Speech factorization for HMM-TTS based on cluster adaptive training.

Javier Latorre, Vincent Wan, Mark J. F. Gales, Langzhou Chen, K.K. Chin [*]
Kate Knill, Masami Akamine

Toshiba Research Europe Ltd, 208 Science Park, Milton Road, Cambridge, UK
{javier.latorre,vincent.wan,mjfg,langzhou.chen,kate.knill}@crl.toshiba.co.uk
masa.akamine@toshiba.co.jp

Abstract

This paper presents a novel approach to factorize and control different speech factors in HMM-based TTS systems. In this paper cluster adaptive training (CAT) is used to factorize speaker identity and expressiveness (i.e. emotion). Within a CAT framework, each speech factor can be modelled by a different set of clusters. Users can control speaker identity and expressiveness independently by modifying the weights associated with each set. These weights are defined in a continuous space, so variations of speaker and emotion are also continuous. Additionally, given a speaker which has only neutral-style training data, the approach is able to synthesise speech with that speaker's voice and different expressions. Lastly, the paper discusses how generalization of the basic factorization concept could allow the production of expressive speech from neutral voices for other HMM-TTS systems not based on CAT.

Index Terms: speech synthesis, cluster adaptive training, expressive synthesis, speech factorization

1. Introduction

Speech synthesizers have improved in terms of intelligibility, naturalness and speech quality. However, as applications like e-book readers and dialogue systems become increasingly popular, even more improvement is needed. In these applications, naturalness implies expressiveness but most high quality speech synthesizers only produce speech in a neutral or "reading" style. This problem was identified more than two decades ago and over the years different approaches have been proposed to bring expressiveness into synthetic speech. See [1] for a review. In early techniques, the intended emotion was produced by modifying the prosody and/or the spectrum of the speech signal according to heuristic rules extracted from a comparison between neutral and acted-emotional speech. However, most modern techniques are purely data driven. Some unit selection systems produced different emotions by mixing databases recorded with acted emotions [2], but they require large expression-dependent databases which are costly and tedious to collect. Approaches based on voice conversion were proposed [3, 4] which reduce the amount of expressive data required considerably, but two problems remain: they can only produce speech with the voice of one speaker and they can only produce a discrete number of emotions. Two solutions based on multiple regression Hidden Semi-markov models were proposed [5], in which speech with a continuous variety of expressions could be produced in the voice of any arbitrary target speaker. Although the results of both methods are impressive, a drawback is that they require target speaker data in all the expressions used to define the expression space axis.

This paper addresses the problem of how to control the speaker voice and expression independently so that speech with different expressions can be generated for any target voice, especially when only neutral style data is available for that voice or when the voice is defined as an arbitrary point in the voice space. In other words, this paper studies how to factorize speaker identity and expression. The proposed technique is based on cluster adaptive training (CAT). The concept is similar language and speaker factorization [6]. The main difference here is that both speaker and expression are defined by a vector of cluster weights. This paper extends work described in [7].

Section 2 describes the factorization of emotion and speaker within a CAT framework. Section 3 describes the results of several subjective experiments. Section 4 discusses how the basic concept of CAT factorization could be generalized to other HMM-based systems. Section 5 concludes.

2. Speaker and emotion factorization

The goal of speaker and expression factorization is to obtain two sets of parameters $\boldsymbol{\lambda}_s$ and $\boldsymbol{\lambda}_e$ that enable the speaker voice and the expression to be modified independently. Conceptually, the problem is similar to that of a polyglot synthesizer [6]. Whereas the goal there is to synthesize the voice of a monolingual speaker in different languages, the goal here is to synthesize a target voice with different expressions given target speaker data spoken in a neutral style only. This can be achieved in the CAT framework by defining $\boldsymbol{\lambda}_s$ and $\boldsymbol{\lambda}_e$ as two vectors of weights for two associated sets of clusters. For such a model, the emission probability of an observation vector given component, speaker, emotion and the set of model parameters is

$$p(\boldsymbol{o}(t) \mid m, s, e, \mathcal{M}) =$$

$$\mathcal{N}\left(\boldsymbol{o}(t); \boldsymbol{\mu}_{c(m,1)} + \boldsymbol{M}_m^{(s)}\boldsymbol{\lambda}_{q(m)}^{(s)} + \boldsymbol{M}^{(e)}\boldsymbol{\lambda}_{q(m)}^{(e)}, \boldsymbol{\Sigma}_{v(m)}\right) \tag{1}$$

with

$$\boldsymbol{\lambda}_{q(m)}^{(s)} = \left[\lambda_{2,q(m)}^{(s)}, \ldots, \lambda_{P_s+1,q(m)}^{(s)}\right]^\top \tag{2}$$

$$\boldsymbol{\lambda}_{q(m)}^{(e)} = \left[\lambda_{P_s+2,q(m)}^{(e)}, \ldots, \lambda_{P_s+P_e+1,q(m)}^{(e)}\right]^\top \tag{3}$$

$$\boldsymbol{M}_m^{(s)} = \left[\boldsymbol{\mu}_{c(m,2)}, \ldots, \boldsymbol{\mu}_{c(m,P_s+1)}\right] \tag{4}$$

$$\boldsymbol{M}_m^{(s)} = \left[\boldsymbol{\mu}_{c(m,P_s+2)}, \ldots, \boldsymbol{\mu}_{c(m,P_s+P_e+1)}\right] \tag{5}$$

where $t \in \{1, \ldots, T\}$, $m \in \{1, \ldots, M\}$, $s \in \{1, \ldots, S\}$ and $e \in \{1, \ldots, E\}$ enumerate the frames, Gaussian components, speakers and emotions, respectively; $q(m) \in \{1, \ldots, Q\}$ and $v(m) \in \{1, \ldots, V\}$ are respectively the m^{th} component's CAT regression classes and leaf node in the covariance matrices' decision tree; $c(m, i) \in \{1, \ldots, N\}$ is the leaf node for cluster

[*] K.K. Chin is currently at Google.

i of component m in decision trees for cluster mean vectors; P_s, P_e are the number of speaker and emotion clusters; $\boldsymbol{o}(t)$ is the observation vector at frame t; $\lambda_{i,q}^{(s,e)}$, $\boldsymbol{\lambda}_q^{(s)}, \boldsymbol{\lambda}_q^{(e)}$ are respectively the i^{th} cluster's CAT weight and the weight vectors for speaker s and emotion e associated with CAT regression class q; $\boldsymbol{\mu}_n$ is the cluster mean vector associated with leaf node n; $\boldsymbol{M}_m^{(s)}, \boldsymbol{M}_m^{(e)}$ are component m's matrices of speaker and emotion cluster mean vectors; $\boldsymbol{\Sigma}_k$ is the covariance matrix of leaf node k and $\mathcal{M}$ is the full set of model parameters.

2.1. Model training

The set of model parameters consists of two parts: the canonical parameters, $\boldsymbol{\Lambda} = \{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_k\}$, comprising cluster mean vectors and covariance matrices; and CAT weights, $\boldsymbol{W} = \{\boldsymbol{\lambda}_q^{(s,e)} = [1, \boldsymbol{\lambda}_q^{(s)\top}, \boldsymbol{\lambda}_q^{(e)\top}]^\top\}$, comprising the speaker- and expression-specific CAT interpolation weight vectors. The training process estimates canonical parameters and CAT weights iteratively.

The canonical parameter updates are the same as those in [6]. The main difference here is that the CAT weights model both expression and speakers. Therefore, no regression matrices are required and the accumulated statistics $\boldsymbol{G}_{ij}^{(m)}$ and $\boldsymbol{k}_i^{(m)}$ are defined as

$$\boldsymbol{G}_{ij}^{(m)} = \sum_{t,s,e} \gamma_m(t,s,e) \lambda_{i,q(m)}^{(s,e)} \boldsymbol{\Sigma}_{v(m)}^{-1} \lambda_{j,q(m)}^{(s,e)} \tag{6}$$

$$\boldsymbol{k}_i^{(m)} = \sum_{t,s,e} \gamma_m(t,s,e) \lambda_{i,q(m)}^{(s,e)} \boldsymbol{\Sigma}_{v(m)}^{-1} \boldsymbol{o}(t). \tag{7}$$

where $\gamma_m(t,s,e)$ is the posterior probability of component m generating $\boldsymbol{o}(t)$ given speaker s and expression e.

During CAT weight update, the speaker and expression portions must be kept independent: each weight vector has to be updated separately while holding the other one fixed, thus for the expression weight vector,

$$\boldsymbol{\lambda}_q^{(e)} = \Big(\sum_{\substack{t,m,s \\ q(m)=q}} \gamma_m(t,s,e) \boldsymbol{M}_m^{(e)\top} \boldsymbol{\Sigma}_{v(m)}^{-1} \boldsymbol{M}_m^{(e)} \Big)^{-1} \sum_{\substack{t,m,s \\ q(m)=q}} \gamma_m(t,s,e) \boldsymbol{M}_m^{(e)\top} \boldsymbol{\Sigma}_{v(m)}^{-1}) \hat{\boldsymbol{o}}_{q(m)}^{(s)}(t) \tag{8}$$

where $$\hat{\boldsymbol{o}}_{q(m)}^{(s)}(t) = \boldsymbol{o}(t) - \boldsymbol{\mu}_{c(m,1)} - \boldsymbol{M}_m^{(s)} \boldsymbol{\lambda}_q^{(s)}. \tag{9}$$

The speaker weight vector $\boldsymbol{\lambda}_q^{(s)}$ is obtained similarly.

2.2. Model initialization and tying structure

To keep the two sets of clusters independent, the initialization of the model consists of two steps. The speaker components are first trained using only data tagged as emotionally neutral[1]. The available emotional data is then arranged into P_e groups[2] and the model is extended with P_e new clusters. For each non-neutral file the expression CAT weights are set to 0 or 1 according to the group to which the file belongs so that the new clusters model speaker-independent differences between neutral and emotional speech.

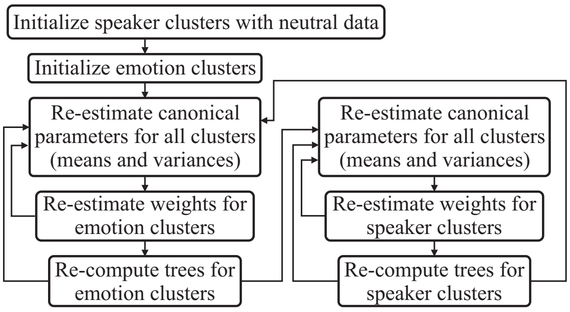
Initialize speaker clusters with neutral data
Initialize emotion clusters
Re-estimate canonical parameters for all clusters (means and variances)
Re-estimate weights for emotion clusters
Re-compute trees for emotion clusters
Re-estimate canonical parameters for all clusters (means and variances)
Re-estimate weights for speaker clusters
Re-compute trees for speaker clusters

Figure 1: Training the speaker and expression CAT model.

Next the parameters of the new cluster are obtained. Similar to [6], the tying structure of each emotion and speaker cluster is defined by its own decision tree. This approach models cluster-specific context dependencies. The decision trees of each cluster are created iteratively cluster-by-cluster [8]: when tying the parameters of one cluster the parameters of all other clusters are held fixed. For each new cluster the decision trees and initial mean vectors are computed simultaneously thus. The accumulated statistics $\boldsymbol{G}_{ij}^{(m)}$ and $\boldsymbol{k}_i^{(m)}$ over all the speaker and emotion clusters are computed. For any given node in the tree, the non-constant part of the log-likelihood can be expressed as

$$\mathcal{L}(n) = \frac{1}{2} \hat{\boldsymbol{\mu}}_n^\top \left(\sum_{m \in \mathcal{S}(n)} \boldsymbol{G}_{ii}^{(m)} \right) \hat{\boldsymbol{\mu}}_n. \tag{10}$$

where $\hat{\boldsymbol{\mu}}_n$ is the ML estimate of $\boldsymbol{\mu}_n$ which is

$$\hat{\boldsymbol{\mu}}_n = \left(\sum_{m \in \mathcal{S}(n)} \boldsymbol{G}_{ii}^{(m)} \right)^{-1} \times \sum_{m \in \mathcal{S}(n)} \left(\boldsymbol{k}_i^{(m)} - \sum_{j \neq i} \boldsymbol{G}_{ij}^{(m)} \boldsymbol{\mu}_{c(m,j)} \right). \tag{11}$$

Therefore the optimum split is given by the question q that maximizes the log-likelihood gain

$$\Delta\mathcal{L}(n;q) = \mathcal{L}(n_+^q) + \mathcal{L}(n_-^q) - \mathcal{L}(n). \tag{12}$$

After the initial trees and models are obtained for each emotion cluster, the canonical parameters and emotion weights are updated iteratively, as shown in fig. 1.

3. Experimental results

3.1. Database

There are 4 speakers in the training data, two male and two female. Neutral data is available for all speakers. For one female and one male speaker, a subset of the sentences are recorded in 8 other styles: angry, happy, sad, surprised, fast, slow, informal and formal. The data has 9697 neutral and 11867 expressive sentences, of which 2378 are happy, 2249 sad, 1225 angry and the rest divided almost equally among the remaining expressions. The speech, sampled at 16kHz, is parameterized as 40 mel-cepstral coefficients, logF0, and 21 bark-scaled aperiodicity bands with Δ and Δ^2. The spectrum is obtained with a pitch synchronous analysis and the aperiodicity with PSHF [9].

Formal subjective experiments showed that emotion-dependent global variance models produce significantly more

[1] Assuming that neutral style data is available for the majority of speakers and that the renditions of neutral are sufficiently consistent.

[2] Emotion may be grouped automatically based on the acoustic features, on some subjective distance or, for acted emotions, on the data's tagging. An observed risk of the latter is that actors' rendition of non-neutral emotions vary greatly.

realistic expressions than emotion-independent ones, especially for 'happy' and 'sad' sentences. Therefore, emotion-dependent global variance models were used in all experiments.

3.2. CAT System configuration

The CAT model has 8 clusters: one bias cluster, 4 for modelling speakers and 3 for emotion. The bias and speaker clusters are initialized with neutral data from all speakers by assigning each speaker to one cluster. The expression clusters are initialized thus: one cluster is initialized with angry data; another with sad and slow; and the last with the remaining 5 styles. Neutral data is not assigned to any expression cluster during initialization as it has been used to initialize the bias and speaker clusters already. For each cluster three classes were defined: silence, pause and speech. For each class an independent set of weights and decision trees were created for spectrum, aperiodicity, lf0 and duration.

3.3. Baseline 1: Speaker-expression AVM

To evaluate the expressive speech synthesised for the speakers that have expressive training data, a speaker-expression average voice model (AVM) with CMLLR/CSMAPLLR transforms [10] is trained on the same data as the CAT model. In the AVM, each speaker/emotion combination is treated as an independent speaker. For a given component, speaker and emotion, the emission probability is

$$p(\boldsymbol{o}(t)\,|m, s, e, \mathcal{M}) = \\ |\boldsymbol{A}^{(e,s)}_{r(m)}|\mathcal{N}\left(\boldsymbol{A}^{(e,s)}_{r(m)}\boldsymbol{o}(t) + \boldsymbol{b}^{(e,s)}_{r(m)}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right) \qquad (13)$$

where $r(m) \in \{1, \ldots, R\}$ is the regression class of component m and $\{\boldsymbol{A}^{(e,s)}_{r(m)}, \boldsymbol{b}^{(e,s)}_{r(m)}\}$ are the set of CMLLR transforms. Block diagonal global transforms for the same 3 classes as per the CAT system are used in the initial stages of training. In the latter stages, 83 classes for the spectrum, 128 for lf0, 85 for aperiodicity and 17 for duration are defined using the model's decision tree pruned to the 8^{th} level. The CMLLR transforms are refined with CSMAPLR to produce $\{\hat{\boldsymbol{A}}^{(e,s)}_{r(m)}, \hat{\boldsymbol{b}}^{(e,s)}_{r(m)}\}$. Lastly, a MAP update of the means is applied:

$$\hat{\boldsymbol{\mu}}^{(s,e)}_m = \frac{\tau\boldsymbol{\mu}_m + \sum_{t\in t(e,s)} \gamma_m(t, s, e)\big(\hat{\boldsymbol{A}}^{(s,e)}_{r(m)}\boldsymbol{o}(t) + \hat{\boldsymbol{b}}^{(s,e)}_{r(m)}\big)}{\tau + \sum_{t\in t(e,s)} \gamma_m(t, s, e)} \qquad (14)$$

where $\boldsymbol{\mu}_m$ is the mean vector of the AVM, $\gamma_m(t, s, e)$ is the state occupancy probability, $t(e, s)$ the data associated with a speaker-expression pair and τ a hyperparameter. The final synthesis model is obtained by transforming the MAP adapted model with the CSMAPLR transforms.

3.4. Baseline 2: Prosody transplant

For some emotions, prosody (lf0, duration and energy) is known to be one of the most important factors. Simultaneously, speaker identity is strongly related to the spectrum. Therefore, a simple baseline to produce expressive speech from a neutral-only target speaker model is to "transplant" the prosody generated by an expressive model into the spectrum.[3] Emotion-dependent models (EDMs) are trained on a female speaker's expressive data and a *target* speaker model is trained on the neutral data of the other female speaker. To transplant prosody, the spectrum, aperiodicity and lf0 are generated from an EDM. The same sentence is generated by the target model with the phone durations forced to match those produced by the EDM. The spectrum generated by the target model is then modified to have the same energy trajectory as that from the EDM. Next, the lf0 generated by the EDM is shifted to match the difference between the average lf0 of the target speaker and the neutral data of the expressive speaker. Lastly, the target aperiodicity and modified spectrum are combined with the modified expressive lf0 to generate speech.

Emotion	AVM	CAT	p-score
Happy	50.5	49.5	0.444
Sad	60.1	39.9	0.001
Angry	23.9	76.1	$< 10^{-3}$
TOTAL	45.0	55.0	0.001

Table 1: ABX similarity to emotion for speaker/emotion pairs that exist in the training data.

E2 E1	happy	sad	angry
neutral	77.3 / 72.5	80.6 / 81.8	81.6 / 77.4
happy	-	90.5 / 88.6	72.6 / 61.6
sad	-	-	85.0 / 85.9

Table 2: Perceptual distance between emotions: 100=completely dissimilar, 50=completely similar. Left numbers correspond to real data samples and right ones to samples synthesized by the CAT model with weights for a "neutral-only" speaker and generic expressions.

3.5. Evaluation methodology

The evaluation method adopted in this work is based on ABX preference tests conducted via *CrowdFlower* using Amazon's Mechanical Turk workers located in the US [11]. Each test evaluated 25 sentences, each of them pronounced in three different expressions: 'happy', 'sad' and 'angry'. In all the tests, the references are real samples from the recorded corpora and the contrasted samples are from the two systems being evaluated. Subjects selected which of the two samples has an emotion closer to that in the reference. This method differs from the traditional one reported in expressive synthesis literature. The reason is that the primary interest is to model the expressions in the data. Considering the continuous nature of the emotion space [12], it makes more sense to ask subjects to evaluate the similarity in expressiveness of some synthetic speech to samples of real speech than to ask them to classify the perceived expression into some categories.

3.6. Results

The AVM may be directly compared with the CAT model using speaker/expression pairs that are present in the training data. An ABX test is run with the reference speaker different from the speaker of the evaluated models. The results are in table 1. Overall, the CAT model with speaker-independent expressions is significantly preferred over AVM, mainly due to a much better performance for angry sentences.

Next, the ability of the CAT system to produce expressive speech with the voice of a neutral-only target speaker is eval-

[3] An AVM factorization that cascades transforms $\boldsymbol{A}^{(e)}(\boldsymbol{A}^{(s)}\boldsymbol{o}(t) + \boldsymbol{b}^{(s)}) + \boldsymbol{b}^{(e)}$ is possible but is outside the scope of this paper.

Emotion	Prosody transp.	CAT Factoriz.	p-score
Happy	30.6	69.4	$< 10^{-3}$
Sad	42.2	57.8	0.027
Angry	24.2	75.8	$< 10^{-3}$
TOTAL	34.0	66.0	$< 10^{-3}$

Table 3: ABX test for similarity to target emotion. Prosody transplant vs CAT with speaker-independent emotion weights.

Emotion	Prosody transp.	CAT Factoriz.	p-score
Happy	63.9	36.1	$< 10^{-3}$
Sad	64.0	36.0	0.001
Angry	40.1	59.9	0.007
TOTAL	51.8	48.2	0.186

Table 4: ABX test for similarity to target speaker. Same samples as in table 3. References are in neutral style.

uated. The first experiment measures the subjective distances between emotions of real and synthetic speech. The subjective distance between two emotions E1 and E2 may be measured using an ABX test. The ABX reference, a real speaker talking in one of the emotions RE1, is compared with two samples uttered by a different speaker (either real or synthetic) SE1 and SE2. This is repeated for RE2. The distance between E1 and E2 is defined as the percentage of stimuli for which the sample with the same emotion as the reference was preferred. The results are in table 2. Compared with the 'real' expressions, those produced by the CAT system for the target speaker are slighly blurred but still acceptable. The strongest degradation is for 'angry' vs. 'happy'. This might be due to the very different way 'angry' was performed by the training speakers: one expresses it by shouting, the other did it with a crescendo of pitch and volume across the sentence, which is very hard to capture using frame-level models.

The second experiment compares the CAT model expression transplant with a basic prosody transplant. In this case the ABX reference was the same speaker as the speaker from which the EDMs for the prosody transplant had been built. In terms of expressiveness, the results (table 3) show a strong general preference for the CAT model. In terms of similarity to the target speaker (table 4), there are no significant differences, despite the prosody transplant model using the spectrum generated by a speaker-dependent model trained with the target speaker data.

Across all the experiments, the results for 'sad' are interesting. The expression of sadness by both training speakers is fundamentally prosodic. Both reduce the tempo and the pitch dynamic. However, whereas the female speaker produces a 'depressed sad' by lowering the pitch, the male speaker produces a 'weeping sad' by raising it. When these data are combined to create the speaker-independent sad weight, it results in a small increase of the pitch which is insufficient to produce the 'weeping sad' effect.

4. Discussion

In the CAT framework, weights for emotion and speaker are independent of each other. Therefore, differences between neutral style and other expressions are represented by a shift in the space of mean vectors as illustrated in fig. 2. An implication is that the same Δ could be ported to other neutral-style HMM-TTS models to produce that expression. More generally, given a neutral target speaker model *and* another able to produce different emotions (including neutral), it should be possible to produce speech of the target speaker voice in different emotions by interpolating the target speaker model with the difference between the models of the desired emotion and the neutral style.

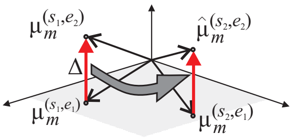

Figure 2: Shifting of the mean vector in the acoustic space.

5. Conclusions

A method to factorize speaker and expression was been proposed. Experiments show that this method produces a continuous variety of expressions even with voices from which only neutral-style data is available. Although the results are interesting, there is still space for improvement. Some supra-segmental models, such as the global variance, still contain important expressive information which was not properly modelled by the frame-level model. An improvement may be expected by combining the current frame-level CAT model with other supra-segmental models [13] also trained with a CAT factorization.

6. References

[1] M. Schroder, F. Burkhardt, and S. Krstulović, "Synthesis of emotional speech," in *Blueprint for Affective Computing*, K. Scherer, T. Banziger, and E. Roesch, Eds. Oxford University Press, 2010.

[2] J. Pitrelli, R. Bakis, E. Eide, R. Fernandez, W. Hamza, and P. M.A., "The IBM expressive text-to-speech synthesis system for American English," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, 2006.

[3] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Communication*, vol. 51, no. 3, 2009.

[4] O. Turk and M. Schroeder, "A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis," in *Proc Interspeech*, 2008.

[5] T. Nose, M. Tachibana, and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation," *IEICE Trans Inf. & Syst*, vol. E92, no. 3, 2009.

[6] 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.

[7] V. Wan, J. Latorre, K. 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.

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

[9] P. Jackson and C. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 7, 2001.

[10] 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 and Language Processing*, vol. 17, 2009.

[11] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Proc. Interspeech*, 2011.

[12] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011.

[13] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis," in *Proc Interspeech*, 2008.