Preserving Word-level Emphasis in Speech-to-speech Translation using Linear Regression HSMMs

Quoc Truong Do, Shinnosuke Takamichi, Sakriani Sakti,
Graham Neubig, Tomoki Toda, Satoshi Nakamura

Graduate School of Information Science
Nara Institute of Science and Technology, Japan
{do.truong.dj3,shinnosuke-t,ssakti,neubig,tomoki,s-nakamura}@is.naist.jp

Abstract

In speech, emphasis is an important type of paralinguistic information that helps convey the focus of an utterance, new information, and emotion. If emphasis can be incorporated into a speech-to-speech (S2S) translation system, it will be possible to convey this information across the language barrier. However, previous related work focuses only on the translation of particular prosodic features, such as F_0, or works with emphasis but focuses on extremely small vocabularies, such as the 10 digits. In this paper, we describe a new S2S method that is able to translate the emphasis across languages and consider multiple features of emphasis such as power, F_0, and duration over larger vocabularies. We do so by introducing two new components: word-level emphasis estimation using linear regression hidden semi-Markov models, and emphasis translation that translates the word-level emphasis to the target language with conditional random fields. The text-to-speech synthesis system is also modified to be able to synthesize emphasized speech. The result shows that our system can translate the emphasis correctly with 91.6% F-measure for objective test, and 87.8% for subjective test.

Index Terms: speech translation, paralinguistic translation, emphasis estimation, emphasis translation

1. Introduction

Modern speech-to-speech (S2S) translation systems [1] have greatly improved in accuracy, and computer-aided. Communication across the language barrier is moving closer to reality. However, most S2S systems can not translate paralinguistic information such as emphasis or emotion, and as a result, communication though S2S systems is more dry and less emotionally engaging than standard speech communication. If it were possible to translate paralinguistic information along with the content, communication though S2S translation could be a much more fulfilling experience. Among the various types of paralinguistic information, in our work we focus on emphasis, which plays an important role in conveying the key words of utterances to make communication smoother.

In speech, emphasis is manifested by changing the duration, power, or F_0 [2]. The challenge in developing an S2S system that can accurately translate emphasis is that we must consider these acoustic features in three components: emphasis extraction, emphasis translation, and synthesis of emphasized speech. Previous work [3] proposed a binary emphasis detection method to find emphasized parts in a speech. However, this work uses only F_0 patterns to detect emphasis with a binary value, it also was strictly mono-lingual. [4] proposed a method to model the word-level emphasis in HMM-based TTS using factorized decision trees, but there is no emphasis estimation or translation. In [5, 6], duration and power are extracted directly from speech input and then translated to the target language by a mapping function. However, this simple method was applied to only very small vocabularies, specifically the 10 digits. As a result, it cannot generalize to unseen words, and has difficulty in modeling emphasis in large vocabulary systems. In [7], a method has been proposed to translate F_0 patterns across languages, but other acoustic parameters such as duration, power, or spectrum that are related to emphasis have not being investigated.

In our work, we take one step further to construct an S2S translation system that conveys emphasis with a much larger vocabulary size than tackled before. The idea is based on the conventional S2S framework, with the incorporation of an additional component to estimate an emphasis level for each word in an utterance by applying linear-regression hidden semi-Markov models (LR-HSMMs), which are a simple form of multi-regression hidden semi-Markov models (MR-HSMMs) [8]. We choose LR-HSMMs for our S2S model for two reasons. First, it allows us to build a single model for both emphasis estimation and synthesis of emphasized speech. Second, it is appropriate for tasks with words that do not exist in the training data, because it allows us to model speech at the phoneme level. By utilizing LR-HSMMs, we can estimate a real-numbered value of emphasis at the word-level. Then, the sequence of word-level emphasis levels is translated to the target language by an emphasis translation model using conditional random fields (CRFs) [9], which allows us flexibly integrate different features to the emphasis translation model. Finally, the text-to-speech system synthesizes emphasized speech using text and the corresponding emphasis sequence.

2. Word-level Emphasis Modeling

In this section, we describe the use of linear regression hidden semi-Markov models (LR-HSMMs) in modeling word-level emphasis.

2.1. LR-HSMM definition

We adopt LR-HSMMs, which are a simple form of MR-HSMMs [8] to model the emphasis speech as follows. We assume a word sequence consists of J words $\mathbf{w} = [w_1, \cdots, w_j, \cdots, w_J]$, and a length T vector sequence of acoustic features of the input utterance $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_t^\top \cdots, \boldsymbol{o}_T^\top\right]^\top$. As the observation feature vector $\boldsymbol{o}_t$ at frame t we use a combination of the spectral feature vector $\boldsymbol{o}_t^{(1)}$ and the F_0 feature vector $\boldsymbol{o}_t^{(2)}$ as described in [10]. The likelihood function of the LR-HSMMs is given by

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

where $\boldsymbol{q} = \{q_1, \cdots, q_t, \cdots, q_T\}$ is the HSMM state sequence, $\boldsymbol{\lambda} = \{\lambda_1, \cdots, \lambda_j, \cdots, \lambda_J\}$ is the word-level emphasis weight sequence, and $\mathcal{M}$ is an HSMM parameter set. Note that in this paper, the emphasis weight is shared over all HSMM states corresponding to a word as shown in Fig. 1. The state output probability density function modeled by a Gaussian distribution[1] is given by

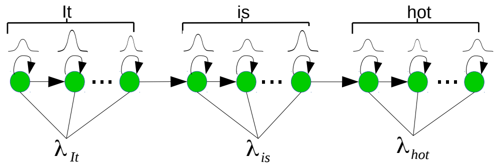

Figure 1: Word-level emphasis.

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

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

where ω_t is frame-level emphasis equivalent to λ_j, where j is the word corresponding to frame t, and s is a stream index (*i.e.*, $s = 1$ for the spectral feature and $s = 2$ for the F_0 features). At HSMM state i for the s^{th} stream, the mean vector is given by a linear combination of the vector $\boldsymbol{\mu}_i^{(s)}$ for normal speech and the vector $\boldsymbol{b}_i^{(s)}$ expressing the difference between normal speech and emphasized speech using ω_t as a weighting value, and the covariance matrix is $\boldsymbol{\Sigma}_i^{(s)}$. Moreover, the duration probability is given by

$$\begin{aligned} P(\boldsymbol{q}|\boldsymbol{\lambda}, \mathcal{M}) &= \prod_{i=1}^{N} P(d_i|\omega_i, \mathcal{M}), && (4) \\ P(d_i|\omega_i, \mathcal{M}) &= \mathcal{N}\left(d_i; \mu_i^{(d)} + \omega_i b_i^{(d)}, {\sigma_i^{(d)}}^2\right), && (5) \end{aligned}$$

where $\{d_1, \cdots, d_i, \cdots, d_N\}$ is a set of HSMM state durations corresponding to $\boldsymbol{q}$, $\omega_i = \lambda_j$ 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). At HSMM state i, the mean of the Gaussian distribution is also given by a linear combination of the mean value $\mu_i^{(d)}$ for normal speech and the value $b_i^{(d)}$ expressing the difference between the normal speech and emphasized speech using ω_i as a weighting value, and the variance is given by ${\sigma_i^{(d)}}^2$.

2.2. Training of the LR-HSMM

The training process mainly follows the standard HMM-based speech synthesis training process [12, 13, 14]. First, the training data is labeled with full contextual factors encoding various features of the sentence. To model emphasis, we use an additional contextual factor encoding the word-level emphasis by adding an emphasis question to the standard question set to cluster context-dependent phoneme HSMM states in each cluster [14]. Because of limitations in training data, we adopt decision-tree-based state tying [15, 16].

Using this decision tree, we can partition the set of Gaussian components into 2 groups, one is normal, and the other is emphasized Gaussians. Finally, the mean vectors of normal Gaussians are set to $\boldsymbol{\mu}_i^{(s)}$ and $\mu_i^{(d)}$, and the difference mean vectors between normal and emphasized Gaussians are set to $\boldsymbol{b}_i^{(s)}$ and $b_i^{(d)}$ so that the mean vectors of the LR-HSMMs are equal to those of emphasized Gaussians if the emphasis weight ω_i is set to 1. The covariance matrices and variances of the LR-HSMMs are set to those of normal Gaussians.

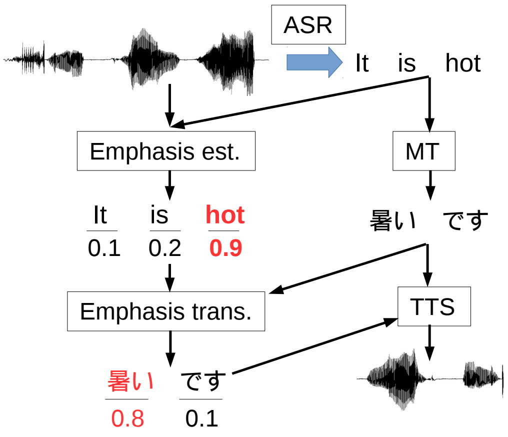

Figure 2: Emphasis speech-to-speech translation

3. Word-level Emphasized Speech Translation

Our proposed emphasized speech translation model consists of a conventional S2S system, an emphasis estimation model, and an emphasis translation model as illustrated in Fig. 2. In this paper, we focus on the emphasis modeling and translation. Therefore, we assume that the ASR and MT systems provide correct transcriptions and translation outputs.

3.1. Emphasis weight sequence estimation

Given an observation sequence $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_t^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top$, and its transcription, the process to estimate the emphasis weight sequence is as follows: First, an LR-HSMM is constructed by selecting the Gaussian distributions corresponding to 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 adaptation process in the cluster adaptive training (CAT) algorithm [17]. The word-level emphasis weight sequence is estimated by maximizing the HSMM likelihood as follows:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{o}|\boldsymbol{\lambda}, \mathcal{M}). \quad (6)$$

This maximization process is performed with the EM algorithm [18]. In the E-step, posterior probabilities are calculated as follows:

$$\begin{aligned} \gamma_{i,t}^{(s)} &= P(q_t = i|\boldsymbol{o}, \boldsymbol{\lambda}, \mathcal{M}), && (7) \\ \gamma_{i,t}^{(d)} &= P(d_i = t|\boldsymbol{o}, \boldsymbol{\lambda}, \mathcal{M}). && (8) \end{aligned}$$

Then, in the M-step, the maximum likelihood estimate of the word-level emphasis weight sequence

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

$\hat{\boldsymbol{\lambda}} = \left\{\hat{\lambda}_1, \cdots, \hat{\lambda}_j, \cdots, \hat{\lambda}_J\right\}$ is determined as

$$\hat{\lambda}_j \quad = \quad g_j^{-1} k_j, \tag{9}$$

where g_j and k_j are calculated by

$$\begin{aligned} g_j \quad &= \quad \sum_{i \in q(j)} \left[\sum_{s=1}^{2} \sum_{t=1}^{T} \gamma_{i,t}^{(s)} \boldsymbol{b}_i^{(s)\top} \boldsymbol{\Sigma}_i^{(s)-1} \boldsymbol{b}_i^{(s)} \right. \\ &\qquad \left. + \sum_{t=1}^{T} \gamma_{i,t}^{(d)} b_i^{(d)2} \sigma_i^{(d)-2} \right], \end{aligned} \tag{10}$$

$$\begin{aligned} k_j \quad &= \quad \sum_{i \in q(j)} \left[\sum_{s=1}^{2} \boldsymbol{b}_i^{(s)\top} \boldsymbol{\Sigma}_i^{(s)-1} \sum_{t=1}^{T} \gamma_{i,t}^{(s)} \left(\boldsymbol{o}_t^{(s)} - \boldsymbol{\mu}_i^{(s)} \right) \right. \\ &\qquad \left. + b_i^{(d)} \sigma_i^{(d)-2} \sum_{t=1}^{T} \gamma_{i,t}^{(d)} \left(d_t - \mu_i^{(d)} \right) \right], \end{aligned} \tag{11}$$

where $q(j)$ indicates a set of HSMM states corresponding to word w_j. It should be noted that in this framework we can easily control the effect of individual acoustic features on emphasis estimation by selecting Gaussian components used in the M-step as shown in Eqs. 10 and 11.

3.2. Emphasis translation with conditional random fields

Our next step is emphasis translation, or to take word-level emphasis estimates in the source languages $\hat{\boldsymbol{\lambda}}^{(f)}$, and convert them to emphasis estimates in the target language $\hat{\boldsymbol{\lambda}}^{(e)}$. We perform estimation of emphasis in the target language using conditional random fields (CRFs) [9], a standard method for discriminative sequential prediction. To train CRFs to predict target side emphasis, we create training data consisting of source and target words $\mathbf{w}^{(f)}$ and $\mathbf{w}^{(e)}$, and the corresponding estimated emphasis values. As $\hat{\boldsymbol{\lambda}}^{(e)}$ is a sequence of continuous values, and CRFs requires discrete state sequences, we first quantize $\hat{\boldsymbol{\lambda}}^{(f)}$ and $\hat{\boldsymbol{\lambda}}^{(e)}$ into buckets, giving us a discrete sequence $\hat{\boldsymbol{\lambda}}^{(f)'}$ and $\hat{\boldsymbol{\lambda}}^{(e)'}$. We then create CRFs training data that consists of N samples $D = [(\mathbf{x}_1, \lambda_1^{(e)'}), \cdots, (\mathbf{x}_n, \lambda_n^{(e)'}), \cdots, (\mathbf{x}_N, \lambda_N^{(e)'})]$, where $\mathbf{x}_n$ is a feature vector for each word in $w_n^{(e)}$ consisting of:

- source word-level emphasis $\lambda_j^{(f)}$, and its context,
- source word $w_j^{(f)}$, and word context,
- source word part of speech (PoS) $pos(w_j^{(f)})$, and PoS context,
- target word $w_n^{(e)}$, and word context,
- target word PoS $pos(w_n^{(e)})$, and PoS context,

where context means the information of one succeeding and one preceding words.

To decide which source features correspond to a target word $w_n^{(e)}$, we use one-to-one word alignments between $w_j^{(f)}$ and $w_n^{(e)}$. The likelihood of CRFs is given by

$$P(\boldsymbol{\lambda}^{(e)'}|\mathbf{x}) = \frac{\prod_{n=1}^{N} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(\lambda_{n-1}^{(e)'}, \lambda_n^{(e)'}, \mathbf{x}_n^{(k)})\right\}}{\sum_{\tilde{\boldsymbol{\lambda}}^{(e)'}} \prod_{n=1}^{N} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(\tilde{\lambda}_{n-1}^{(e)'}, \tilde{\lambda}_n^{(e)'}, \mathbf{x}_n^{(k)})\right\}}, \tag{12}$$

where θ_k is the weight parameter. The feature function f_k combines the word-level emphasis bi-gram $\lambda_{n-1}^{(e)'}, \lambda_n^{(e)'}$, and the input feature $\mathbf{x}_n^{(k)}$. The CRFs model parameters are optimized by maximizing the likelihood function in Equation (12) using Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [19] as implemented in CRFSuite [20].

3.3. Speech synthesis

As described in Section 2, the TTS system relies on LR-HSMMs to synthesize the emphasized speech. Unlike the word based model used in previous works [5, 6], our system uses a phoneme based model, allowing the proposed method to work with larger vocabularies.

The output speech parameter vector sequence $\boldsymbol{o}^{(e)}$ is determined by maximizing the likelihood function given the state sequence $\boldsymbol{q} = [q_1, \cdots, q_T]$, the word-level emphasis sequence $\hat{\boldsymbol{\lambda}}^{(e)'}$, and the HSMM model set $\mathcal{M}$

$$\hat{\boldsymbol{o}}^{(e)} = \underset{\boldsymbol{o}^{(e)}}{\operatorname{argmax}} P(\boldsymbol{W}\boldsymbol{o}^{(e)}|\boldsymbol{q}, \hat{\boldsymbol{\lambda}}^{(e)'}, \mathcal{M}), \tag{13}$$

where $\boldsymbol{W}$ is the weighting matrix for calculating the dynamic features [21]. We also adopt the Global Variance method [22] to alleviate the over-smoothness of the generated parameters. The STRAIGHT analysis-synthesis system [23] was employed for parameter extraction and waveform generation. The feature vector consists of spectral and excitation parameters and their delta and delta-delta features. Five-state left-to-right HSMMs were used.

4. Experiments

4.1. Experimental Setup

Experiments were performed using a bilingual English-Japanese emphasis corpus [24]. The corpus consists of 1015 parallel utterances of English and Japanese. In each language, at least one of the content words in the sentence is emphasized. In our experiments, we use the data from 2 speakers, a native English speaker, and a Japanese native speaker as the training and testing data. After filtering out long sentences over 10 words, we obtained 966 utterances, which we divided into 916 sentences with 1,186 emphasized words for training and 50 sentences with 62 emphasized words for testing. The word-level emphasis is quantized to the closest of {0, 0.3, 0.6, 0.9} unless stated otherwise, which we describe in detail in Section 4.3.

To measure the accuracy, we calculate emphasis F-measure, the harmonic mean of the precision and recall with which the system detects emphasis.

4.2. Emphasis translation evaluation

In our first experiment, we evaluate the ability of the proposed method to reproduce emphasis in the target language. In addition, we also evaluate the effect of the combination of input features described in Section 3.2 to find out which features gives the highest F-measure.

The translated word-level emphasis is classified into binary values (1 and 0) using a threshold 0.5. We have evaluated different thresholds, and found out that 0.5 is the best value to classify emphasized and normal words. We also use a baseline "All Emphasis" that predicts that every word is emphasized. The result is shown in Table 1.

Comparing the first row with the others, we can see that the emphasis prediction model outperforms the chance rate by approximately 40%. This indicates that our proposed system can produce emphasis in the target language relatively accurately. Next, we can see that the model that use the features including the source language information (3rd-7th row) is better than

Table 1: F-measure for different combinations of input features. *e_en* and *e_ja* denote word-level emphasis, *w_en* and *w_ja* denote word information, and *t_en* and *t_ja* denote PoS tag of English, and Japanese, respectively.

Feature type	***F*-measure** (%)
All Emphasis	42.5
w_ja, *t_ja*, *t_ja_c*, *e_ja*	81.6
e_en, *e_ja*	82.8
e_en, *e_ja*, *e_en_c*	82.8
e_en, *e_ja*, *w_en*, *w_ja*	84.8
e_en, *e_ja*, *w_en*, *w_ja*, *t_en*, *t_ja*	90.0
e_en, *e_ja*, *w_en*, *w_ja*, *t_en*, *t_ja*, *t_ja_c*	**91.6**

Table 2: F-measure for different quantization methods.

System	***F*-measure** (%)
0/1 Quant	85.5
0.1 Quant	90.8
0.3 Quant	**91.6**
Labels	90.7

the model use only target information (2nd row). This demonstrates that our model is effectively translating emphasis from the source, as opposed to simply predicting based on the target.

Looking at the third and fourth rows, we can see that the emphasis context in the source language does not help the word-level emphasis translation, indicating that word-level emphasis in the target language depends mainly on the emphasis of the corresponding source word. By adding the word information in both languages, the accuracy increased by 2%, and further increased when adding the PoS tag information by approximately 6%. Finally, we add the context of the PoS tags in Japanese, yielding the best system with 91.6% accuracy. This is consistent with the characteristic of the corpus that content words are usually emphasized. We also tested with other combinations of the features, but none of them gave the accuracy higher than 91.6%. Overall, the result indicates that along with acoustic features (emphasis level), the linguistic features such as word, PoS tag are also contributing to the improvement of the translation model.

4.3. Word-level emphasis quantization evaluation

Next, we examine the effect of word-level emphasis quantization to the translation model. When creating the CRFs model, we use 4 different quantization schemes.

0/1 Quant: The word-level emphasis is quantized into the closest of 1 and 0

0.3 Quant: The word-level emphasis is quantized into the closet of {0, 0.3, 0.6, 0.9}.

0.1 Quant: The word-level emphasis is quantized into bucket of 0.1

Labels: The word-level emphasis in source language is quantized according to "0.3 Quant" and the emphasis in target language is derived from the labels from the corpus and has binary values, 1 for emphasis, 0 for normal.

From the result, we can see that the quantization scheme "0.3 Quant" gives the best result, likely because it provides an appropriate amount of training data. And more importantly, it even outperforms the manually created "Labels," suggesting that training the system using quantized word-level emphasis can be more effective than binary values.

4.4. Manual evaluation

In the final experiment, we performed a manual evaluation to determine how well the end-to-end system can translate emphasis. We asked to 6 native Japanese speakers to listen to the emphasis translated utterances, and select the words that they think are emphasized in 150 randomized testing utterances from the following 3 systems.

Baseline: No emphasis translation is performed. The TTS is trained using a normal decision tree.

CRF-based: Emphasis is translated from English to Japanese using the CRF model, which is trained using the best features in Table 1.

Natural: Natural speech spoken speech by a Japanese speaker.

Fig. 3 shows the accuracy for all 3 systems. We can see that the proposed emphasis translation model achieves a large improvement over the baseline system by 11.8% F-measure. The audio generated by the baseline system have many words that are randomly emphasized, because there is no emphasis control based on the source utterance.

Comparing these results with the automatic evaluation of Section 4.3, we can still see a gap of approximately 4% between the results. This is likely due to problems of speech synthesis. When listening to the natural and synthetic audio, we found that there are often pauses inserted in natural speech in order to emphasize words, which the synthetic audio does not have. This problem can be addressed by introducing a pause prediction model in the target language.

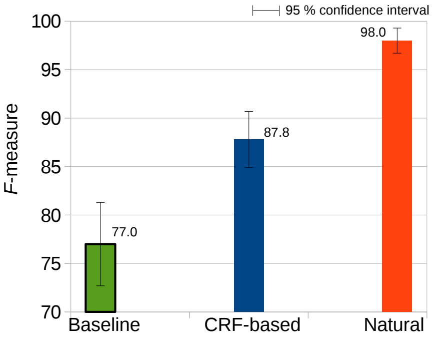

Figure 3: Emphasis prediction F-measure for manual evaluation

5. Conclusion

In this paper, we present a new speech translation method that is able to translate the emphasis across languages. Compared to previous works, our proposed method can works with larger vocabularies, and consider many factors of emphasis such as duration, power, and F_0. Although the experiment data is not really big, the method can works with unseen words without any modifications. Future work will improve the emphasized speech synthesis by adding a pause prediction model, improve the emphasis translation, and include an investigation of ASR and MT errors.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 24240032 and by the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

7. References

[1] S. Nakamura, “Overcoming the language barrier with speech translation technology,” *Science & Technology Trends - Quarterly Review No.31*, April 2009.

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

[3] B. Arons, “Pitch-based emphasis detection for segmenting speech recordings,” in *Proceedings of ICSLP*, 1994, pp. 1931–1934.

[4] K. Yu, F. Mairesse, and S. Young, “Word-level emphasis modelling in HMM-based speech synthesis,” in *Processing of ICASSP*, March 2010, pp. 4238–4241.

[5] T. Kano, S. Sakti, S. Takamichi, G. Neubig, T. Toda, and S. Nakamura, “A method for translation of paralinguistic information,” in *Proceedings of IWSLT*, 2012.

[6] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, “Generalizing continuous-space translation of paralinguistic information.” in *Proceedings of INTERSPEECH*, 2013, pp. 2614–2618.

[7] P. D. Aguero, J. Adell, and A. Bonafonte, “Prosody generation for speech-to-speech translation,” in *Proceedings of ICASSP*, vol. 1, 2006.

[8] H. Zen, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” *IEICE*, vol. E90-D, no. 5, pp. 825 – 834, 2007.

[9] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proceedings of ICML*, 2001, pp. 282–289.

[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*. ISCA, 1999.

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

[12] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[13] K. Yu, H. Zen, F. Mairesse, and S. Young, “Context adaptive training with factorized decision trees for HMM-based statistical parametric speech synthesis,” *Speech Communication*, vol. 53, no. 6, pp. 914 – 923, 2011.

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

[15] S. J. Young, J. J. Odell, and P. C. Woodland, “Tree-based state tying for high accuracy acoustic modelling,” in *Proceedings of Human Language Technology Workshop*, 1994, pp. 307–312.

[16] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” *ASJ (E)*, vol. 21, pp. 79–86, 2000.

[17] M. Gales, “Cluster adaptive training of hidden Markov models,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2000.

[18] 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.

[19] J. Nocedal, “Updating Quasi-Newton Matrices with Limited Storage,” *Mathematics of Computation*, vol. 35, no. 151, pp. 773–782, 1980.

[20] N. Okazaki, “Crfsuite: a fast implementation of conditional random fields (crfs),” 2007. [Online]. Available: http://www.chokkan.org/software/crfsuite/

[21] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proceedings of ICASSP*, vol. 3, 2000, pp. 1315–1318 vol.3.

[22] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis,” *IEICE*, pp. 816 – 824, 2007.

[23] H. Kawahara, I. Masuda-Katsuse, and A. 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.

[24] D. Q. Truong, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, “Collection and analysis of a Japanese-English emphasized speech corpus,” in *Proceedings of Oriental COCOSDA*, September 2014.