented as a log ratio in dB. HNR has been found to vary with "vocal age" [12] and relate to "vocal attractiveness" [13].

- **Spectral Tilt**: Spectral tilt is a measure of a speech signal's distribution of power against frequency. The spectral tilt may be used as a cue by listeners [14] and is one of the key properties of Lombard speech [15]. The calculation of the spectral tilt was based on the definition in [15].

The first two variables are self-reported meta-data features and the remainder are continuously-values acoustic properties. We performed three sets of stepwise multiple linear regression: one including both the meta features and acoustic features, one using only the meta features, and one using only the acoustic features. They will be referred to as the "Full," "Meta," and "Acoustic" regression models, respectively.

3.2. Results of Stepwise Multiple Linear Regression

Table 1 shows the results of the stepwise multiple linear regression. The numbers are the weight of explanatory variables automatically chosen using the forward algorithm based on the F-score. + denotes the cases where the Mel-cepstral distance using low-mid dimensions (1-40) was the chosen variable. Adjusted R^2 values are given in the last column. We can see that meta-data-based multiple regression obtains better R^2 value than acoustic feature-based regression, but that they are complementary: combining meta-data and acoustic features results in a multiple regression model that can explain 63% of the variation in the perceptual judgements.

In contrast to the results of others (e.g. [16]), Mel-cepstral features were not chosen in our "Full" regression model. This is presumed to be because we have an accent variable that itself implicitly explains phonetic (segmental) differences among accents. It is also interesting that our regression models (full and acoustic) did not choose the 0th DCT coefficient (which is the average of log F_0), but rather the 1st DCT coefficient. Since our speakers have varied accents, F0 movements represented by the 1st DCT coefficient appear to be more useful to listeners in distinguishing between speakers than differences of global average log F_0.

4. Average Voice Models Using Speaker Clustering

4.1. Speaker Clustering for Average Voice Model Training

In order to verify the effect of speaker clustering for average voice model training, a second perceptual similarity test was conducted using HMM-based speech synthesis. Three of the speakers used in the first test were selected as target speakers for speaker adaptation, one from the SE cluster (v26), one OU (v3) and one NA (v15). For each target speaker, the seven average voice models shown in Table 2 were constructed. All models, except "Average", were trained on eight speakers using 400 sentences each giving a total of 3200 sentences. For each condition the average voice model was adapted to the target speaker using only 25 sentences. For full details of the HMM-based speech synthesis and speaker adaptation methods used in the experiment, please refer to [17].

Table 2: Average Voice Model Definition

Condition	Description
Average	Global average voice model trained on 29 female speakers and a total of 9966 sentences
	Perceptual selection
Perceptual	Eight closest speakers selected using the perceptual similarity scores
MDS	Eight closest speakers selected using Euclidian distance in the 2d MDS perceptual space
	Automatic selection
Full	Eight closest speakers selected using the Full regression model
Meta	Eight speakers selected using the Meta regression model
Acoustic	Eight closest speakers selected using the Acoustic regression model
Random	Eight randomly chosen speakers

4.2. Listening Test

Natural speech from each target speaker was used as the reference (X). All permutations, except identical samples, of synthetic sentence pairs matching each target speaker were created and presented in both orders (XAB and XBA), to avoid recency effects, for a total of 146 target/samples-pairs. 40 native speakers of English were recruited to take part. Each participant was presented with all 146 stimuli and asked to judge which they found to be the most similar to the reference natural speech. The test was divided into three sections, in which participants made 50 judgements in section 1 and 2 and the remaining 46 in section 3; the order of presentation was randomized per participant. Between each section, participants had a small break with shorter breaks during each section. The listening test took approximately 50 minutes to complete. With 40 participants each making 146 judgements each the total number of judgements

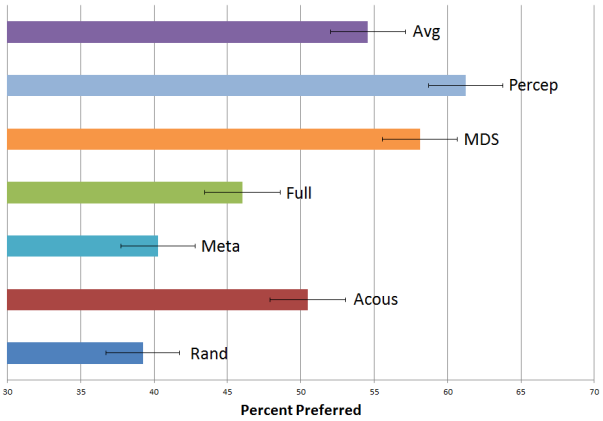

Figure 2: Results of the ABX test

was 5840.

4.3. Results and Discussion

Figure 2 shows the averaged preferred percentage scores with 95% confidence intervals. First we can see that (as expected) the Percep condition provides the best results. The MDS condition is also as good as the Percep condition. Both are significantly better than all other conditions. This shows that choosing perceptually similar speakers improves speaker adaptation performance, supporting our main hypothesis. More specifically, an average voice model trained on perceptually similar speakers using only 3,200 sentences provides better speaker adaptation performance than adaptation from a global average voice model using 9,966 sentences – about three times amount of data.

However, all of the automatic methods for choosing perceptually similar speakers (Full, Acoustic, Meta) have worse performance than the methods based directly on perceptual data (Percep, MDS). Despite seemingly reasonable R^2 values (0.63), the linear regression models are not able to choose perceptually similar speakers. Surprisingly, the preference score of the Meta model is worse than that of the Acoustic model and in fact just as as bad as the Rand condition, even though the Meta model has a higher R^2 value than the Acoustic one. A possible explanation could be that average voice models using the Meta-feature regression do not always reduce the required transform distance; this method makes no use of acoustic information.

5. Conclusion

This paper has introduced several methods for speaker clustering for speaker-adaptive HMM-based speech synthesis. We have confirmed our main hypothesis: it is better to use a smaller number of carefully chosen speakers than a large number of arbitrary speakers. However, the only methods which can successfully identify such speakers require perceptual judgements. Automatic methods, even though learned from such perceptual data, fail to select appropriate speakers. It remains future work to find an automatic method for speaker selection: this is necessary to scale the method up, and in particular, to be able to create average voice models for new target speakers in cases where we have no perceptual data on that target speaker.

6. Acknowledgment

The research leading to these results was partly funded from EPSRC grants EP/I031022/1 and EP/J002526/1.

7. References

[1] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[2] 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.

[3] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, 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.*, vol. 18, pp. 984–1004, Jul. 2010.

[4] S. Andraszewicz, J. Yamagishi, and S. King, "Vocal attractiveness of statistical speech synthesisers," in *Proc. ICASSP 2011*, May 2011, pp. 5368–5371.

[5] S. Singh and T. Murry, "Multidimensional classification of normal voice qualities," *Journal of the Acoustical Society of America*, vol. 64, no. 1, pp. 81–87, 1978.

[6] 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 Commun.*, vol. 27, pp. 187–207, 1999.

[7] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *Proc. Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, pp. 1–6, 2001.

[8] J. Teutenberg, C. Wason, and P. Riddle, "Modelling and synthesising F0 contours with the discrete cosine transform," in *Proc ICASSP 2008*, 2008, vol. 2008, pp. 3973–3976.

[9] F. Hodge, R. Colton, and R. Kelley, "Vocal intensity characteristics in normal and elderly speakers," *Journal of Voice*, vol. 15, no. 4, pp. 503–511, 2001.

[10] M. Brockmann, C. Storck, P. N. Carding, and M. J. Drinnan, "Voice loudness and gender effects on jitter and shimmer in healthy adults," *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 1152–1160, 2008.

[11] P. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchmann, and B. Millet, "Differentiated perceptual evaluation of pathological voice quality: Reliability and correlations with acoustic measurements," *Revue de laryngologie, dotologie et de rhinologie*, vol. 117, no. 3, pp. 219–224, 1996.

[12] C. T. Ferrand, "Harmonics-to-noise ratio: An index of vocal aging," *Journal of Voice*, vol. 16, no. 4, pp. 480–487, 2002.

[13] Laetitia Bruckert, Patricia Bestelmeyer, Marianne Latinus, Julien Rouger, Ian Charest, Guillaume A. Rousselet, Hideki Kawahara, and Pascal Belin, "Vocal attractiveness increases by averaging," *Current Biology*, vol. 20, no. 2, pp. 116–120, 2010.

[14] E. D. Thiessen and J. R. Saffran, "Spectral tilt as a cue to word segmentation in infancy and adulthood," *Perception and Psychophysics*, vol. 66, no. 5, pp. 779–791, 2004.

[15] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, pp. 1253–1262, 2009.

[16] Y. Ijima, M. Isogai, and H. Mizuno, "Correlation analysis of acoustic features with perceptual voice quality similarity for similar speaker selection," in *Proc. Interspeech 2011*, 2011, vol. 2011, pp. 2237–2240.

[17] Junichi Yamagishi and Oliver Watts, "The CSTR/EMIME HTS system for Blizzard Challenge 2010," in *Proc. Blizzard Challenge 2010*, 2010.