This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

Preserving Word-level Emphasis in Speech-to-speech Translation

Quoc Truong Do, *Nonmember, IEEE,* Tomoki Toda, *Member, IEEE,* Graham Neubig, *Member, IEEE,* Sakriani Sakti, *Member, IEEE,* and Satoshi Nakamura, *Member, IEEE*

***Abstract*—Speech-to-speech translation (S2ST) is a technology that translates speech across languages, which can remove barriers in cross-lingual communication. In conventional S2ST systems, the linguistic meaning of speech was translated, but paralinguistic information conveying other features of the speech such as emotion or emphasis were ignored. In this paper, we propose a method to translate paralinguistic information, specifically focusing on emphasis. The method consists of a series of components that can accurately translate emphasis using all acoustic features of speech. First, linear-regression hidden semi-Markov models (LR-HSMMs) are used to estimate a real-numbered emphasis value for every word in an utterance, resulting in a sequence of values for the utterance. After that, the emphasis translation module translates the estimated emphasis sequence into a target language emphasis sequence using a conditional random field (CRF) model considering the features of emphasis levels, words, and part-of-speech tags. Finally, the speech synthesis module synthesizes emphasized speech with LR-HSMMs, taking into account the translated emphasis sequence and transcription. The results indicate that our translation model can translate emphasis information, correctly emphasizing words in the target language with 91.6% F-measure by objective evaluation. A listening test with human subjects further showed that they could identify the emphasized words with 87.8% F-measure, and that the naturalness of the audio was preserved.**

***Index Terms*—Emphasis estimation, word-level emphasis, intent, emphasis translation, speech-to-speech translation.**

I. Introduction

SPEECH is one of the richest and most powerful communication channels used by mankind. It allows the speaker to express not only the content that they want to convey, but also paralinguistic information such as emotion and emphasis. This paralinguistic information is useful in a broad variety of situations, just one example of which is shown in Fig. 1. In this example, the listener has misheard some of the information provided by the speaker, and the speaker repeats the information, intentionally emphasizing the misheard words, making it possible for the listener to fully understand what was missed before. The example is simple, but it demonstrates the complexities in human communication, where both linguistic and paralinguistic information can be transferred in a single utterance. This communication is even more complex in cross-lingual situations because of differences in languages and cultures.

For many years, scientists have tried developing automatic S2ST translation systems [1] that help bring communication across the language barrier closer to reality. A S2ST system consists of 3 main components, automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS). However, most S2ST systems cannot translate paralinguistic information such as emphasis or emotion, and as a result, communication though traditional S2ST systems is less engaging than natural speech communication. If it were possible to translate paralinguistic information along with the content, communication though S2ST translation could be a more fulfilling experience.

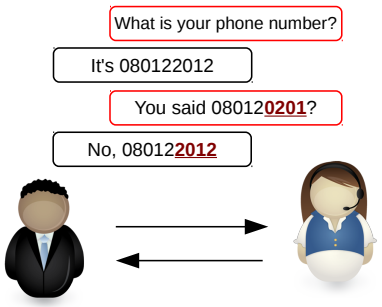

Fig. 1. An example of human communication where emphasis plays a crucial role in conveying intention of speakers (emphasized parts are written in bold with underline).

Among the various types of paralinguistic information, in our work we focus on emphasis, which plays a crucial role in conveying the keywords or focus of utterances. We follow Tsiartas et al. [2] in defining emphasis as the perceived loudness of a word or phrase. This emphasis is often used to distinguish between what is the focus and not focus part of an utterance [3] (Fig. 1), speakers can also express their emotions, with more emotional voices often employing more emphasis than neural voices. In this work, emphasis is taken into account as the factor that is intentionally expressed by speakers to convey the focus of utterances.

The difficulty in handling emphasis is that it can be manifested by changing different types of the acoustic features such as the duration, power, or F_0 of the emphasized words [4]. The challenge in developing an S2ST system that can accurately translate emphasis is that we must consider these acoustic features of emphasis in three components: emphasis extraction, emphasis translation, and synthesis of emphasized speech.

There are several works that have tried to address emphasis in individual components or throughout the whole S2ST translation pipeline. However, they are either limited domain or do not considers all the acoustic features of emphasis.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

Arons [5] proposed a binary emphasis detection method to find emphasized words in speech. However, this work uses only F_0 patterns to detect emphasis with a binary value, and also was strictly monolingual. Yu et al. [3] proposed a method to model word-level emphasis in hidden Markov model (HMM)-based TTS using factorized decision trees, but there is no emphasis estimation or translation involved. Later on, Tsiartas et al. [2] found that there is a relation between emphasis transfer and speech translation quality by conducting a study on multi-lingual speech corpora. After that, the authors presented an approach to map acoustic features into a discrete set of units in an attempt to translate emphasis across languages [6]. Although there was no real emphasis translation was performed, the paper introduced a potentially useful idea of representing emphasis as discrete values so that it is easier to map emphasis between languages. The works of Kano et al. [7,8] translate emphasis in a limited domain, the 10 digits. The methods model speech differently for each word in the vocabulary, and therefore cannot generalize to unseen words, and also have difficulties in modeling emphasis in large vocabulary systems. Anumanchipalli et al. [9] and Aguero et al. [10] proposed approaches to translate F_0 patterns across languages, but other acoustic parameters such as duration, power, or spectrum that are related to emphasis have not been investigated.

In our work, we take one step further to construct an S2ST translation system that conveys emphasis in an open vocabulary task and considers all the acoustic features of emphasis. To do so, we create models that handle emphasis in each of the three components of the S2ST translation framework that allow both the handling of all acoustic features, and can model large vocabularies (Section III). First, in the ASR phase, we estimate a real-numbered value for emphasis for each word in an utterance by applying linear-regression hidden semi-Markov models (LR-HSMMs). LR-HSMMs are a simple form of multi-regression hidden semi-Markov models (MR-HSMMs) [11], and work by automatically estimating an interpolation coefficient between multiple HSMM models (in our case, models for emphasized or non-emphasized words). Then in the MT component, the sequence of word-level emphasis levels is translated to the target language by an emphasis translation model using conditional random fields (CRFs) [12], which allows us flexibly integrate different features in the emphasis translation model. Finally, the text-to-speech system uses LR-HSMMs to synthesize emphasized speech using text and the corresponding emphasis sequence[1].

Additionally, we also construct a bilingual English-Japanese emphasized speech corpus, in which the speaker expresses emphasis intentionally (Section IV)[2].

II. Conventional Speech-to-speech Translation

The conventional S2ST pipeline [1] is illustrated in Figure 2. In an S2ST translation system, first the ASR module transcribes audio from a source language into a transcription, which is then translated by the MT system into a target language sentence. This is finally synthesized into target language speech by the TTS module. In the following subsections, we give a short description for each component, and define terminology and formulas that we will use in describing our emphasis translation model.

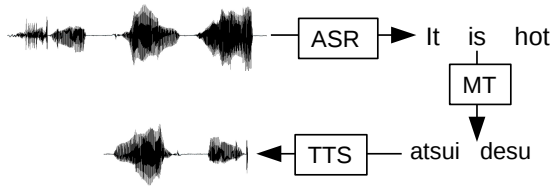

Fig. 2. Work-flow of a conventional S2ST system

A. Automatic Speech Recognition (ASR)

ASR aims to convert the speech signal into the corresponding word sequence. The first step of speech recognition takes the speech signal $\boldsymbol{x}$ and extracts speech features $\boldsymbol{o}$, such as mel-frequency cepstral coefficients (MFCCs). Given these features, ASR predicts the most plausible word sequence $\boldsymbol{w}$ that maximizes the conditional probability

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} P(\boldsymbol{w}|\boldsymbol{o}). \quad (1)$$

In this paper, we adopt DNN-HMM ASR, implemented in the Kaldi speech recognition tookit [15].

B. Statistical Machine Translation (MT)

The MT system lies in the middle of the S2ST system, and has a job to translate the hypothesis from the ASR module to a particular target language sentence. There are many methods that can be applied to MT task such as phrase-based [16], tree-based [17], and neural network [18] translation models.

Given a source language sentence $\boldsymbol{w}^{(f)}$, the MT system finds the highest probability target language sentence $\boldsymbol{w}^{(e)}$ as follows,

$$\hat{\boldsymbol{w}}^{(e)} = \underset{\boldsymbol{w}^{(e)}}{\operatorname{argmax}} P(\boldsymbol{w}^{(e)}|\boldsymbol{w}^{(f)}). \quad (2)$$

In the case of phrase-based models, which we use in this paper, this probability is calculated using a log-linear model with features including language model, translation model, and reordering model probabilities.

[1]It also may be possible to conceive of a joint approach that merges all three components together and trains them using a single objective function. However, because standard S2ST systems are still based on the 3-step approach, devising a method for end-to-end training is not trivial. In addition joint optimization also requires a large amount of parallel speech data, which may not be simple to collect. Therefore, we also model our emphasis translation method after the 3-step approach used in standard S2ST systems.

[2]Parts of this work have been presented in [13,14]. The work here provides a more comprehensive and systematic description of the method, presents a deeper analyses of the collected corpus, and conducts additional experiments on emphasis detection, as well as on translating emphasis while taking into account errors from ASR and MT systems.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

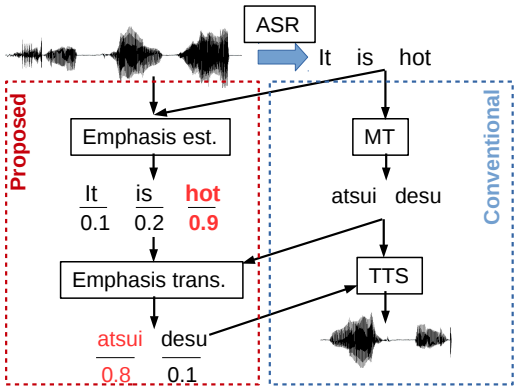

Fig. 3. An illustration of an emphasis S2ST system.

C. Text-to-speech Synthesis (TTS)

Text-to-speech is the last component in the S2ST system that synthesizes the target audio given the translated hypothesis. This paper adopts an HSMM-based TTS model. The reason is not only to inherit the advantages of the HSMM-based method, but also the flexibility to modify it to model the emphasized speech described in subsection III-A.

In this framework, the output speech parameter vector sequence $\boldsymbol{v}$ is determined by maximizing the likelihood function given the state sequence consists of T states $\boldsymbol{q} = [q_1, \cdots, q_T]$, and the HSMM model set $\mathbf{M}$

$$\hat{\boldsymbol{v}} = \underset{\boldsymbol{v}}{\operatorname{argmax}} P(\boldsymbol{W}\boldsymbol{v}|\boldsymbol{q}, \mathbf{M}), \tag{3}$$

where $\boldsymbol{W}$ is the weighting matrix for calculating the dynamic features [19].

III. Proposed Method for Preserving Word-level Emphasis

In speech, emphasis is manifested by changing the duration, power, or F_0 [20]. The challenge in developing an S2ST 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. Figure 3 illustrates the whole framework that this study proposes. Emphasis modeling and estimation utilize linear-regression hidden semi-Markov models (LR-HSMMs), which are a simple form of multi-regression hidden semi-Markov models (MR-HSMMs) [21]. We adopt conditional random fields (CRFs) [12] to translate the estimated emphasis sequence from a source language utterance into the emphasis sequence in the target language.

A. Emphasis Modeling

Previous work [21] proposed the MR-HSMM approach to control speaking styles in speech synthesis systems, where the styles are controlled by a parameter vector called a style control vector. The approach has shown not only promising results in generating emotional speech and various speaking styles, but also provides a flexible way to control the effect of individual acoustic features. In this work we adopt LR-HSMMs, which are a specific version of MR-HSMMs, that control a single scalar value instead of a vector of real values. As the scalar number that we control, we choose a scalar emphasis level for each word. LR-HSMMs give us both the ability to control the emphasis level of words and an easy way to analyse and control the effect of individual acoustic features such as duration, power, and fundamental frequency (F_0) for both emphasis estimation and synthesis. It is crucial to know these effects in order to better understand how emphasis is expressed in individual languages and across languages. LR-HSMMs also allow us to build a single model for both emphasis estimation and synthesis of emphasized speech. More importantly, they are appropriate for tasks with words that do not exist in the training data, because they allow us to model speech at the phoneme level.

1) Linear-regression hidden semi-Markov model (LR-HSMM): We use LR-HSMMs to model the emphasized 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 [22]. The likelihood function of the LR-HSMMs is given by

$$P(\boldsymbol{o}|\boldsymbol{\lambda}, \mathcal{M}) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{\lambda}, \mathcal{M}) P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}, \mathcal{M}), \tag{4}$$

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. 4. The state

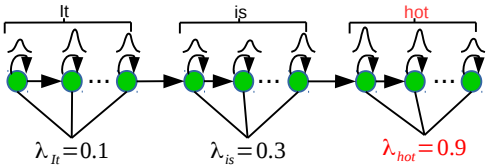

Fig. 4. An example of word-level emphasis (λ). Each word has its own emphasis level, and the emphasis level of one word is shared among all the HMM states associated to that word. In this example, the word "hot" is emphasized, so it has higher emphasis level than the other words.

output probability density function is modeled by a Gaussian distribution[3] as follows

$$P(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}, \mathcal{M}) = \prod_{t=1}^{T} P(\boldsymbol{o}_t|q_t, \omega_t, \mathcal{M}), \tag{5}$$

$$P(\boldsymbol{o}_t|q_t = i, \omega_t, \mathcal{M}) = \prod_{s=1}^{2} \mathcal{N}\left(\boldsymbol{o}_t^{(s)}; \boldsymbol{\mu}_i^{(s)} + \omega_t \boldsymbol{b}_i^{(s)}, \boldsymbol{\Sigma}_i^{(s)}\right), \tag{6}$$

where ω_t is frame-level emphasis equivalent to λ_j, j is the word corresponding to frame t, and s is a stream index (*i.e.*,

[3]Specifically, because F_0 features are discontinuous, so they are modeled by multi-space probability distributions [23].

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

$s = 1$ for the spectral features 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. The covariance matrix is $\boldsymbol{\Sigma}_i^{(s)}$. Moreover, the duration probability is given by

$$P\left(\boldsymbol{q}|\boldsymbol{\lambda},\mathcal{M}\right) = \prod_{i=1}^{N} P\left(d_i|\omega_i,\mathcal{M}\right), \quad (7)$$

$$P\left(d_i|\omega_i,\mathcal{M}\right) = \mathcal{N}\left(d_i;\mu_i^{(d)}+\omega_i b_i^{(d)},{\sigma_i^{(d)}}^2\right), \quad (8)$$

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) LR-HSMM Training: The training process mainly follows the standard HMM-based speech synthesis training process [24–26]. 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 [26] (An example is shown in Fig. 5).

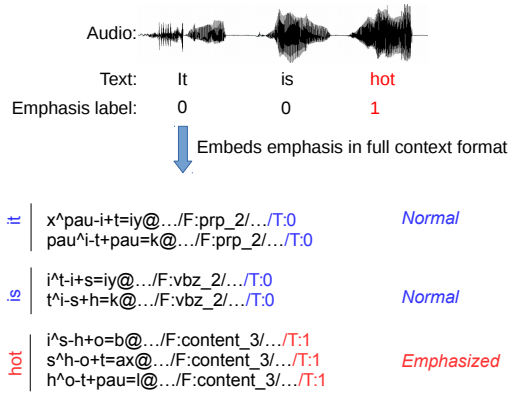

Fig. 5. An example of training samples consists of the audio signal, text, and emphasis labels. In this example, the word "hot" is the emphasized word, therefore it has an emphasis label of 1. In order to perform the model training, text and emphasis labels are converted into full contextual labels where emphasis labels are embedded at the end of the context.

In the HSMM acoustic modeling technique, each full contextual label is modeled by two HMM-GMM models for duration features and MFCC+F_0+aperiodic features. The challenge is that the number of labels is usually huge, so in practice Gaussian models are tied together using decision-tree-based state tying [27,28] technique to reduce the number of Gaussian models. This technique is very effective and suitable for small amounts of training data.

By adding the emphasis context to the full contextual label (Fig. 5), the decision tree can also learn to partition the leaf nodes (Gaussians) into two groups of normal and emphasized Gaussians as illustrated in Figure 6.

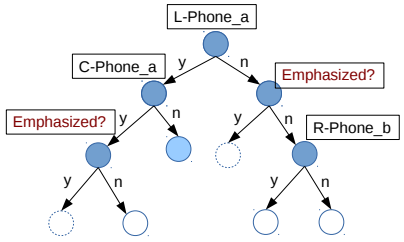

Fig. 6. An example of a decision tree with emphasis questions. The white dashed nodes are emphasized nodes and white solid nodes are normal nodes.

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.

3) Emphasis 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 the context of the given transcription. Then, emphasis is estimated by determining maximum likelihood estimates of the emphasis weight sequence. This can be done using the adaptation process in the cluster adaptive training (CAT) algorithm [29]. In the CAT algorithm, given a set of trained Gaussian clusters and new speaker data, it estimates interpolation parameters between the clusters in the way that maximizes the probability of the data given the model. To adopt it to estimate emphasis weights, we treat emphasized and normal Gaussians as the "Gaussian clusters" and emphasis weights are the interpolation parameters.

The word-level emphasis weight sequence is then estimated by maximizing the HSMM likelihood as follows:

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} P\left(\boldsymbol{o}|\boldsymbol{\lambda},\mathcal{M}\right). \quad (9)$$

This maximization process is performed with the EM algorithm [30]. In the E-step, posterior probabilities are calculated as follows:

$$\gamma_{i,t}^{(s)} = P(q_t = i|\boldsymbol{o},\boldsymbol{\lambda},\mathcal{M}), \quad (10)$$

$$\gamma_{i,t}^{(d)} = P(d_i = t|\boldsymbol{o},\boldsymbol{\lambda},\mathcal{M}). \quad (11)$$

Then, in the M-step, the maximum likelihood estimate of the word-level emphasis weight sequence $\hat{\boldsymbol{\lambda}} = \left\{\hat{\lambda}_1,\cdots,\hat{\lambda}_j,\cdots,\hat{\lambda}_J\right\}$ is determined as

$$\hat{\lambda}_j = g_j^{-1}k_j, \quad (12)$$

where g_j and k_j are calculated by

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

$$g_j = \sum_{i \in q(j)} \left[\sum_{s=1}^{2} \sum_{t=1}^{T} \gamma_{i,t}^{(s)} \boldsymbol{b}_i^{(s)\top} \boldsymbol{\Sigma}_i^{(s)-1} \boldsymbol{b}_i^{(s)} + \sum_{t=1}^{T} \gamma_{i,t}^{(d)} b_i^{(d)2} \sigma_i^{(d)-2} \right], \quad (13)$$

$$k_j = \sum_{i \in q(j)} \left[\sum_{s=1}^{2} \boldsymbol{b}_i^{(s)\top} \boldsymbol{\Sigma}_i^{(s)-1} \sum_{t=1}^{T} \gamma_{i,t}^{(s)} \left(\boldsymbol{o}_t^{(s)} - \boldsymbol{\mu}_i^{(s)} \right) + b_i^{(d)} \sigma_i^{(d)-2} \sum_{t=1}^{T} \gamma_{i,t}^{(d)} \left(d_t - \mu_i^{(d)} \right) \right], \quad (14)$$

where $q(j)$ indicates a set of HSMM states corresponding to word w_j, and s is a stream index (i.e., $s = 1$ for spectral (power) features and $s = 2$ for F_0 features).

Looking at the equations (13) and (14), we can interpret the first part of the summation as the model of F_0 and power features, and the second part as the model of duration features. Because they are all independent, we can easily control the effect of individual acoustic features on emphasis estimation just by removing or adding them from or to the calculation.

B. Emphasis Translation

Emphasis translation is the task of translating the estimated emphasis sequence for the source language to an emphasis sequence in the target language. It can also be viewed as a sequence labeling task where we want to label a sequence of target language emphasis weights given source language information. One conceivable approach is directly map emphasis values from the source language to the target language using word alignments[4]. However, similarly to emphasis estimation, in which many acoustic features affect the performance, we can hypothesize that emphasis will be expressed differently in different languages, and thus emphasis translation depend on not only emphasis weights, but also other linguistic information such as words and part-of-speech (PoS) tags. To capture these interlingual nuances, it is important to have a translation model that is suitable for sequence labeling tasks and can handle many different types of features together.

Lafferty et al. [12] proposed the Conditional Random Field (CRF) approach for sequence labeling and showed that it outperforms conventional HMM-based approaches for part-of-speech tagging problems. Gregory et al. [31] extends the use of CRFs from text to speech for pitch accent prediction in conversational speech and showed promising results. This work is related to emphasis because pitch is one feature that changes when a word is emphasized.

The major advantage of CRFs comes from the fact that they allow for the flexible incorporation of many different features into the model. This is advantageous for the emphasis translation task because we can easily take into account many features including words, part-of-speech tags, and their context units into the translation model.

[4]Word alignments capture information of corresponding word pairs and are used to extract cross-language features.

1) Conditional Random Fields: The linear-chain CRF can be depicted as an undirected graph in Figure 7. θ and μ correspond to the transition probability and emission probabilities in the hidden Markov model. Given a training data that consists

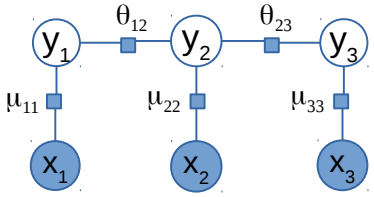

Fig. 7. The linear-chain conditional random field with the model parameters $\{\theta, \mu\}$.

of T samples $D = [(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_t, y_t), \cdots, (\mathbf{x}_T, y_T)]$, the conditional probability is calculated as,

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp \left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}, \quad (15)$$

where f is a feature function, K is number of feature functions, and $Z(\mathbf{x})$ is normalization function

$$Z(\mathbf{x}) = \sum_{\mathbf{y}'} \prod_{t=1}^{T} \exp \left\{ \sum_{k=1}^{K} \theta_k f_k(y'_t, y'_{t-1}, \mathbf{x}_t) \right\} \quad (16)$$

The model parameters θ are optimized by maximizing the conditional probability

$$\mathcal{L}(\theta) = P(\mathbf{y}|\mathbf{x}). \quad (17)$$

2) Emphasis Translation with Conditional Random Fields: We use CRFs to perform emphasis translation, or to take word-level emphasis estimates in the source language $\hat{\boldsymbol{\lambda}}^{(f)}$, and convert them to emphasis estimates in the target language $\hat{\boldsymbol{\lambda}}^{(e)}$.

Given word alignments derived from machine translation systems and a set of input features $\mathbf{X}$ including source language and target language features, the emphasis translation procedure works as follows. First, for each word in the target language sentence, we extract a subset of input features $\mathbf{X}$ corresponding to the word. Then, the input $\mathbf{X}$ is fed into CRFs to produce the target emphasis level. For the target words that do not align with any words in the source language, their emphasis levels are simply set to 0. Basically, CRFs run on each word, but they can also capture emphasis dependencies among neighbor words by using context units. Depending on the task and the amount of data, the model can perform better or worse when using more input features. Details about feature extraction are described as below.

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 (e.g., 0.473231 $\rightarrow$ 0.5), giving us a discrete sequence $\hat{\boldsymbol{\lambda}}^{(f)'}$ and $\hat{\boldsymbol{\lambda}}^{(e)'}$. Various quantization schemes are evaluated in experiments. We then create CRF training data that consists of N samples $D =$

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

$[(\mathbf{x}_1, \lambda_1^{(e)'}), \cdots, (\mathbf{x}_n, \lambda_n^{(e)'}), \cdots, (\mathbf{x}_N, \lambda_N^{(e)'})]$, 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 word. To decide which source features correspond to a target word $w_n^{(e)}$, we use word alignments between $w_j^{(f)}$ and $w_n^{(e)}$, as illustrated in Fig. 8.

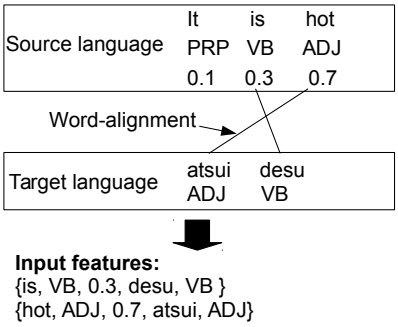

Fig. 8. An example of CRFs features for emphasis prediction task. Note that contextual information is not shown in the figure for simplification.

The CRF model parameters are optimized using limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm implemented in the CRFSuite toolkit [32].

IV. Corpus Collection

A. Construction of Emphasized Speech Corpora

In this subsection, we describe the creation of emphasized utterances from a well-known speech corpus BTEC [33]. BTEC is the Basic Travel Expression Corpus, a Japanese-English corpus covering a wide variety of content in the travel domain (samples are shown in Table I). Using this corpus as a basis for our recording material reduces the burden of corpus construction, as we can just choose appropriate sentences from the sentences in the original corpus. We focused on BTEC because they have the parallel sentences for English and Japanese and contain relatively short utterances, allowing for easier analysis. We construct a corpus in the manner shown in Figure 9.

TABLE I
Examples of English-Japanese bilingual BTEC sentences

English	Japanese
Could you recommend a good restaurant?	Doko ka yoi resutoran o shoukai shite moraemasen ka?
Do you feel weary?	Daruidesu ka?

First, we selected randomly 16,000 pairs of sentences from the BTEC corpus, and performed part-of-speech tagging on both languages. We used NLTK [34] for English and Mecab [35] for Japanese.

Next, we performed word alignment between the sentences using a nonparametric Bayesian inversion transduction grammars algorithm implemented in the pialign tool [36]. The alignment helps to determine the emphasized units in the target language given the emphasized units in the source language. In order to make this decision easily, we only keep the pairs of sentences which have alignments where if the emphasized word in the source language is a noun, the corresponding emphasized word in the target language is also a noun, and similarly for adjectives and adverbs. We focus only on content words because they are the most important words in the sentence. The example in Figure 9 illustrates the selection of the sentences where emphasize words are noun. This step is repeated for adjectives and adverbs.

After this, we had 2500 sentences. These sentences were verified manually to ensure the correctness and naturalness of emphasized units.

After manual verification, a total of 1015 pairs of sentences remained, and the detail of the corpus is shown in Table II. Almost all sentences had only one emphasized unit. This is natural because we often emphasize the important information in the sentence, and the number of important words are often one or two. In a limited number of cases, we emphasized more than two words.

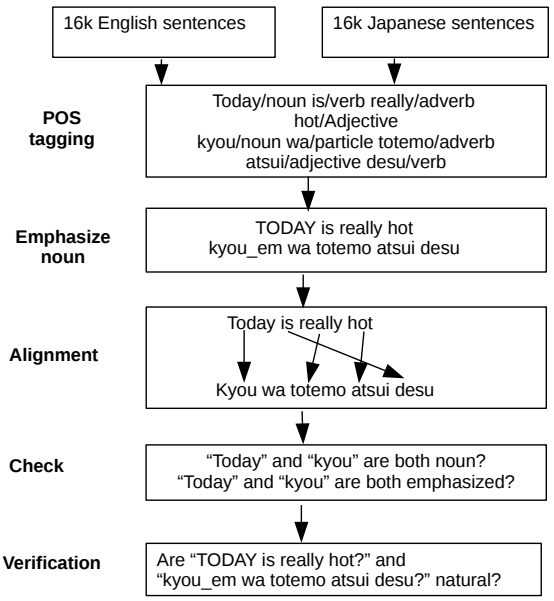

Fig. 9. Creation of emphasized sentences in the conversation corpus.

TABLE II
The conversation corpus materials

Utterances		1015
Emphasized units		1305
Utterances have X emphasized units	1	776
	2	193
	3	41
	4	5

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

B. Recording

The recording step required speakers who can speak both Japanese and English with good pronunciation and naturalness.

We have selected in total 3 bilingual, 1 native English, and 5 native Japanese speakers. The speakers were asked to read the text carefully, remember which words need to be emphasized, and emphasize them. After the recording session finished, we manually listened to the recorded audio to verify that proper emphasis had been placed on the selected word.

In the English text, the emphasized words are in upper case. Because in the Japanese text, as there is no notion of "upper case," we instead choose to attach a marker "_em" after the emphasized units[5]. An example of the displayed text is shown in Table III[6]

TABLE III
TRANSCRIPTION OF THE EMPHASIZED CORPUS

Language	Label
Japanese	kamera_em desu. suteki_em na hi_em ne.
English	It's a CAMERA. BEAUTIFUL DAY, isn't it.

The recording step was performed in a quiet environment. The audio was recorded with a frequency of 16 KHz, 16 bits, and single channel. After recording, all audio files were verified to ensure that there is no clipping caused by the speaker speaking too loudly.

The numbers of utterances and emphasized words for the corpus are shown in Table IV.

TABLE IV
RECORDED SPEECH DATA FOR THE CORPUS

Corpus	Utterances	Emphasized words	Speakers
Conversation	1015	1305	9

V. EXPERIMENTAL EVALUATION

A. Emphasis Corpus Analysis

In our first experiment, we examine the difference of power and duration between normal and emphasized words, in order to better understand how emphasis is expressed in and across languages. The analysis was performed on the conversation corpora described in Section IV.

In order to extract information of power and duration for words, we first perform forced alignment on the corpus to obtain the timing information for every word. Then, based on the timing information we compute the power (amplitude) and duration of each word. Figures 10 show the duration distributions. We can see that the emphasized words have longer duration than normal words. However, the Japanese speakers tend to use less duration than the English speakers. This is due to Japanese being a mora-timed language, in which mora (similar to syllables) tend to have the same duration [37].

Figure 11 shows the amplitude distribution of normal and emphasized words for both English and Japanese. We can see clearly that emphasized words have higher amplitude than normal words, and that the amplitude distributions are similar in both languages[7].

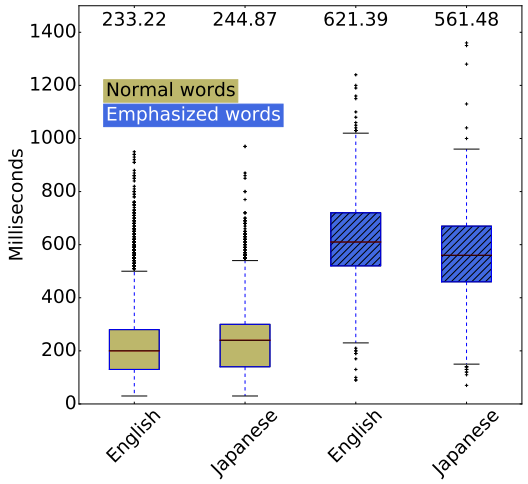

Fig. 10. Duration distribution of the conversation corpus.

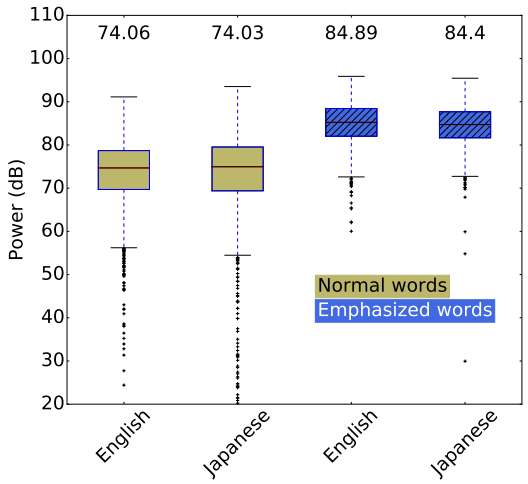

Fig. 11. Amplitude distribution of the conversation corpus.

B. Emphasis Modeling Evaluation

In this experiment, we validate that the proposed emphasis estimation method is able to detect emphasis and find which acoustic features (spectral, F0, duration) are more useful to

[5]The marker does not pose difficulties to the speakers because they are asked to read the sentences carefully before uttering them out loud.

[6]For convenience, the Japanese characters are written as their pronunciation, but the actual texts for the speaker are written with Kanji characters.

[7]It should be noted that the data used here is read speech, and it is possible that somewhat different tendencies would be seen in spontaneous speech corpora.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

estimate emphasis or distinguish between the emphasized and normal words. We do so by optimizing emphasis weight sequences using different settings of acoustic features:

- **dur**: using only the duration feature.
- **lf0**: using only log $F0$ (lf0) feature.
- **mgc**: using only spectral features.
- **mgc_dur**: using spectral and duration features.
- **mgc_lf0**: using spectral and lf0 features.
- **lf0_dur**: using lf0 and duration features.
- **mgc_lf0_dur**: combine all features.

The estimated word-level emphasis is then classified into labels of 0 and 1 indicating normal and emphasized words by using an emphasis threshold of 0.5. Then, we calculate the F-measure to show how accurately the system can detect emphasis. The process is illustrated in Figure 12.

In addition, we also perform an analysis on the correlation of emphasis between source and target languages to investigate to what extent emphasis translation based on the direct mapping approach[8], described in the above section, is sufficient based on the correlation of word-level emphasis between languages. Experimental setup and results are shown in detail as below.

1) Experimental Setup: The original corpus has 1015 sentences. After filtering out long utterances over 10 words which might not be good for model training[9], we obtain 966 utterances, which we divided into 916 utterances with 1,186 emphasized words for training and 50 utterances with 62 emphasized words for testing. All speakers in the conversation corpus including 3 bilingual, and 7 monolingual speakers are used for experiments. Thereby, we have in total 500 testing samples. The speech features include 25 dimension spectral parameters, 1 dimension log-scaled F_0, and 5 dimension aperiodic features. Each speech parameter vector included the static features and their delta and delta-deltas. The frame shift was set to 5ms. Each HSMM model has 7 states including initial and final states. We adopt STRAIGHT [38] for speech analysis.

In order to measure the relationship of the word-level emphasis across languages, we calculate the Pearson correlation coefficient to measure the strength of the linear association between them,

$$r = \frac{\sum_i (\lambda_i^{(en)} - \bar{\lambda}^{(en)})(\lambda_i^{(ja)} - \bar{\lambda}^{(ja)})}{\sqrt{\sum_i (\lambda_i^{(en)} - \bar{\lambda}^{(en)})^2}\sqrt{\sum_i (\lambda_i^{(ja)} - \bar{\lambda}^{(ja)})^2}}, \quad (18)$$

where r is the Pearson correlation coefficient, $\lambda_i^{(en)}$ is the emphasis level for the i-th word 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 emphasis level of English and Japanese, respectively.

2) Word-level Emphasis Estimation Evaluation: The result of this experiment is shown in Figure 13. Looking at the duration column, we can see that classification with the duration feature alone works relatively well in English, but does not work well in Japanese. We also observed this situation when combined with lf0-duration or mgc-duration; the performance did not increase significantly compared to lf0 or mgc only. This is consistent with the duration analysis result in the previous section, in which Japanese is a mora-timed language, so stretching out duration of words might change the meaning. The result also showed that in English all three features–duration, F_0, and spectral features–play the same role in term of emphasis prediction because they gave fairly equal performance. However, for Japanese, the spectral features are more significant compared to the other two.

By combining all features together. We achieved 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 emphasis translation experiments.

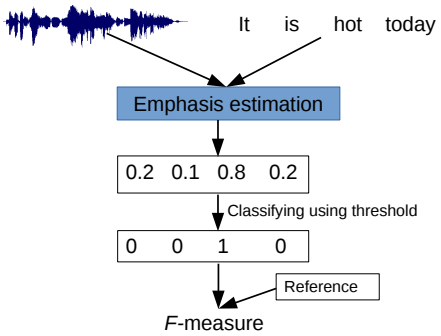

Fig. 12. Word-level emphasis estimation evaluation process.

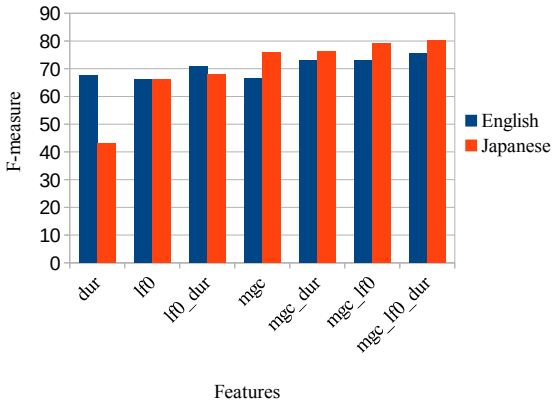

Fig. 13. F-measure of emphasis prediction.

3) Analysis of Emphasis Across English and Japanese: As we described in the section III-B, one way to translate emphasis is directly mapping emphasis weights using word alignments. If correlation coefficient of emphasis weights across languages is high, we can conclude that only emphasis weights (acoustic features) are adequate for emphasis translation, and if it is low, it means more linguistic features are needed for accurate translation.

In order to calculate the correlation coefficient, we first estimate emphasis sequences for both source and target languages. Then, we extract emphasis weights of corresponding

[8]The direct mapping approach is based on linear regression method $y = w.x + b$ that map the source emphasis level x to the target emphasis level y with parameters w and b.

[9]Specifically, we found that speakers sometime put wrong emphasis on the selected words or put extra pauses when uttering long sentences.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

words using word alignments. We use the training data for this experiment to minimize the effect of emphasis estimation's errors.

Figure 14 shows a scatter plot of the relationship between emphasis in the two languages. The Pearson correlation coefficient is 0.625. It means that emphasis is expressed differently in languages, and only a moderate correlation could be achieved by linearly mapping emphasis from the source to target languages. This also indicates that we need more sophisticated approaches to translate emphasis across languages.

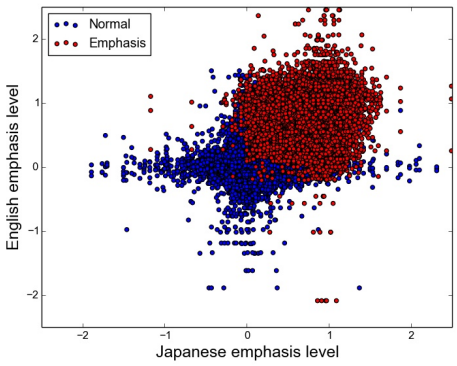

Fig. 14. Relationship between English and Japanese word-level emphasis

C. Emphasis Translation Evaluation Without ASR and MT Errors

In this subsection, we evaluate the performance of emphasis translation based on conditional random fields (CRFs), which we hypothesize will be more effective than the direct mapping approach. Moreover, we also evaluate the effect of using each set of features on emphasis translation. Specifically, we expect that linguistic features will improve the performance over using only acoustic features due to the complexity of emphasis expression across languages.

In this experiment, we assumed the ASR and MT system output correct hypotheses, so all errors are due to emphasis translation only. We also evaluate the system considering ASR and MT errors in Section V-D.

1) Experimental Setup: Due to the limitation of the speaker dependence of the current emphasis estimation and synthesis methods, we used only 1 native English speaker and 1 native Japanese speaker. We used 916 utterances for CRF training, and two testing sets–50 utterances and 95 utterances in the following sections.

Because the CRFs only deal with discrete values, we had to quantize the continuous word-level emphasis. Different quantization schemes and their influence in emphasis translation performance are described in detail in Section V-C3.

The performance is measured objectively by F-measure, which is calculated in the similar way as the previous experiment. This score represents for how accurately we can preserve emphasis information in the target language.

2) Emphasis Translation Results: The first testing set, which consists of 50 utterances, was used in this evaluation. The result is shown in Table V. Comparing the direct mapping approach (the bottom line) and CRF approaches, we can see that when using only emphasis information, the CRF approach seems to be struggling to learn the correlation between source and target languages, while the direct mapping approach can capture this information better. However, when using more linguistic features such as words, PoS tags and their context units, the CRF starts showing its advantages in flexibly handling many features together, where it is difficult to do the same thing in the direct mapping approach. This indicates that along with the acoustic feature (emphasis level), the linguistic features are also important pieces of information in capturing the complexity in cross-lingual emphasis translation. The result also follows our expectation in the previous experiment that linear regression model do not work well in this task where correlation coefficient is at the moderate level.

In addition, we also evaluate the effect of the combination of input features described in section III-B2 to find out which features give the best performance. We found out that the model using only emphasis in the source language (2nd rows) performs better than the model using only target information (1st row). This demonstrates that our model is effectively translating emphasis from the source, as opposed to simply predicting based on the target. Furthermore, when adding target language-specific feature such as word contexts and PoS tag contexts, the performance is getting even better from 82.8% to 91.6% F-measure[10].

Looking at the 2nd and 3rd rows, we can see that emphasis context in the source language does not help for emphasis translation, indicating that word-level emphasis in the target language depends mainly on 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 words, and PoS tags are also contributing to the improvement of the translation model.

3) Word-level Emphasis Quantization Evaluation: As described in section III-B2, CRFs require discrete state sequence, so we need to quantize emphasis weights. Because it is difficult to intuitively determine which quantization scheme is best, we perform an experiment to investigate how emphasis quantization affects emphasis translation performance by using 4 different quantization schemes as follows,

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

[10]We also tried to use features that express information about longer context and other target language-specific features, but they did not lead to increase performance, perhaps because the training data is not sufficient to learn these sparser features

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

TABLE V
F-MEASURE FOR DIFFERENT COMBINATION OF INPUT FEATURES ON EMPHASIS TRANSLATION TASK WITHOUT ASR AND MT ERRORS. THE FEATURE INCLUDING WORDS, PART-OF-SPEECH TAGS, AND THEIR PRECEDING+SUCCEEDING UNITS ANNOTATED WITH THE "CONTEXT" SUFFIX.

Emphasis		**Word**		**Tag**		**Emphasis context**		**Word context**		**Tag context**		F**-measure**
En	Ja	En	Ja	En	Ja	En	Ja	En	Ja	En	Ja	
	✓		✓		✓							81.6 (±10.2)
✓	✓											82.8 (±10.1)
✓	✓					✓						82.8 (±10.1)
✓	✓	✓	✓									84.8 (±9.8)
✓	✓	✓	✓	✓	✓							90.0 (±9.2)
✓	✓	✓	✓	✓	✓			✓	✓			88.7 (±9.3)
✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	90.0 (±9.2)
✓	✓	✓	✓	✓	✓						✓	**91.6** (±8.7)
Direct mapping approach												86.8 (±10.2)

- **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 buckets of 0.1.
- **Labels:** The word-level emphasis in the source language is quantized according to "0.3 Quant" and emphasis in the target language is derived from the labels from the corpus and has binary values, 1 for emphasis, 0 for normal.

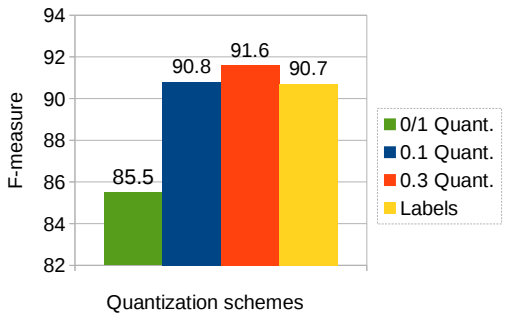

Fig. 15. F-measure for different quantization methods.

The result is shown in Figure 15. We can see that the quantization scheme "0.3 Quant" gives the best result, likely because it provides an appropriate amount of training data for each class. And more importantly, it even performs slightly better the manually created "Labels," suggesting that training the system using the quantized word-level emphasis can be more effective than binary values.

4) Subjective Emphasis Prediction on Emphasis Translation: The final goal of emphasis translation is to help target language listeners be able to capture source language speakers' emphasis information. In this experiment, we performed a manual evaluation to determine how well the listener can detect emphasis translated by the end-to-end system. We hypothesize that subjective evaluation will have slightly lower performance compared to objective evaluation due to the imperfection of speech synthesis systems, which might synthesize normal sounds for emphasized words and vice-versa. We asked 6 native Japanese speakers to listen to 150 translated emphasized utterances from the following 3 systems, and select the words that they think are emphasized.

- **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 V.
- **Natural:** Natural speech by a Japanese speaker.

Fig. 16 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 has many words that are randomly emphasized, because it was trained on emphasized utterances, but there is no emphasis control based on the source utterance.

Comparing these results with the automatic evaluation, 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.

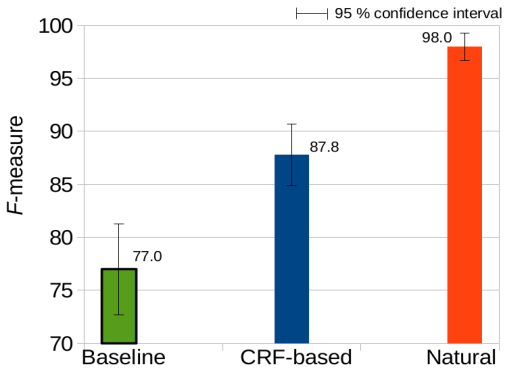

Fig. 16. Emphasis prediction F-measure for manual evaluation

D. Emphasis Estimation and Translation with Imperfect ASR and MT

The above experiments are run on perfect ASR and MT hypotheses to verify the performance of the emphasis translation part only. However, in reality, it is not possible to have

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

TABLE VI
F-MEASURE OF EMPHASIS TRANSLATION WITH ASR AND MT ERRORS (0.1 QUANT.). THE TEST SET *A* CONSISTS OF 95 UTTERANCES AND THE TEST SET *B* CONSISTS OF 86 UTTERANCES EXCLUDING THE ONE LOST EMPHASIZED WORDS DUE TO MT ERRORS.

Emphasis		**Word**		**Tag**		**Word context**		**Tag context**		F**-measure**	
En	Ja	En	Ja	En	Ja	En	Ja	En	Ja	Set *A*	Set *B*
✓	✓									63.1 (±8.8)	64.1 (±9.0)
✓	✓	✓	✓							70.2 (±8.5)	72.6 (±8.8)
✓	✓	✓	✓	✓	✓					69.9 (±8.2)	72.8 (±8.8)
✓	✓	✓	✓	✓	✓	✓	✓			**77.1** (±8.3)	**80.0** (±8.2)
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	75.2 (±8.2)	78.0 (±8.5)

perfect ASR and MT accuracy. In this experiment, we evaluate the ability of the proposed method to reproduce emphasis in the target language when errors from ASR and MT are taken into account. This is a fully end-to-end emphasis translation system.

We first generate emphasis labels from MT translated transcriptions by aligning them with original labels in the corpus to find emphasized words. However, we can only generate the labels that have emphasized words for 89 out of 95 translated utterances, the other 6 utterances lost the emphasized words due to the errors from ASR and MT modules and are considered as normal utterances.

The result is shown in Table VI[11]. Although suffering from ASR and MT errors, the system still achieves a relatively high performance of 77.06% F-measure on the test set *A* and 80.00% on the test set *B*. The difference of 1-3% between the test set *A* and *B* indicates the loss due to losing the emphasized word. One more time, the linguistic information shows its importance in emphasis translation tasks by 6% F-measure improvement (line 1^{st} and 4^{th}) on the test set *B*. However, the performance dropped when we use so many input features (line 5^{th}), this is likely due to over-fitting problems because the training data is limited.

In addition, we also perform a subjective evaluation, which is similar with the previous experiment. We observed a similar tendency that the subjective result has 4% lower F-measure compared to the objective result.

1) Naturalness Evaluation: In this experiment, we evaluate the naturalness of the following three systems to find out if the quality and naturalness of translated emphasized speech is degraded compared to conventional S2ST systems,

- **Baseline:** No emphasis translation is performed. The text-to-speech module is trained on the emphasis corpus.
- **0-emphasis:** All emphasis weights in the target language are set to 0. The output of this system is similar to the conventional S2ST system where there is no emphasis in the speech.
- **CRF:** Our proposed emphasis translation method.

In order to evaluate the naturalness of the proposed method. We first generate 3 sets of speech, each containing 93 speech samples from 3 systems above. We asked 7 native Japanese listeners to listen to each pair of audio, which are played in random order, and select a preferred one. The preference (A/B) score is shown in Figure 17. As we can see, the

[11]CRF models are trained using "0.1 Quant." We also evaluated others quantization schemes, but this is the best one.

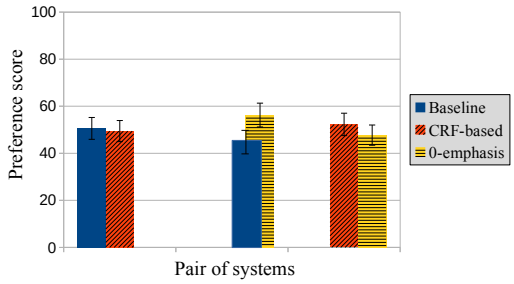

Fig. 17. Preference for each pair of methods from 6 listeners

scores for each method are similar to each other, indicating that emphasis translation does not degrade the naturalness of the synthetic speech. Furthermore, because the **0-emphasis** system represents a conventional S2ST system that does not consider emphasis information, the result also demonstrates that emphasis translation does not harm the conventional S2ST systems in terms of naturalness.

VI. CONCLUSION

In this paper, we have proposed a method to translate emphasis across languages. Unlike previous works where the emphasis translation model relied on translating only F_0 patterns, or can only handle small vocabularies (10 digits), our proposed method modeled emphasis at word-level, which is a natural way as how human emphasize speech, and also considered all acoustic features such as duration, power, and F_0. Moreover, each of translation components (emphasis estimation, translation, synthesis) can also handle tasks with open vocabularies.

In order to achieve the goal, we have collected and analyzed emphasized speech from English-Japanese bilingual speech corpora to find out how people emphasized speech across languages. The analysis on the corpus showed that there are significant differences in terms of duration and power between normal and emphasized words. And an analysis of word-level emphasis showed a medium linear correlation, meaning that the way people emphasize words are different between languages.

With regards to emphasis modeling and analysis. The linear-regression hidden semi-Markov model was shown effective in modeling emphasized speech, making it possible to utilize all acoustic features rather than individual ones. The evaluation on emphasis prediction also demonstrated that word-level em-

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

phasis can be estimated accurately when using all the speech features including spectral, log F_0, and duration features.

The work on integrating emphasis estimation, and emphasis translation into a conventional S2ST system demonstrated that the translation model based on conditional random fields (CRFs) can translate emphasis information accurately, and does not degrade the naturalness of the synthetic speech. We also observed that along with acoustic features (word-level emphasis), the linguistic features such as word and part of speech tags also contribute to the improvement of the translation model. The proposed emphasis translation model can also be applied directly to open vocabulary tasks because individual components including emphasis estimation and translation are not trained on word-level, but on phoneme-level.

However, there are still some limitations. First, the same LR-HSMM model is used for emphasis estimation and synthesis components. Although this simplifies the translation model, it restricts the model to be speaker dependent. One solution is performing speaker adaptation to adapt the current model to a specific speaker, or training a speaker-independent model by increasing the number of Gaussian components of each HSMM state. Second, emphasis translation based on CRFs requires emphasis-level quantization and each output labels are also independent. It might be possible to improve the performance by applying other cost functions to take into account correlation between emphasis levels.

Future works will include collecting of more data, such as spontaneous speech, and speech that contains other information such as emotion. We also hope to create systems that handle different varieties of paralinguistic information in a single S2ST system. Additionally, other linguistic features such as emotional words might also be useful for the translation model because those words are more likely to be emphasized than others. We also plan to develop a speaker-independent emphasis estimation component that would generalize the translation model to be more robust on out-of-training-data speakers.

VII. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 24240032, by the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan, and by a joint research project with ATR-Trek.

References

[1] S. Nakamura, "Overcoming the language barrier with speech translation technology," *Science & Technology Trends - Quarterly Review No.31*, April 2009.
[2] A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "A study on the effect of prosodic emphasis transfer on overall speech translation quality," in *Proceedings of ICASSP*, 2013.
[3] 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.
[4] E. Fudge, *English Word-stress*. Routledge, 2015.
[5] B. Arons, "Pitch-based emphasis detection for segmenting speech recordings," in *Proceedings of ICSLP*, 1994, pp. 1931–1934.
[6] A. Tsiartas, P. Georgiou, and S. S. Narayanan, "Toward transfer of acoustic cues of emphasis across languages," in *Proceedings of InterSpeech*, 2013.
[7] T. Kano, S. Sakti, S. Takamichi, G. Neubig, T. Toda, and S. Nakamura, "A method for translation of paralinguistic information," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2012.
[8] 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.
[9] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Intent transfer in speech-to-speech machine translation," in *Proceedings of SLT*, Dec 2012, pp. 153–158.
[10] P. D. Aguero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *Proceedings of ICASSP*, vol. 1, 2006.
[11] 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.
[12] 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.
[13] 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.
[14] 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 *INTERSPEECH*, 2015.
[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*. IEEE Signal Processing Society, Dec. 2011.
[16] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *NAACL*, 2003, pp. 48–54.
[17] Y. Liu, Q. Liu, and S. Lin, "Tree-to-string alignment template for statistical machine translation," in *ACL*, 2006, pp. 609–616.
[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of NIPS*, 2014, pp. 3104–3112.
[19] 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.
[20] H. Fujisaki, "Information, prosody, and modeling - with emphasis on tonal features of speech," in *Proceedings of Speech Prosody*, 2004, pp. 1–10.
[21] 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, Sep. 2007.
[22] 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.
[23] 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.
[24] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.
[25] 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 – 92, 2011.
[26] 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.
[27] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of HLT*, 1994, pp. 307–312.
[28] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *ASJ (E)*, vol. 21, pp. 79–86, 2000.
[29] M. Gales, "Cluster adaptive training of hidden Markov models," *IEEE*, vol. 8, no. 4, pp. 417–428, 2000.
[30] 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.
[31] M. L. Gregory and Y. Altun, "Using conditional random fields to predict pitch accents in conversational speech," in *Proceedings of ACL*, 2004.
[32] N. Okazaki, "Crfsuite: a fast implementation of conditional random fields (CRFs)," 2007. [Online]. Available: http://www.chokkan.org/software/crfsuite/
[33] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proceedings of EUROSPEECH*, 2003, pp. 381–384.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TASLP.2016.2643280

[34] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009.
[35] T. Kudo, *MeCab: Yet Another Part-of-Speech and Morphological Analyzer*. http://mecab.sourceforge.jp, 2006.
[36] G. Neubig, T. Watanabe, E. Sumita, S. Mori, and T. Kawahara, "An unsupervised model for joint phrase alignment and extraction," in *Proceedings of ACL*, 2011, pp. 632–641.
[37] P. Ladefoged and K. Johnson, *A Course in Phonetics*. Cengage Learning, 2010.
[38] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187 – 207, 1999.

Quoc Truong Do received his B.E. from University of Engineering and Technology, Hanoi, Vietnam, in 2013, and his M.S. from the Graduate School of Information Science, NAIST, Nara, Japan in 2015. He is currently in the doctoral course at NAIST, Japan. He interested in speech and natural language processing, with a focus on speech recognition, and speech translation. He is a student member of ISCA, and ASJ.

Tomoki Toda received his B.E. degree from Nagoya University, Japan, in 1999 and his M.E. and D.E. degrees from Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science from 2003 to 2005. He was then an Assistant Professor (2005-2011) and an Associate Professor (2011-2015) at NAIST. From 2015, he has been a Professor in the Information Technology Center at Nagoya University. His research interests include statistical approaches to speech processing. He received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).

Graham Neubig received his B.E. from University of Illinois, Urbana-Champaign in 2005, and his M.S. and Ph.D. in informatics from Kyoto University in 2010 and 2012 respectively. He is currently an assistant professor at the Nara Institute of Science and Technology, Nara, Japan. His research interests include speech and natural language processing, with a focus on machine learning approaches for applications such as machine translation, speech recognition, and spoken dialog.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. From 2011, she has been an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. She is now also a board member of SLTU (Spoken Language Technologies for Under-resourced languages) and a committee member of SIG ELRA-LRL (Low Resourced Languages). Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampoli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.

Copyright (c) 2017 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.