INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A Hybrid System for Continuous Word-level Emphasis Modeling Based on HMM State Clustering and Adaptive Training

Quoc Truong Do[⋆], *Tomoki Toda*[†], *Graham Neubig*[⋆], *Sakriani Sakti*[⋆], *Satoshi Nakamura*[⋆]

[⋆] Nara Institute of Science and Technology, Japan
{do.truong.dj3,neubig,ssakti,s-nakamura}@is.naist.jp
[†] Nagoya University, Japan
tomoki@icts.nagoya-u.ac.jp

Abstract

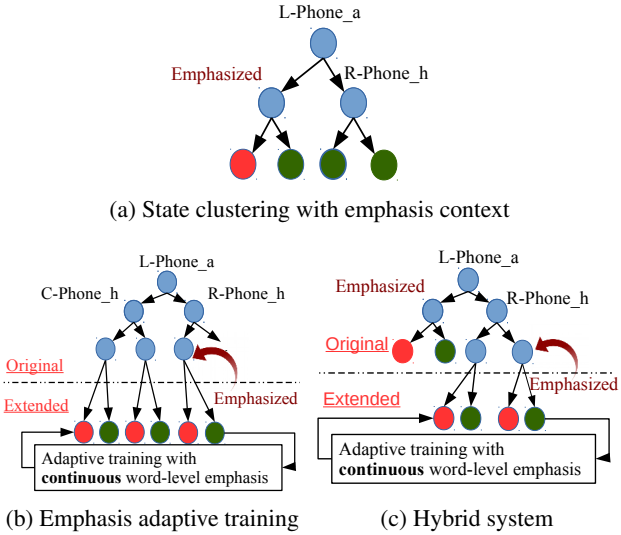
Emphasis is an important aspect of speech that conveys the focus of utterances, and modeling of this emphasis has been an active research field. Previous work has modeled emphasis using state clustering with an emphasis contextual factor indicating whether or not a word is emphasized. In addition, cluster adaptive training (CAT) makes it possible to directly optimize model parameters for clusters with different characteristics. In this paper, we first make a straightforward extension of CAT to emphasis adaptive training using continuous emphasis representations. We then compare it to state clustering, and propose a hybrid approach that combines both the emphasis contextual factor and adaptive training. Experiments demonstrated the effectiveness of adaptive training both stand-alone or combined with the state clustering approach (hybrid system) with it improving emphasis estimation by 2-5% F-measure and producing more natural audio.
Index Terms: Emphasized speech, word-level emphasis, continuous representation, emphasis adaptive training.

1. Introduction

Emphasis is an indispensable aspect of speech that conveys a variety of information including focus and emotion. For example, speakers often put more emphasis on particular words to help the listener understand which information in the sentence is the most important. Emphasis is also used to express emotion by putting more power, higher pitch, or longer duration on particular words [1] ("it is *REALLY* hot today"). There are many situations where emphasis can be applied, such as speech translation, where recent works have attempted to preserve emphasis at word level in translation, bringing opportunities to perceive emphasis from other languages [2, 3, 4]. All applications attempting to use emphasis rely on an emphasis modeling strategy, the most important component.

With regards to speech modeling techniques, hidden semi-Markov models (HSMMs) are a common approach in speech synthesis, and provide a data-driven framework and flexibility to model the different varieties of speech [5]. Previous work on word-level emphasis modeling based on HSMMs has relied on state clustering with emphasis contextual factors. A simple approach uses emphasis contextual factors indicating whether or not a word and its neighbor words are emphasized, and creates an *emphasis decision tree* [6] or a *factorized decision tree* [7] with some nodes having an emphasized question, as illustrated in Fig. 1 (a). While these methods are both expressive and effective, they have a disadvantage in that they make a hard zero-one distinction between unemphasized and emphasized words. However, in reality, emphasis is more subtle, and can be better represented using a continuous variable where a larger number indicates a higher level of emphasis.

Previous work [4] has proposed a model of continuous word-level emphasis using linear-regression HSMMs, which



(a) State clustering with emphasis context

(b) Emphasis adaptive training

(c) Hybrid system

Figure 1: Emphasis modeling techniques including (a) state clustering using contextual information, (b) emphasis adaptive training without emphasis context (b), and (c) a hybrid system that adopts both strategies.

are a simple form of multi-regression HSMMs [8] that effectively model varieties of speech with different characteristics (such as speed, speaking style or in this case, emphasis) by using a regression parameter to combine different Gaussian components. LR-HSMMs have the strong advantage of being able to model emphasis with continuous variables. However, at training time, these models still only use zero-one hard decisions about emphasis to optimize model parameters.

In this paper, we make two improvements to LR-HSMM-based emphasis modeling to solve this problem. First, to improve the parameter optimization process, we make an extension of cluster adaptive training (CAT) [9] to *emphasis adaptive training* as illustrated in Fig. 1 (b). Next, to take advantage of both parameter optimization and expressive decision tree modeling, we propose a *hybrid approach* that considers both the state clustering using the emphasis contextual factor and emphasis adaptive training, as illustrated in Fig. 1 (c).

2. Emphasis Modeling using HSMM State Clustering

State clustering is an approach that helps to reduce the number of HSMM states and the need for a large amount of training

http://dx.doi.org/10.21437/Interspeech.2016-930

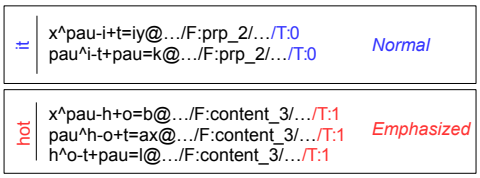


Figure 2: An example of full contextual labels for the word "it" and "hot" where the word "it" is normal and "hot" is emphasized. The context "T:*" is the additional emphasis factor, and the remaining items are traditional contextual information.

data by clustering HSMM states using some cluster criterion. This clustering is generally performed by using decision trees, which decide the cluster of HMM states based on a number of contextual factors. By simply adding an emphasis contextual factor to the cluster criterion, as illustrated in Fig. 2, we can model *normal* and *emphasized* HSMM states [6][1]. The decision tree constructed by having additional emphasis context (Fig. 1 (a)) can separate Gaussians components into *normal* and *emphasized* ones. Although this approach is simple and easy to implement, there are three problems: (1) it does not guarantee that the emphasis question appears in all paths starting from the root to leaf nodes, causing a problem that there are some nodes that make no distinction between emphasized and non-emphasized words; (2) it separates the training data into normal and emphasized parts, causing emphasized and normal nodes to only be trained with emphasized and normal data, respectively; and (3) emphasis is treated as a binary value indicating emphasized or not, and thus it is not possible to model emphasis at a "medium" level using continuous values.

3. Continuous Emphasis Modeling with LR-HSMMs and State Clustering

The state clustering approach described in the previous section can model zero-one emphasis. However, in reality, emphasis is more subtle. For example, one sentence might have two emphasized words with one having a smaller level of emphasis than the other. Therefore, it may be better to represent emphasis as a continuous variable where a larger number indicates a higher emphasis level. In this section, we describe continuous word-level emphasis modeling [4] using linear-regression hidden semi-Markov models (LR-HSMMs) [8] with HSMM state clustering.

3.1. LR-HSMM definition

We assume a word sequence consists of J words $\mathbf{w} = [w_1, \cdots, w_J]$, and a length T acoustic feature vector $\boldsymbol{o} = [\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T]$. As the observed feature vector $\boldsymbol{o}_t$ at time t, we use a combination of a spectral feature vector $o_t^{(1)}$ and F_0 feature vector $o_t^{(2)}$ as described in [10]. The likelihood function of the LR-HSMMs is given by

$$P(\boldsymbol{o}|\boldsymbol{\Lambda}, \mathcal{M}) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{\Lambda}, \mathcal{M}) P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\Lambda}, \mathcal{M}), \quad (1)$$

where $\boldsymbol{q} = [\boldsymbol{q}_1, \cdots, \boldsymbol{q}_T]$ is the HSMM state sequence, $\boldsymbol{\Lambda} = [\lambda_1, \cdots, \lambda_j, \cdots, \lambda_J]$ is the word-level emphasis sequence, and $\mathcal{M}$ is an HSMM parameter set. The LR-HSMM has two separate Gaussian components, normal and emphasized Gaussians, which are derived by using a decision tree constructed using HSMM state clustering, which described in the above section.

Because emphasis is defined at the word level, all linear-regression states that belong to one word will share the same emphasis level, as illustrated in Fig. 3. The state output probability density function modeled by a Gaussian distribution[2] is given by

$$P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\Lambda}, \mathcal{M}) = \prod_{t=1}^{T} P(\boldsymbol{o}_t|q_t, \omega_t, \mathcal{M}), \quad (2)$$

$$P(\boldsymbol{o}_t|q_t = i, \omega_t, \mathcal{M}) = \prod_{s=1}^{2} \mathcal{N}\left(\boldsymbol{o}_t^{(s)}; \boldsymbol{\mu}_{n,i}^{(s)} + \omega_t \boldsymbol{b}_i^{(s)}, \boldsymbol{\Sigma}_i^{(s)}\right), \quad (3)$$

where $\boldsymbol{\mu}_{n,i}^{(s)}$ is the normal Gaussian mean vector at state i and stream s, ω_t is frame-level emphasis equivalent to λ_j if state $i \in w_j$; and s is a stream index (*i.e.*, $s = 1$ for the spectral features and $s = 2$ for the F_0 features), and $\boldsymbol{b}_i^{(s)}$ a vector expressing the difference between the normal and emphasized Gaussian mean,

$$\boldsymbol{b}_i^{(s)} = \boldsymbol{\mu}_{e,i}^{(s)} - \boldsymbol{\mu}_{n,i}^{(s)}, \quad (4)$$

where $\boldsymbol{\mu}_{e,i}^{(s)}$ is the emphasized Gaussian mean vector. The duration probability $P(\boldsymbol{q}|\boldsymbol{\Lambda}, \mathcal{M})$ is also derived in a similar way to the state output probability,

$$P(\boldsymbol{q}|\boldsymbol{\lambda}, \mathcal{M}) = \prod_{i=1}^{N} P(d_i|\omega_i, \mathcal{M}), \quad (5)$$

$$P(d_i|\omega_i, \mathcal{M}) = \mathcal{N}\left(d_i; \mu_i^{(d)} + \omega_i b_i^{(d)}, {\sigma_i^{(d)}}^2\right), \quad (6)$$

where $\mu_i^{(d)}$ and $b_i^{(d)}$ are the normal Gaussian mean and the difference between emphasized and normal Gaussian means, respectively; d_i is an HSMM state duration, $\omega_i = \lambda_j$ if $d_i \in w_j$; and N is the number of states in the sentence HSMM sequence (i.e., the sum of d_i over N HSMM states is equivalent to T).

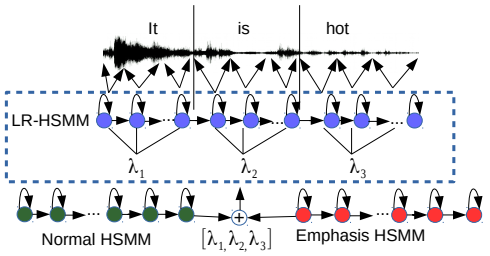


Figure 3: An example of linear-regression HSMMs. Each word has its own emphasis level λ_j, and all HMM states that belong to the same word will share the same emphasis level.

3.2. Word-level emphasis sequence estimation

Given an observation sequence $\boldsymbol{o}$, and its transcription, the process to estimate the word-level emphasis sequence is as follows [4]: first, an LR-HSMM is constructed by selecting the Gaussian distributions corresponding to the context of the given transcription. Then, emphasis is estimated by determining maximum likelihood estimates of the emphasis weight sequence, which is the same as the cluster weight estimation process in

[1] Of course, it is possible to use more contextual factors, i.e. indicating whether preceding and succeeding words are emphasized. However, in this work we omit these factors to maintain comparability of the evaluation with other approaches using the same context factors.

[2] Specifically, a multi-space probability distribution [11] is used for the F_0 component in this paper.

the cluster adaptive training (CAT) algorithm [9]. The word-level emphasis weight sequence is estimated by maximizing the HSMM likelihood as follows:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} P\left(\boldsymbol{o}|\boldsymbol{\lambda}, \mathcal{M}\right). \quad (7)$$

This maximization process is performed with the EM algorithm [12].

4. Emphasis Adaptive Training

First, we make an extension of CAT [9] to allow it to perform emphasis adaptive training. The idea of the proposed method is to iteratively estimate and update the word-level emphasis sequences and model parameters, respectively.

Given the estimated word-level emphasis sequence $\boldsymbol{\Lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_J]$, we want to find the model parameters that maximize the likelihood function

$$\hat{\mathcal{M}} = \operatorname*{argmax}_{\mathcal{M}} P\left(\boldsymbol{o}|\boldsymbol{\lambda}, \mathcal{M}\right). \quad (8)$$

The maximization process is performed with the EM algorithm as follows:

1. Use the existing model parameters to estimate word-level emphasis sequences $\hat{\boldsymbol{\Lambda}}$ as described in the previous section. In other words, this step automatically generates pseudo-labels for the training data.
2. Update the mean of the normal Gaussian component at stream s, $\boldsymbol{\mu}^{(s)}_{n,m}$, and duration state d, $\mu^{(d)}_{n,m}$, given estimated word-level emphasis sequences $\hat{\boldsymbol{\Lambda}}$. Below is the formula used to update $\boldsymbol{\mu}^{(s)}_{n,m}$.

$$\boldsymbol{\mu}^{(s)}_{n,m} = \boldsymbol{G}^{-1}\boldsymbol{K} \quad (9)$$

$$\boldsymbol{G} = \sum_{m' \in m} [(1-\omega^{(m')}) \sum_t \gamma_t^{(m')} \boldsymbol{o}_t - \omega^{(m')}(1-\omega^{(m')}) \sum_t \gamma_t^{(m')} \boldsymbol{\mu}_e^{(m)}], \quad (10)$$

$$\boldsymbol{K} = \sum_{m' \in m} (1-\omega^{(m')})^2 \sum_t \gamma_t^{(m')}, \quad (11)$$

where m' is the untied model of linear Gaussian component m, $\omega^{(m')}$ is Gaussian-level emphasis that is equivalent to λ_j if the untied model m' belongs to the word w_j. The mean of emphasis Gaussian $\boldsymbol{\mu}^{(s)}_{e,m}$ and duration model $\mu^{(d)}_{n,m}$ can be updated in a similar way.

3. Go back to step 1 until the model is converged.

Note that in this paper, the covariance matrices of Gaussian components are kept unchanged for simplification.

Based on emphasis adaptive training, we propose an approach to model emphasis without the need of state clustering with emphasis context as illustrated in Fig. 1 (b). In this approach, the decision tree is constructed without any emphasis context, as illustrated in Fig. 1 (b) – Original. After that, the original leaf nodes are turned into intermediate nodes by adding an emphasis question splitting each of them into normal and emphasized nodes (Fig. 1 (b) – Extended). At this point, the emphasized and normal leaf nodes are equivalent. Then, to ensure that the parameters of emphasized and normal Gaussians are different, we add to the emphasized Gaussians a mean difference vector $\bar{\boldsymbol{b}}^{(s)}$, which is calculated based on the tree created from the state clustering approach. Finally, we adopt emphasis adaptive training described above to further optimize the parameters.

Unlike the previous approach where emphasized and normal Gaussians are trained only on emphasized or normal speech respectively, emphasis-adaptive-training-based approaches are able to utilize all the training data to train the model parameters. When training on emphasized samples, the emphasized Gaussian components get more weight (emphasis level) than normal Gaussians, and vice versa.

However, the simple approach described here also has a weakness in that it forces emphasis questions to always be asked right before the leaf nodes. This has the potential to result in sub-optimal decision tree structure. We resolve this weakness in the following section.

5. Hybrid System for Continuous Emphasis Modeling

Next, we propose a hybrid approach that takes advantage of both of the above approaches. First, a decision tree (the original tree) with emphasis questions asked at some intermediate nodes is constructed as in Section 2. Then, we extend leaf nodes that belong to paths that do not have an emphasis question asked in any of the intermediate nodes as shown in Algorithm 1.

Algorithm 1 State splitting algorithm.

```
1:  procedure StateSplitting(s)
2:      if s has emphasis question then
3:          return
4:      else
5:          if s is leaf node. then
6:              SET s as intermediate node.
7:              ADD emphasis question to s.
8:              SPLIT s into 2 leaf nodes.
9:              return
10:         else
11:             StateSplitting(left node of s).
12:             StateSplitting(right node of s).
```

The state splitting process 6-8 will duplicate the mean and covariance matrix of Gaussian components of the state being split. After splitting the tree, every leaf node is guaranteed to represent either a normal or emphasized Gaussians. Then, to ensure that emphasized and normal Gaussians are different, the same procedure as the previous section is applied.

Finally, we perform emphasis adaptive training with continuous emphasis representations to further optimize model parameters for the nodes split by the line 8 of Algorithm 1.

6. Experiments

6.1. Experimental setup

In this section, we evaluate the performance of emphasized speech modeling using state clustering, emphasis adaptive training, and the hybrid approach. The experiments were conducted using a bilingual English-Japanese emphasized speech corpus [13], which has emphasized content words that were carefully selected to maintain the naturalness of emphasized utterances. The corpus consists of 966 pairs of utterances that were spoken by 3 bilingual speakers, 6 monolingual Japanese, and 1 monolingual English speaker. In the experiments, we selected 2 speakers for each language with 916 utterances for training and 50 utterances for testing. Thus, we have 100 testing samples in total for each language. The LR-HSMM model was trained for each speaker separately, resulting in 4 models in total. The speech features were extracted using 31 mel-cepstral coefficients including 25 dimension spectral parameters, 1 dimension log-scaled F_0, and 5 dimension aperiodic features. Each speech parameter vector includes static features and their delta

and delta-deltas. The frame shift was set to 5 ms. Each HSMM model is modeled by 7 HMM states including initial and final states. We adopt STRAIGHT [14] for speech analysis.

With regards to emphasis adaptive training, we performed the adaptive training for the first 6 iterations, then re-estimate word-level emphasis sequences. These are then used to perform emphasis adaptive training until the model converges.

6.2. Word-level emphasis prediction evaluation

In the first experiment, we evaluate the performance of the different models in emphasis prediction, where we are given an input speech signal and would like to predict whether each word is emphasized. For each system, we estimate the word-level emphasis sequences for the testing data, then classify them to normal and emphasized labels using a threshold of 0.5. Then, we calculate the F-measure for all systems. The result is shown in Fig. 4.

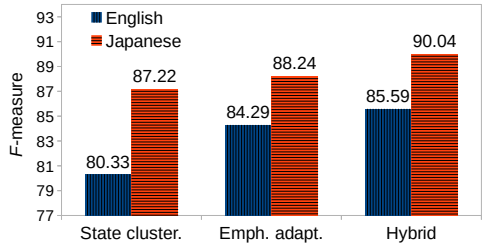


Figure 4: Emphasis prediction accuracy.

As we can see, the model using emphasis adaptive training and the hybrid approach outperform the state clustering approach in both languages by 2-5% F-measure[3]. One possible reason for this is that in state clustering approaches, emphasis questions do not appear at all paths starting from the root to leaf nodes, leading to some emphasized words having weak emphasis levels. To test this hypothesis, we perform an analysis showing the percentage of the number of decision tree traversing without asking for emphasized questions in the state clustering approach for both languages. The result shown in Fig. 5 indicates that many times we traverse through the tree to derive emphasized and normal Gaussian components, emphasis questions are not asked at in all three acoustic feature streams for 11.37% times in English, and no emphasis question is asked more than half the time in at least one of the feature streams. On the other hand, the proposed approach guarantees that we can always derive different emphasized and normal Gaussian components. This is also one explanation for why the improvement of the hybrid system compared to other methods is larger in English than Japanese.

6.3. Naturalness evaluation

In the next experiment, we use the models to synthesize speech of the Japanese data and perform a preference test evaluation to evaluate the naturalness of the synthetic speech. 50 utterances in the testing data were synthesized with each system using the ground-truth emphasis labels (e.g., "it is *really* hot today" with emphasis label "0 0 1 0 0"). 7 Japanese native listeners performed a pairwise evaluation over all pairs of systems.

As shown in Fig. 6, the hybrid approach generated more natural audio compared to all others. We hypothesize the reasons are as follows:

[3]We did not carry out subjective evaluation explicitly, however, our previous work [4] has shown that the human emphasis prediction has about a 4% reduction of F-measure compared to objective evaluation due to the lack of pauses in the synthetic speech.

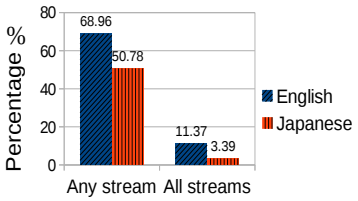


Figure 5: The percentage of decision tree traversing without asking for emphasized questions in the state clustering approach. "All streams" and "Any stream" indicate situations in which emphasis questions are not asked in all feature streams (lf0, duration, and spectral) or any of them, respectively.

- **State clustering**: The *emphasized* and *normal* Gaussians are trained using only *emphasized* or *normal* speech, respectively. Although emphasis questions are placed in the decision tree according to the likelihood function, due to the limitation of training data, some Gaussian components do not get a sufficient amount of training samples, leading to low quality synthetic speech.
- **Adaptive training**: In this approach, we can utilize all training data to train Gaussian components. When training on emphasized speech samples, the *emphasized* Gaussian components get more weight (emphasis level) than *normal* Gaussians, and vice versa, leading to higher quality than the state clustering approach. However, the emphasis questions are forced to be asked right before the leaf node (not according to likelihood function), this potentially makes the audio become unnatural.
- **Hybrid approach**: This approach inherits advantages from both above systems. The decision tree has emphasis questions are placed according to likelihood function, some paths that do not have emphasis questions are refined using state splitting, and the model parameters are further optimized using adaptive training.

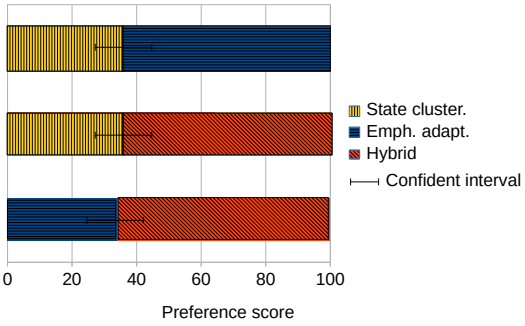


Figure 6: Preference score of synthetic speech for each system.

7. Conclusion

In this paper, we have proposed methods for emphasis adaptive training and a hybrid system that combine state clustering and adaptive training approaches. Experiments showed that the proposed model outperforms other methods by 2-5% F-measure of emphasis estimation accuracy, and produces more natural audio. Future work will incorporate emphasis adaptive training with more sophisticated clustering such as factorized decision trees, and MLLR adaptation.

8. References

[1] H. Fujisaki, "Information, prosody, and modeling - with emphasis on tonal features of speech," in *Proceedings of Speech Prosody*, 2004, pp. 1–10.

[2] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Generalizing continuous-space translation of paralinguistic information," in *Proceedings of Interspeech*, August 2013.

[3] T. Kano, S. Sakti, S. Takamichi, G. Neubig, T. Toda, and S. Nakamura, "A method for translation of paralinguistic information," in *Proceedings of IWSLT*, December 2012, pp. 158–163.

[4] Q. T. Do, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs," in *Proceedings of InterSpeech*, September 2015.

[5] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE*, vol. 88, no. 11, pp. 2484–2491, 2005.

[6] K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Emphasized speech synthesis based on hidden Markov models," in *Proceedings of COCOSDA*, August 2009, pp. 76–81.

[7] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proceedings of ICASSP*, 2010, pp. 4238–4241.

[8] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE*, vol. E90-D, no. 9, pp. 1406–1413, September 2007.

[9] M.J.F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE*, vol. 8, no. 4, pp. 417–428, 2000.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proceedings of Eurospeech 1999*, 1999.

[11] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *The royal statistical society*, vol. 39, no. 1, pp. 1–38, 1977.

[13] Q. T. Do, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Collection and analysis of a japanese-english emphasized speech corpus," in *Proceedings of COCOSDA*, Phuket, Thailand, September 2014.

[14] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 34, pp. 187 – 207, 1999.