Study on Word-Level Emphasis Across English and Japanese *

◎ Do Quoc Truong, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig,
Tomoki Toda, Satoshi Nakamura (NAIST)

1 Introduction

Speech-to-speech (S2S) translation systems [1] allow us to communicate with other people in different languages. However, conventional S2S systems ignore the emphasis, which is an important factor of paralinguistic information. In order to construct such a S2S system, it is important to study on how emphasis is expressed in each language and also across languages. In this paper, the emphasis is analyzed in word-level. We first estimate a real-numbered value of word-level emphasis using many speech features such as F_0, duration and power. Based on this estimation we perform an analysis of how emphasis is expressed in individual language and also across languages.

2 Word-level emphasis modeling and estimation with LR-HSMM

In this section, we describe about the use of linear-regression hidden semi-Markov models (LR-HSMM) [2] in modeling and estimating word-level emphasis.

2.1 LR-HSMM

The observation feature vector sequence for each sentence that consists of W words is given by $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \cdots, \boldsymbol{o}_t^\top, \cdots, \boldsymbol{o}_T^\top\right]^\top$, where the observation feature vector $\boldsymbol{o}_t$ at frame t consists of the spectral feature vector $\boldsymbol{o}_t^{(1)}$ and the F_0 feature vector $\boldsymbol{o}_t^{(2)}$ in this paper. The number of frames in the sequence is T. The likelihood function of the LR-HSMM 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, $\boldsymbol{\lambda} = \{\lambda_1, \cdots, \lambda_w, \cdots, \lambda_W\}$ 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. The state output probability density function modeled by a Gaussian distribution is given by

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

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

where s is a stream index (*i.e.*, $s = 1$ for the spectral features and $s = 2$ for the F_0 feature), $\boldsymbol{\mu}_i^{(s)}$ is the mean vector for normal speech and $\boldsymbol{b}_i^{(s)}$ is the difference vector between normal speech and emphasized speech using λ_w as a weighting value, and the covariance matrix is $\boldsymbol{\Sigma}_i^{(s)}$. The duration probability function is also derived in the same fashion with the state output probability function. The emphasis sequence $\boldsymbol{\lambda}$ is estimated by maximizing the likelihood function 1.

The LR-HSMMs model is trained by following the standard HMM-based speech synthesis training process [3]. To model the emphasis and normal speech, we use an additional contextual factor encoding the word-level emphasis, i.e., "current word is emphasized?", as also used in [4].

3 Experiment set-up and evaluation

The experiments were performed using a bilingual English-Japanese emphasis corpus [5], in which the emphasized words were carefully selected to maintain the naturalness of the emphasized utterances. The corpus consists of 966 pairs of utterances that were spoken by 3 bilingual speakers; 6 mono-lingual Japanese and 1 mono-lingual English speakers. The LR-HSMMs were trained using 916 utterances for each speaker. All 966 emphasis sequences were used for the analysis. We adopt STRAIGHT [6] for the speech analysis.

3.1 Word-level Emphasis Estimation Evaluation

First, we validate that the proposed method is able to detect emphasis and find which acoustic features (spectral, F0, duration) are more useful to estimate the emphasis or distinguish between the emphasized and normal words. We do so by optimizing the emphasis weight sequence using different settings of acoustic features:

- **dur**: using only the duration feature.
- **lf0**: using only log $F0$ (lf0) feature.
- **sp**: using only spectral features.
- **sp_dur**: using spectral and duration features.
- **sp_lf0**: using spectral and lf0 features.
- **lf0_dur**: using lf0 and duration features.
- **sp_lf0_dur**: combine all features.

* 日英対訳コーパスにおける強調音声の分析

The word-level emphasis is then classified into labels of 0 and 1 indicating normal and emphasized words by using emphasis threshold 0.5. Then, we calculate the F-measure to show how accurate the system can detect the emphasis. The result is shown in Figure 1. First, by looking at the duration column, the duration feature works not so bad in English, but does not work well in Japanese. We can also observed this situation when combine lf0 and duration. This problem caused by the characteristic Japanese where people can not use long duration to emphasise words. Looking at the performance of individual feature, we can see that for English all three feature duration, F_0, and spectral play the same role in term of emphasis prediction. However, for Japanese, the spectral feature is more significant compared to the other two.

By combining all features together. We archived the best performance for both languages. The F-measure for English is 75.63% and Japanese is 80.36%. Therefore, we will use this combination for the emphasis translation experiments.

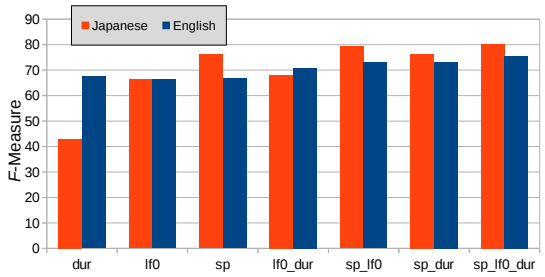

Fig. 1 F-measure of emphasis prediction

3.2 Emphasis across English and Japanese

In this experiment, we analyse the correlation of emphasis weights in English and Japanese. The Pearson correlation coefficient is calculated between the English and Japanese emphasis weight to measure the strength of the linear association between them

$$r = \frac{\sum_i (\lambda_i^{(en)} - \bar{\lambda}^{(en)})(\lambda_i^{(ja)} - \bar{\lambda}^{(ja)})}{\sqrt{\sum_i (\lambda_i^{(en)} - \bar{\lambda}^{(en)})^2}\sqrt{\sum_i (\lambda_i^{(ja)} - \bar{\lambda}^{(ja)})^2}}. \quad (4)$$

where r is the Pearson correlation coefficient, $\lambda_i^{(en)}$ is the emphasis level for the word i-th in English and $\lambda_i^{(ja)}$ is the emphasis level for the corresponding Japanese word which is determined by one-to-one word alignment, $\bar{\lambda}^{(en)}$ and $\bar{\lambda}^{(ja)}$ is the mean of emphasis level of English and Japanese, respectively.

As the result, the correlation of the word-level emphasis between English and Japanese is shown in Figure 2, the Pearson correlation coefficient is 0.625, indicating that there is some correlation coefficient of emphasis level between two languages, but it is not so high. It might possible to develop a emphasis translation by using a linear function. Of course, a more sophisticated method could be used to make a better generalization.

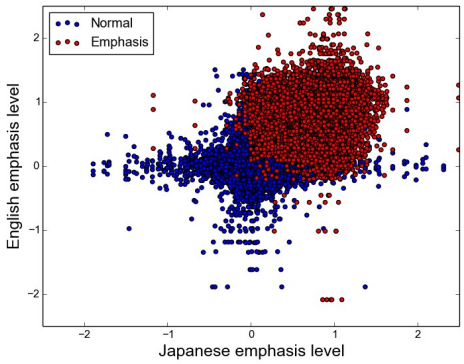

Fig. 2 Relationship between English words and Japanese words emphasis.

4 Conclusion

This paper has conducted a bilingual analysis of emphasis. We found that the combination of all speech features outperforms other combination in term of emphasis detection. The analysis of emphasis across languages found a relatively high correlation of emphasis vectors between English and Japanese. The future works will construct an emphasis speech translation that translates the word-level emphasis vectors across languages.

5 Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 24240032.

参考文献

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

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

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

[4] 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.

[5] 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.

[6] 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.