CLUSTER ADAPTIVE TRAINING OF AVERAGE VOICE MODELS

Vincent Wan, Javier Latorre, Kayoko Yanagisawa, Mark Gales, Yannis Stylianou

Toshiba Research Europe Ltd., Cambridge, UK.

ABSTRACT

Hidden Markov model based text-to-speech systems may be adapted so that the synthesised speech sounds like a particular person. The average voice model (AVM) approach uses linear transforms to achieve this while multiple decision tree cluster adaptive training (CAT) represents different speakers as points in a low dimensional space. This paper describes a novel combination of CAT and AVM for modelling speakers. CAT yields higher quality synthetic speech than AVMs but AVMs model the target speaker better. The resulting combination may be interpreted as a more powerful version of the AVM. Results show that the combination achieves better target speaker similarity when compared with both AVM and CAT while the speech quality is in-between AVM and CAT.

Index Terms— Speech synthesis, cluster adaptive training, average voice model, voice cloning

1. INTRODUCTION

One of the strengths of hidden Markov model (HMM) based parametric speech synthesis systems is their flexibility to be adapted to a target speaker using linear approaches. This allows a new voice to be created rapidly. To achieve good quality synthesis, a substantial quantity of speech data is needed to cover all the sounds in a language. Early systems [1, 2] were trained on hours of data from one speaker only but more data may be obtained from several speakers. Average voice models (AVMs) [3] are more sophisticated. They use linear transforms to capture speaker dependent attributes leaving the HMMs to model the sounds of the language. Removing speaker specific attributes from the speech allows many speakers to be pooled in order to achieve the required phonetic coverage. A new voice may then be created without completely retraining the HMMs by applying a linear transform, which may be learned from samples of the target speech. Linear transforms are quite powerful because they allow a full matrix transformation and a translation in the parameter space. However, they have a sizable number of parameters. Thus, a moderate quantity of speech may still be needed. Furthermore, noise in the adaptation data may distort the speech. Another limitation of the AVM is in the decision tree that controls the grouping or clustering of speech units within the model. The AVM learns a single structure for all speakers in the training data. However for a new speaker that structure may be suboptimal.

Recently, cluster adaptive training (CAT) was proposed for speech synthesis. It was originally used for rapid speaker adaptation in automatic speech recognition [4] and is analogous to Eigenvoices [5, 6]. It was modified for speech synthesis by the addition of multiple decision trees. It has been used to capture language variation in polyglot synthesis [7], to model speaker variation [8], expressiveness in audiobooks [9] and to factorise speaker and emotion [10]. Unlike AVMs the means of the probability density functions are an interpolation of those obtained from several decision trees. To adapt a CAT model to a new speaker, only the interpolation weights need to be estimated. Since these are few in number, CAT models may be adapted using very little data [8] and are more robust to noise during adaptation [11]. These weights may be interpreted as a point in a speaker space. However, the range of voices that can be synthesised from the speaker space is more restricted compared with the range of voices from an AVM. This is because AVMs incorporate a matrix transformation that is missing from the formulation of CAT used in [8, 9, 10]. The matrix transformation is important: it is known that a change in vocal tract length is equivalent to a matrix transformation applied to Mel-cepstral features [12, 13].

This paper therefore studies a combination of CAT with linear transforms (described in section 4) to address the short-comings of the CAT formulation. The proposed combination may be interpreted as a more general form of AVM. Sections 2 and 3 describe the AVM and CAT models. Section 5 describes the experiments and results and section 6 concludes.

2. AVERAGE VOICE MODEL (AVM)

The standard AVM [3] uses speaker dependent constrained maximum likelihood linear regression (CMLLR) transforms [14] to remove speaker dependent traits and normalise the features. The CMLLR transform, defined by $\{\boldsymbol{A}_r^{(s)}, \boldsymbol{b}_r^{(s)}\}$, maps an observation vector $\boldsymbol{o}(t)$ at frame t according to

$$\hat{\boldsymbol{o}}_{r(m)}^{(s)}(t) = \boldsymbol{A}_{r(m)}^{(s)}\boldsymbol{o}(t) + \boldsymbol{b}_{r(m)}^{(s)} \tag{1}$$

where s denotes the speaker and $r(m)$ the regression class of the m^{th} Gaussian component. The emission probability of $\boldsymbol{o}(t)$, uttered by s and given component m may be expressed as

$$p(\boldsymbol{o}(t) \mid m, s, \mathcal{M}) = \left|\boldsymbol{A}_{r(m)}^{(s)}\right| \mathcal{N}\left(\hat{\boldsymbol{o}}_{r(m)}^{(s)}(t); \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right) \tag{2}$$

where $\mathcal{M} = \{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \boldsymbol{A}_{r(m)}^{(s)}, \boldsymbol{b}_{r(m)}^{(s)}\}$ is the set of model parameters; $\boldsymbol{\mu}_m$ is the mean vector and $\boldsymbol{\Sigma}_m$ the covariance matrix of the m^{th} Gaussian component. The parameters are typically split into two distinct parts which are estimated separately: the Gaussian (or canonical) parameters $\{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$ and the transform parameters $\{\boldsymbol{A}_{r(m)}^{(s)}, \boldsymbol{b}_{r(m)}^{(s)}\}$.

3. CAT MODEL

The CAT model used here is the multiple decision tree model from [8]. The underlying idea, illustrated in Fig. 1, is that the means of the distributions are defined as a linear combination of multiple mean vectors from different clusters, where each cluster's structure is defined by its own decision tree. In this way, the mean $\boldsymbol{\mu}_m^{(s)}$ of component m, for speaker s is obtained by interpolating the mean vector of each cluster,

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

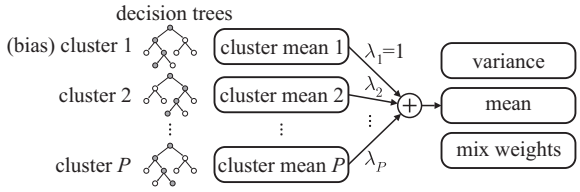

Fig. 1. CAT with cluster-dependent decision trees.

$$\boldsymbol{\mu}_m^{(s)} = \sum_{i=1}^{P} \boldsymbol{\mu}_{c(m,i)} \lambda_{i,w(m)}^{(s)} \tag{3}$$

where i is the cluster index, P the total number of clusters, $\boldsymbol{\mu}_{c(m,i)}$ the cluster mean vector for component m of the i^{th} cluster and $\lambda_{i,w(m)}^{(s)}$ the interpolation weight (CAT weight) associated with cluster i for the $w(m)$ weights regression class. The associated output probability is

$$p(\boldsymbol{o}(t) \mid m, s, \mathcal{M}) = \mathcal{N}\left(\boldsymbol{o}(t); \boldsymbol{\mu}_m^{(s)}, \boldsymbol{\Sigma}_{v(m)}\right) \tag{4}$$

where $v(m)$ is the index of the covariance associated with component m and $\mathcal{M} = \{\boldsymbol{\mu}_{c(m,i)}, \boldsymbol{\Sigma}_{v(m)}, \boldsymbol{\lambda}_{r(m)}^{(s)}\}$ are the model parameters. The optimisation of $\mathcal{M}$ is done iteratively. Given an initial set of CAT weights, the canonical parameters are first estimated. Decision trees are built iteratively on a cluster by cluster basis [15] so that when building the trees for one cluster, the other trees and their canonical parameters are held fixed. A greedy algorithm is used to grow the trees. The question asked at each node is chosen to maximise the log-likelihood gain. In order to maintain a balance between complexity and accuracy, tree growing is stopped by a combination of MDL [16] and 10-fold cross-validation [17]. Then, the CAT weights are estimated as per [4]. In order to keep the scale of the CAT weights the weights of the first cluster are fixed $\lambda_{1,w(m)} = 1.0$. This implies that the first cluster is always added. Therefore, it is called the bias cluster. Covariances and mixture weights could have their own decision tree but they are usually tied together with the bias cluster so that $v(m) = c(m, 1) \forall m$.

4. COMBINING CAT WITH CMLLR

One of the reasons why people sound different is the difference in the size of their vocal tracts [18]. In the spectral domain, a change of the vocal tract length is equivalent to a warping of the frequency axis. For Mel-cepstral coefficients this warping may be modelled as a matrix transformation [12, 13] that rotates the coefficient vectors at each frame. One way to implement such a matrix is via a CMLLR transform applied globally to the speech. Although a CAT transformation can produce a set of context specific shifts of the Mel-cepstral coefficients, it cannot produce such rotations. Therefore, a natural way to improve similarity to target speakers is by combining the fine context dependency provided by CAT with the global warping provided by a CMLLR. The CMLLR is defined without a translation term:

$$\hat{\boldsymbol{o}}_{r(m)}^{(s)}(t) = \boldsymbol{A}_{r(m)}^{(s)} \boldsymbol{o}(t) \tag{5}$$

with regression classes $r(m)$ for silence, pause and speech. Combining equations (5) and (3), the emission probability becomes

$$p(\boldsymbol{o}(t) \mid m, s, \mathcal{M}) = \left|\boldsymbol{A}_{r(m)}^{(s)}\right| \mathcal{N}\left(\hat{\boldsymbol{o}}_{r(m)}^{(s)}(t); \sum_{i=1}^{P} \boldsymbol{\mu}_{c(m,i)} \lambda_{i,w(m)}^{(s)}, \boldsymbol{\Sigma}_{v(m)}\right) \tag{6}$$

4.1. CAT+CMLLR

A simple way to combine CAT and CMLLR is to apply the CMLLR transform on top of a CAT model at adaptation time only. This approach is analogous to applying a CMLLR to a speaker independent model. Estimation of the CMLLR and CAT weights is performed iteratively. First, the CAT model is adapted to the target by estimating a set of CAT weights for that speaker (as per section 5.2 baseline CAT model). Then an identity matrix CMLLR transform is introduced. The CMLLR transform and CAT weights are then updated alternately until the log-likelihood converges. The final CMLLR transform and CAT weights are used for synthesis.

4.2. CATAVM

In the CAT+CMLLR approach the CMLLR transform is only used during adaptation synthesis. This introduces a mismatch with the training procedure. To avoid it, the CMLLR transforms can be applied during training within an integrated speaker adaptive training (SAT) framework. This model is conceptually like a standard AVM model in which the transformed space is modelled by a CAT model instead of by a standard single tree model.

4.2.1. Parameter estimation

In training, the expectation-maximisation (EM) algorithm is used to maximise the log likelihood of the model given the training speech data, associated transcriptions and speaker labels. The canonical parameters, CMLLR transforms, CAT weights and decision trees are each updated separately in an iterative fashion. The process is that used for speaker adaptive training. First, given a set of canonical parameters, the CMLLR and CAT weights are estimated iteratively: For a given set of CAT weights, the interpolated mean vectors $\boldsymbol{\mu}_m^{(s)}$ are used to estimate the CMLLR transforms, as in a standard AVM [14]. Then, given the transformed observations $\hat{\boldsymbol{o}}_{r(m)}^{(s)}(t)$, the CAT weights are obtained as in the normal CAT model. After the new CAT weights and CMLLR transforms have been obtained, new canonical parameters and decision trees are computed in the same way as for a standard CAT model. The process is repeated a fixed number of iterations or until the log-likelihood converges. A detailed description of the CATAVM framework and its training can be found in [7]. There, standard CMLLRs were used to model speaker identity and CAT weights to model languages. In this paper a specific form of CMLLR and the CAT weights are used in combination to model speaker identity.

4.2.2. Model initialisation

There are several ways to initialise the proposed CATAVM model. One option is to start from a CAT model (described in section 3), estimate CMLLR transforms for it, then iteratively update decision trees, canonical parameters, CAT weights and CMLLR transforms in a speaker adaptive training scheme. Since the CMLLRs act upon the observation vectors, the trees may require many iterations of updates to converge since they were originally optimised on the untransformed observation vectors.

In this paper, the CATAVM is initialised from a standard AVM with separate CMLLR regression classes for speech, silence and pause. An identity transform is enforced for silence and pause. The first CAT cluster is initialised from the AVM canonical model. The AVM's CMLLR transforms have their translation terms stripped away leaving only the $\boldsymbol{A}_{r(m)}$ matrices. The other CAT clusters are

then initialised with a zero mean root node[1]. With the CAT weights set to one/zero values depending upon each speaker's cluster assignment [8], the decision trees for each cluster are estimated using the alignments from the AVM with unstripped CMLLRs. Once the clusters are initialised then the speaker adaptive training scheme mentioned above is applied.

4.3. Mathematical interpretation

Since the bias cluster is always added with weight $\lambda^{(s)}_{1,w(m)} = 1.0$, and the tying structure of the covariance is the same as that of the bias cluster, equation (6) could be rewritten as

$$p(\boldsymbol{o}(t) \mid m, s, \mathcal{M}) = \left|\boldsymbol{A}^{(s)}_{r(m)}\right| \mathcal{N}\left(\bar{\boldsymbol{o}}^{(s)}_{m}(t); \boldsymbol{\mu}_{c(m,1)}, \boldsymbol{\Sigma}_{c(m,1)}\right) \tag{7}$$

where

$$\bar{\boldsymbol{o}}^{(s)}_{m}(t) = \boldsymbol{A}^{(s)}_{r(m)}\boldsymbol{o}(t) + \boldsymbol{b}^{(s)}_{m} \tag{8}$$

$$\boldsymbol{b}^{(s)}_{m} = -\sum_{i=2}^{P} \boldsymbol{\mu}_{c(m,i)} \lambda^{(s)}_{i,w(m)} \tag{9}$$

In other words, a combination of CAT with CMLLR may be viewed as a CMLLR transform having different regression structures for its rotation matrix $\boldsymbol{A}$ and translation vector $\boldsymbol{b}$. If the matrix transformation is only needed to account for vocal tract changes then a global $\boldsymbol{A}$ matrix should suffice. However, there are other context-dependent differences which might be hard to model with a CMLLR transform, unless there is a large amount of adaptation data. For this, CAT-weights are easy to obtain. Moreover, the regression classes in an AVM are the same for both $\boldsymbol{A}$ and $\boldsymbol{b}$ terms and are derived from the canonical model's tree, which is built considering all training speakers. In contrast, the decision trees of each CAT cluster are grown to explicitly maximise the log likelihood of the data assigned to it by the CAT weights. Therefore the trees may be better tuned to accommodate speaker specificities.

5. EXPERIMENTS

5.1. Experimental setup

The same training data that was used in [8] is used here. The speech data are recorded from 4 professional voice talents in studio conditions (2 males each contributing 1.0 and 2.0 hours of speech, and 2 females contributing 3.5 and 4.2 hours of speech). Each spoke with a General American accent of US English in a neutral style.

Adaptation and testing data consists of 10 General American accented speakers selected from four age ranges for each gender: 1 from 13-18, 2 from 20-30, 1 from 30-50 and 1 from 50-70. Each speaker recorded a common set of 50 phonetically rich sentences covering all phones and a second set of 50 sentences, which were chosen from various domains and differed between speakers. The amount of speech ranges between 8 and 15 minutes per speaker. Recordings were made using headset microphones in a standard recording room with low reverberation and no background noise. To avoid effects due to variation in recording conditions, the same microphone and recording room was used for all speakers. Ten sentences chosen at random for each speaker were used in listening tests and the remaining 90 were used for adaptation. The waveforms were down-sampled to 16kHz and parameterised using 40 dimensional Mel-cepstral coefficients, log-F0 and 21 Bark-scale band aperiodicities, each with their first and second order deltas.

All samples were synthesised by a speech generation algorithm including a speaker independent global variance term [19] estimated from the four training speakers. Speech waveforms were synthesised from the generated speech parameters using the source-filter model.

Subjective listening tests were conducted via the crowd-sourcing website *CrowdFlower* using Mechanical Turk workers located in the US [20]. Standard paired-comparison preference tests asked listeners to choose the waveform that sounded better and they were given the option of choosing "no preference". In ABX similarity tests, listeners were played a reference sample from the voice talent followed by two synthesised samples. They were asked to choose the sample that sounded more similar to the speaker in the reference. In this case, listeners could not choose "undecided". Some judgements were discarded according to criteria described in [20].

5.2. Models

The baseline CAT model is identical to that used in [8]. It consists of 5 clusters: a bias plus one cluster per training speaker. The model is initialised considering the imbalance in the amount of training data for each speaker: the initial CAT weights are set to 0/1 values corresponding to the speaker's cluster assignment; the bias cluster is initialised from the speaker dependent model of the dominant (with the most training data) speaker; clusters for the other three speakers are built; the bias cluster is updated on the data of all speakers; then the last cluster is built for the dominant speaker. After initialisation, the canonical parameters and CAT weights are updated iteratively. The process of rebuilding the decision trees and updating the model parameters is repeated twice. To synthesise a new speaker an initial set of CAT weights are copied from one of the training speakers. They are then repeatedly updated to maximise the likelihood given the adaptation data until convergence. The CAT weights converge to the same values irrespective of the starting point.

The baseline AVM is also identical to that used in [8]. It is trained on the four training speakers thus: a speaker independent monophone maximum likelihood model is built then CMLLR speaker adaptive training is applied. The monophone models are cloned to full context models which are clustered using decision trees. Speaker adaptive training continues with block diagonal CMLLR transforms with regression classes for speech, silence and pause. The decision trees, canonical model and CMLLR transforms are updated several times iteratively. A further round of speaker adaptive training is run using regression tree CMLLR transforms, where the regression tree is derived from the AVM's decision tree. The state-duration distributions are treated in the same way. To synthesise a new voice, an initial CMLLR transform is estimated from samples of the target speaker and are refined using CSMAPLR. This is followed by a speaker dependent MAP adaptation of the means. The MAP adapted model is combined with the CSMAPLR transform for synthesis. AVMs occasionally produce distorted durations if the adaptation data contain noise such as microphone popping introduced by poor recording techniques. To prevent systematic biasing of listening tests by distorted durations, the AVM samples were synthesised using durations from the baseline CAT model, which is more robust to noise [11].

The CAT+CMLLR combines a CMLLR transform with the baseline CAT model at adaptation and synthesis time only as described in section 4.1.

[1] If exactly one speaker is assigned to each cluster then it is possible to construct each cluster to replace the translation term of the CMLLR. In this case the initial model would be numerically identical to the AVM. Although the condition is true for this paper it may not be in general.

	AVM	CAT	CAT+CMLLR	CATAVM
mean	2163	36	1632	1857
min	653	11	472	283
max	5042	95	3840	4409
MCD	0.245	0.262	0.237	0.250

Table 1. Objective analyses. *Top:* Mean, min and max inter-speaker distances ($\times 10^3$). *Bottom:* Mel-cepstral distortion (MCD).

		CAT		No	
AVM	CAT	+CMLLR	CATAVM	pref	p
28.8	67.8	-		3.4	< 0.001
28.8	-	49.8	-	21.4	< 0.001
-	48.3	39.7	-	12.0	0.017
-	-	39.8	41.7	18.5	0.316

Table 2. Preference test comparing quality of speech.

		CAT		
AVM	CAT	+CMLLR	CATAVM	p
49.6	50.4	-	-	0.403
41.9	-	58.1	-	< 0.001
-	40.0	60.0	-	< 0.001
-	-	50.3	49.7	0.450

Table 3. ABX test comparing similarity to target speaker.

The CATAVM is trained as described in section 4.2 where all model parameters are updated within an integrated speaker adaptive training set up. Adaptation to test speakers is carried out using the same approach as the CAT+CMLLR model as described in section 4.1 with the exception that the CMLLR and CAT weights are initialised from the closest training speaker.

5.3. Objective evaluations

To get an indication of the range of speakers that each model can synthesise, the inter-speaker distance between the synthesised outputs of each method is measured. Each model synthesises a given sentence for different speakers but with durations fixed at the context level so that the same number of frames are generated; the generated Mel-cepstral features are concatenated into a super-vector; the inter-speaker distance is the Euclidean distance between the super-vectors; inter-speaker distance statistics are computed over five sentences. Results for each model are shown in table 1. The smaller values of CAT indicate that the generated speech of different target speakers are closer together and, therefore, not so distinct from each other. In contrast, AVM, CAT+CMLLR and CATAVM have larger values indicating the samples are more distinctive from each other. This highlights the importance of the matrix transform.

To get an objective measure of the performance of the different models, the Mel-cepstral distortion (MCD) were computed between the original signal and the synthesised version. For this purpose, the test sentences were first aligned with the adapted models and the 'original' duration was enforced during the synthesis. These tests were run on the adaptation sentences and non speech sections (silence and pause) were excluded. Table 1 shows the MCD results. The CAT model has the greatest distortion while CAT+CMLLR has the least. However, the ranking from objective measures does not correlate with the subjective test ranking below.

5.4. Subjective evaluation

Table 2 compares the different approaches by subjective pairwise preference test to measure speech quality and table 3 shows the ABX test results for similarity to target speaker. Comparing the AVM and CAT baselines, CAT is shown to have better quality but, despite the results shown in section 5.3, neither is closer than the other to the target speaker. The judgement may be tricky because CAT and AVM samples sound very different from each other.

The CAT+CMLLR approach sounds similar to the AVM and, in tests, is preferred over the AVM both in terms of quality and similarity. Compared to CAT, CAT+CMLLR has better similarity but has lower quality. CAT synthesis sounds better in some unvoiced regions which may account for the result. Thus, adding a matrix transform to CAT improves speaker similarity but sacrifices quality.

No difference between the CAT+CMLLR and CATAVM is observed despite the fact that there is a mismatch between training and adaptation in the CAT+CMLLR model. The complexity of CAT+CMLLR is lower than that of CATAVM since the former does not require the estimation of any CMLLRs at training time. However, this result may partly be due to the fact that each cluster is initialised with exactly one speaker so the role of the matrix transformation may be performed adequately by the individual clusters. If each cluster had to represent more than one speaker then the matrix transformation may have a greater impact.

From the mathematical interpretation described in section 4.3 the improved results may be attributed to two possibilities. Firstly, the regression structures of the $\boldsymbol{b}$ vectors in the AVM and their equivalent in the CATAVM are different: a tree interpolation structure may be better than a single tree. Secondly, it may be more effective to have fewer $\boldsymbol{A}$ matrix transforms that operate on the speech at a higher level; fewer matrices is certainly more efficient because they contain the bulk of the parameters in a CMLLR transform. The CATAVM has fewer parameters that need to be estimated. It has one $\boldsymbol{A}$ matrix and a small number of interpolation weights, while a standard AVM with R regression classes has up to R CMLLR transforms each with their own set of $\{\boldsymbol{A}, \boldsymbol{b}\}$ parameters.

6. CONCLUSION

This paper combines CAT with CMLLR transforms for modelling speakers in HMM-TTS. On a voice cloning task adapting to 90 sentences, listening tests show that the proposed CAT plus a model is better than both AVM and CAT in terms of speaker similarity; it has better quality than the AVM but lower quality than CAT.

The proposed combination may be interpreted as a generalisation of the AVM with multiple decision trees in the canonical model. It is shown that, mathematically, the proposed formulation may be interpreted as an AVM that allows different regression structures for the $\boldsymbol{A}$ and $\boldsymbol{b}$ terms of a CMLLR transform. This is a powerful aspect: given that the number of parameters in the $\boldsymbol{A}$ matrices is much larger than in the $\boldsymbol{b}$ vectors and given that the two terms potentially model different aspects of speech, having them share the same regression structures may be inefficient and/or suboptimal.

7. REFERENCES

[1] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, 1995, pp. 660–663.

[2] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, 2002, CD-ROM Proceeding.

[3] 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. Audio Speech Lang. Process.*, vol. 17, pp. 66–83, 2009.

[4] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, 2000.

[5] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proc. ICSLP*, 1998.

[6] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP*, 2002, pp. 1269–1272.

[7] H. Zen, N. Braunschweiler, S. Buchholz, M. Gales, K. Knill, S. Krstulović, and J. Latorre, "Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 5, 2012.

[8] V. Wan, J. Latorre, KK Chin, L. Chen, M. J. F. Gales, H. Zen, K. Knill, and M. Akamine, "Combining multiple high quality corpora for improving HMM-TTS," in *Proc. Interspeech*, 2012.

[9] L. Chen, M. J. F. Gales, V. Wan, J. Latorre, and M. Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Proc. Interspeech*, 2012.

[10] J. Latorre, V. Wan, M. J. F. Gales, L. Chen, K. Chin, K. Knill, and M. Akamine, "Speech factorization for HMM-TTS based on cluster adaptive training," in *Proc. Interspeech*, 2012.

[11] K. Yanagisawa, J. Latorre, V. Wan, M. J. F. Gales, and S. King, "Noise robustness in HMM-TTS speaker adaptation," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013, pp. 119–124.

[12] M. Pitz and H. Ney, "Vocal tract normalisation equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 930–944, 2005.

[13] D. R. Sanand and S. Umesh, "Study of Jacobian compensation using linear transformation of conventional MFCC for VTLN," in *Interspeech*, 2008.

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

[15] K. Saino, *A clustering technique for factor analyzed voice models*, Master thesis, Nagoya Institute of Technology, 2008.

[16] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. Eurospeech*, 1997, pp. 99–102.

[17] T. Shinozaki, "HMM state clustering based on efficient cross-validation," in *Proc. ICASSP*, 2006, pp. 1157–1160.

[18] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP*, 1996, pp. 346–349.

[19] T. Toda, A.W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. ICASSP*, 2005, pp. 9–12.

[20] S. Buchholz, J. Latorre, and K. Yanagisawa, "Crowdsourced assessment of speech synthesis," in *Crowdsourcing for Speech Processing*, M. Eskenazi, G.-A. Levow, H. M. Meng, G. Parent, and D. Suendermann, Eds., pp. 173–216. John Wiley & Sons, Chichester, 2013.