An End-to-end Model for Cross-Lingual Transformation of Paralinguistic Information

Takatomo Kano, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology, Japan

Abstract

Speech translation is a technology that helps people communicate across different languages. The most commonly used speech translation model is composed of Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-To-Speech synthesis (TTS) components, which share information only at the text level. However, spoken communication is different from written communication in that it uses rich acoustic cues such as prosody in order to transmit more information through non-verbal channels. This paper is concerned with speech-to-speech translation that is sensitive to this paralinguistic information. Our long-term goal is to make a system that allows users to speak a foreign language with the same expressiveness as if they were speaking in their own language. Our method works by reconstructing input acoustic features in the target language. From the many different possible paralinguistic features to handle, in this paper we chose duration and power as a first step, proposing a method that can translate these features from input speech to the output speech in continuous space. This is done in a simple and language-independent fashion by training an end-to-end model that maps source language duration and power information into the target language. Two approaches are investigated: linear regression and neural network models. We evaluate the proposed method and show that paralinguistic information in the input speech of the source language can be reflected in the output speech of the target language.

Keywords: Paralinguistic Information, Speech to speech translation, Emotion, Automatic Speech Recognition, Machine Translation, Text to Speech synthesis

1. Introduction

When we speak, we use many different varieties of acoustic and visual cues to convey our thoughts and emotions. Many of those paralinguistic cues transmit additional information that cannot be expressed in words. While these cues may not be a critical factor in written communication, in spoken communication they have great importance; even if the content of the words are the same, if the intonation and facial expression are different an utterance can take an entirely different meaning. As a result, it would be advantageous to take into account these paralinguistic features of speech in any system that is constructed to aid or augment human-to-human communication.

Speech-to-speech translation helps people communicate across different languages, and is thus one prime example of such a system. However, standard speech translation systems only convey linguistic content from source languages to target languages without considering paralinguistic information. Although the input of ASR contains rich prosody information, the words output by ASR are in written form that have no indication of the prosody included in the original speech. As a result, the words output by TTS on the target side will thus be given the canonical prosody for the input text, not reflecting the prosodic traits of the original speech. In other words, because information sharing between the ASR, MT, and TTS modules is limited to only lexical information, after the ASR conversion from speech to text, source-side acoustic details such as rhythm, emphasis, or emotion are lost.

This paper is concerned with speech-to-speech translation that is sensitive to paralinguistic information, with the long-term goal of making a system that allows a user to speak a foreign language with the same expressiveness as if they were speaking in their own language. The proposed method works by recognizing acoustic features (duration and power) in the source language, then reconstructing them in the target language. From the many different possible paralinguistic features to handle, in this paper we chose duration and power as a first step, proposing a method that can translate these features from the input speech to the output speech in continuous space.

First, we extract features at the level of Hidden Markov Model (HMM) states, the use a paralinguistic translation model to predict the duration and power features of HMM states of the output speech. Specifically, we use two approaches: a linear regression model that predicts separately predicts prosody for each word in the vocabulary, and a model that can adapt to more

general tasks by training a single model that is applicable to all words in the vocabulary using neural networks.[1]

2. Conventional Speech-to-Speech Translation

In conventional speech-to-speech translation systems, the ASR module decodes the text of the utterance from input speech. Acoustic features are represented as $\boldsymbol{A} = [a_1, a_2 ... a_{N_a}]$ and the corresponding words are represented as $\boldsymbol{E} = [e_1, e_2, ..., e_{N_e}]$. N_a and N_e are the lengths of the acoustic feature vectors and spoken words respectively.

The ASR system finds $\boldsymbol{E}$ that maximizes $P(\boldsymbol{E}|\boldsymbol{A})$. By Bayes' theorem, we can convert this to

$$P(\boldsymbol{E}|\boldsymbol{A}) \propto P(\boldsymbol{A}|\boldsymbol{E})P(\boldsymbol{E}), \tag{1}$$

where $P(\boldsymbol{A}|\boldsymbol{E})$ is the Acoustic Model (AM) and $P(\boldsymbol{E})$ is the Language Model (LM). The MT module finds the target words sequence $\boldsymbol{J}$ that maximizes probability $P(\boldsymbol{J}|\boldsymbol{E})$:

$$\hat{\boldsymbol{J}} = \underset{\boldsymbol{J}}{\operatorname{argmax}}\, P(\boldsymbol{J}|\boldsymbol{E}). \tag{2}$$

Similarly to what was done for ASR, we can convert $P(\boldsymbol{J}|\boldsymbol{E})$ as follows:

$$\hat{\boldsymbol{J}} = \underset{\boldsymbol{J}}{\operatorname{argmax}}\, P(\boldsymbol{E}|\boldsymbol{J})P(\boldsymbol{J}), \tag{3}$$

where $P(\boldsymbol{E}|\boldsymbol{J})$ is a translation model and $P(\boldsymbol{E})$ is a language model.

The TTS module generates speech parameters $\boldsymbol{O} = [o_1, o_2, ..., o_{N_o}]$ given HMM AM states $\boldsymbol{H_x} = [h_1, h_2, ..., h_{N_h}]$ that represent $\boldsymbol{J}$. Here N_o and N_h is the length of the generated speech parameter sequence and the number of states of the HMM AM. The output $\boldsymbol{O} = [o_1, o2, ..., o_{N_o}]$ can be represented by

$$\hat{\boldsymbol{O}} = \operatorname{argmax} P(\boldsymbol{O}|\boldsymbol{H}) \tag{4}$$

These three modules, only share information through $\boldsymbol{E}$ or $\boldsymbol{J}$, which are strings of text in the source and target languages respectively. As a result, all non-verbal information that was original expressed in source speech $\boldsymbol{A}$ is lost the moment it is converted into source text $\boldsymbol{E}$ by ASR.

[1] Part of the content of this article is based on content that has been published in IWSLT and InterSpeech [10, 11]. In this paper describe these methods using an unified formulation, adds a more complete survey, and discuss the results in significantly more depth.

3. Speech Translation considering Paralinguistic Information

In order to perform speech translation in a way that is also able to consider paralinguistic information, we need consider how to handle paralinguistic features included in $\boldsymbol{A}$. Specifically, we need to extract acoustic features during ASR, translate them to another language during MT, and then reflect them in the target speech during TTS.

The first design decision we need to make is at what granularity at which to represent paralinguistic features: phoneme, word, phrase, or sentence level. In the ASR and TTS modules, phonemes are the smallest lexical unit that represent speech, and in the MT module, words are the smallest unit handled by the system. From the point of view of speech processing phonemes are a good granularity with which to handle paralinguistic features. However, in human speech, paralinguistic features such as emphasis, surprise, and sadness can be more intuitively attributed to the word, phrase and sentence level [19]. Thus, as the main focus of our work is on methods for translation of emphasis between languages, for this paper we decide to construct our models purely on the word level. We create word-level AMs for ASR and TTS, extract the paralinguistic features $\boldsymbol{X}$ belonging to each word, and translate these word-level acoustic features from the source to target directly using a regression model in the MT module. Finally we use translated acoustic features to generate output speech in the TTS module.

While the overall framework here is independent of the speech translation task, as the research is ambitious, our experiments below focus on a limited setting of translating digits. This digit translation task can be motivated by a situation where a customer is contacting a hotel staff member attempting to make a reservation. The customer conveys the reservation number, and the hotel staff member confirms, but the number turns out to be incorrect. In this case, the customer would re-speak the number, using prosody to emphasize the missing information. The problem formulation below will also use this setting as an example, specifically the example of English-Japanese translation.

3.1. Speech Recognition

The first step of the process uses ASR to recognize the lexical and paralinguistic features of the input speech. This can be represented formally as

$$\hat{\boldsymbol{E}}, \hat{X} = \underset{\boldsymbol{E},\boldsymbol{X}}{\operatorname{argmax}} P(\boldsymbol{E}, \boldsymbol{X} | \boldsymbol{A}), \tag{5}$$

where $\boldsymbol{A}$ indicates the input speech, $\boldsymbol{E}$ indicates the words included in the utterance and $\boldsymbol{X}$ indicates paralinguistic features of the words in $\boldsymbol{E}$. In order to recognize this information, we construct a word-based HMM AM. The AM is trained with audio recordings of speech and the corresponding transcriptions $\boldsymbol{E}$ using the standard Baum-Welch algorithm. Once we have created our model, we perform simple speech recognition using the HMM AM and a language model that assigns a uniform probability to all digits. Viterbi decoding can be used to find $\boldsymbol{E}$. Finally we can decide the duration vector x_i of each word e_i based on the time spent in each state of the HMM AM in the path found by the Viterbi algorithm. The power component of the vector is chosen in a similar way, and by taking the mean power value over frames that are aligned to the same state of the AM. We express power as $[power, \Delta power, \Delta\Delta power]$ and join these features together as a super-vector to control power in the translation step. Δ indicates dynamic features. It should be noted that in contrast to other work such as [2], for the ASR part, we dont need a manual labeling the prosody of speech and simply segment each word and extract observed acoustic features.

3.2. Lexical and Paralinguistic Translation

Lexical translation finds the best translation $\boldsymbol{J}$ of recognized source sentence $\boldsymbol{E}$. Generally we can use any variety of statistical machine translation to obtain this translation in standard translation tasks, but for digit translation we can simply write one-to-one lexical translation rules with no loss in accuracy such as $j_i = e_i$ where i is word index. Paralinguistic translation converts the source-side acoustic feature vector $\boldsymbol{X}$ into the target-side acoustic feature vector $\boldsymbol{Y}$ according to the following equation

$$\hat{\boldsymbol{Y}} = \underset{\boldsymbol{Y}}{\operatorname{argmax}} P(\boldsymbol{Y}|\boldsymbol{X}) \tag{6}$$

There are many types of acoustic features used in ASR and TTS systems, including MFCC, MGC, Filter-bank, F0, power, and duration. In this work we use power and duration to express "emphasis information". We make this decision due to the fact that MFCC, Filter-bank, and MGC features are more strongly connected to lexical information related to the content of the utterance. F0, power and duration are more correlated with paralinguistic information regarding the method of speech, but because Japanese is a tonal language where F0 has a strong relationship with content distinctions, in this work we focus on duration and power. We control duration and power of each

word using a source-side duration and power super-vector $\boldsymbol{x_i} = [x_1, ..., x_{N_x}]$ and a target-side duration duration and power super-vector $\boldsymbol{y_i} = [y_1, ..., y_{N_y}]$. Here N_x and N_y represent the length of the paralinguistic feature vector for each word i.

In these vectors N_x represents the number of HMM states on the source side and N_y represents the number of HMM states on the target side. The sentence duration and power vector consists of the concatenation of the word duration and power vectors such that $\boldsymbol{Y} = [y_1, \ldots, y_n, \ldots, y_{N_y}]$. We can assume that duration and power translation of each word pair is independent from that of other words, allowing us to find the optimal $\boldsymbol{Y}$ using the following equation

$$\hat{\boldsymbol{Y}} = \underset{\boldsymbol{Y}}{\operatorname{argmax}} \prod_n P(y_n|x_n) \tag{7}$$

The word-to-word acoustic translation probability $P(y_n|x_n)$ is calculated according to a linear regression matrix that indicates that y_i is distributed according to a normal distribution

$$P(y_i|x_i) = N(y_i; \boldsymbol{W}_{e_i,j_i}, x_i', \boldsymbol{A}) \tag{8}$$

where x' is transposed x and $\boldsymbol{W}_{e_i,j_i}$ is a regression matrix with bias defining a linear transformation expressing the relationship in duration and power between e_i and j_i. An important point here is how to construct regression matrices for each of the words we want to translate. In order to do so, we optimize each regression matrix in the translation model training data by minimizing root mean squared error (RMSE) with a regularization term

$$\hat{W}_{e_i,j_i} = \underset{\boldsymbol{W}_{e_i,j_i}}{\operatorname{argmax}} \sum_{n=1}^{N} \|y_n^* - y_n\|^2 + \alpha \|\boldsymbol{W}_{e_i,j_i}\|^2, \tag{9}$$

where N is the number of training samples, n is the id of a training sample, $y*$ is target language reference word duration and power vector, and α is a hyper-parameter for the regularization term to prevent over-fitting. This maximization can be solved in closed form using simple matrix operations.

3.3. Speech Synthesis

In the TTS part of the system we use an HMM-based speech synthesis system [24], and reflect the duration and power information of the target word paralinguistic information vector onto the output speech.

$$\hat{H}_y = \operatorname{argmax} P(\boldsymbol{H_y}|\boldsymbol{Y}) \tag{10}$$

The output speech parameter vector sequence $\boldsymbol{O} = [o_1, ..., o_{N_o}]$ is determined by maximizing the target HMM AM $\hat{H}_y$ likelihood function given the target language sentence $\hat{\boldsymbol{J}}$ as follows:

$$\hat{\boldsymbol{O}} = \underset{\boldsymbol{O}}{\operatorname{argmax}} P(\boldsymbol{C}|\hat{\boldsymbol{J}}, \hat{\boldsymbol{H}}_{\boldsymbol{y}}) \tag{11}$$

$$subject \quad to \quad \boldsymbol{C} = \boldsymbol{M}\boldsymbol{O} \tag{12}$$

where $\boldsymbol{C}$ is a joint static and dynamic feature vector sequence of the target speech parameters and $\boldsymbol{M}$ is a transformation matrix from the static feature vector sequence into the joint static and dynamic feature vector sequence. When generating speech, the corresponding HMM AM parameters and the length of the target language state sequence are determined by $\hat{\boldsymbol{Y}}$ resulting from the paralinguistic translation step. While TTS generally uses phoneme-based HMM models, we instead used a word-based HMM to maintain the consistency of feature extraction and translation. Usually, in TTS phoneme-based HMM AMs, the current HMM AM is heavily influenced by the previous and next phonemes, making it necessary to consider context information from input sentence. However, in the digit translation task the vocabulary is small, so we construct an word level independent context HMM AM.

4. End-to-end Paralinguistic Translation Methods

In this section we describe two ways to translate paralinguistic features of the source words to target words. The first is simple linear regression model that trains a separate model for each word in the vocabulary, and another is neural network model that trains a single model for the entire vocabulary but provides the model with information of the word identity.

4.1. Linear Regression Models

Paralinguistic translation converts the source-side paralinguistic features $\boldsymbol{X}$ into the target-side paralinguistic features $\boldsymbol{Y}$, in a manner inspired by previous work on voice conversion [1, 21]

$$\hat{\boldsymbol{Y}} = \underset{\boldsymbol{Y}}{\operatorname{argmax}} \quad P(\boldsymbol{Y}|\boldsymbol{X}) \tag{13}$$

In particular, we control duration and power using the source-side word feature vector $x_i = [x_1, ..., x_{N_h}]$ and target-side word feature vector $y_i = [y_1, ..., y_{N_h}]$. Here i represents the word id within the vocabulary. In these

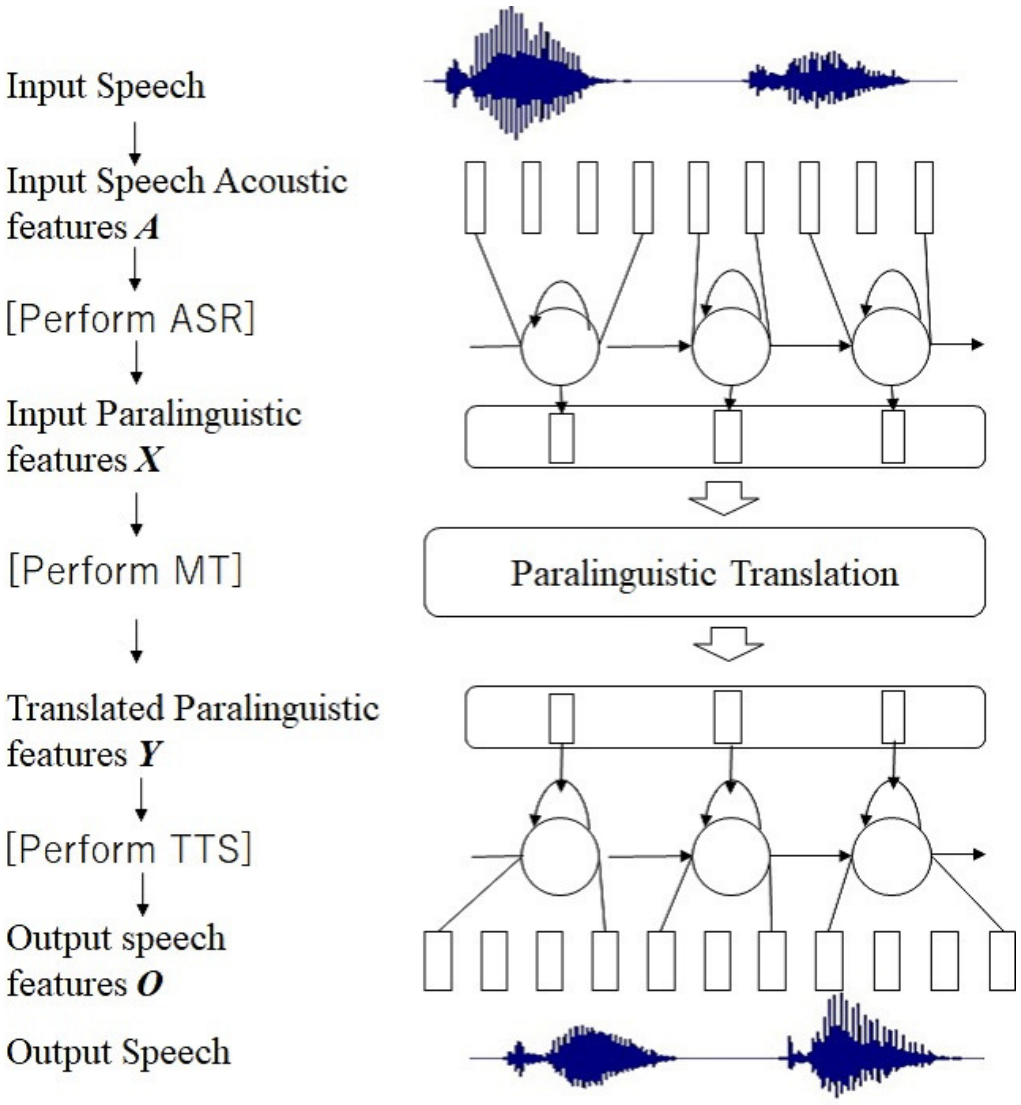

Figure 1: Overview of the proposed method

vectors N_h represents the number of HMM states on the source and target sides. The sentence feature vector consists of the concatenation of the word duration and power vectors such as $\boldsymbol{Y}$ where I is the length of the sentence. We assume that duration and power translation of each word pair

is independent, giving the following equation:

$$\hat{\boldsymbol{Y}} = \underset{\boldsymbol{Y}}{\text{argmax}} \quad \prod P(y_i|x_i) \tag{14}$$

This can be defined with any function, but we choose to use linear regression, which indicates that y_i is distributed according to a normal distribution

$$P(y_i|x_i) = N(y_i; \boldsymbol{W}_{e_i,j_i}, x_i', S) \tag{15}$$

where, x' is transposed x and $\boldsymbol{W}_{e_i,j_i}$ is a regression matrix with bias defining a linear transformation expressing the relationship in duration and power between e_i and j_i.

An important point here is how to construct regression matrices for each of the words we want to translate. In order to do so, we optimize each regression matrix on the translation model training data by minimize RMSE with a regularization term. This separate training of a model for each word pair allows the model to be expressive enough to learn how each words' acoustics are translated into the target language. However, this has serious problems with generalization, as we will not be able to translate any words that have not been observed in our training data a sufficient number of times to learn the transformation matrix. The simplest way to generalize this model is by not training a separate model for each word, but a global model for all words in the vocabulary. This can be done by changing the word-dependent regression matrix $\boldsymbol{W}_{e_i,j_i}$ into a single global regression matrix $\boldsymbol{W}$ and training the matrix over all samples in the corpus. However, this model can be expected to not be expressive enough to perform paralinguistic translation properly. For example, the mapping of duration and power from a one-syllable word to another one-syllable word, and from a one-syllable word to a two-syllable word would vary greatly, but the linear regression model only has the power to perform the same mapping for each word.

4.2. Global Neural Network Models

As a solution to the problem of the lack of expressiveness in linear regression, we additionally propose a global method for paralinguistic translation using neural networks. Neural networks have higher expressive power due to their ability to handle non-linear mappings, and are thus an ideal candidate for this task. In addition, they allow for adding features for many different types of information following ASR, MT, and TTSs common practice, such

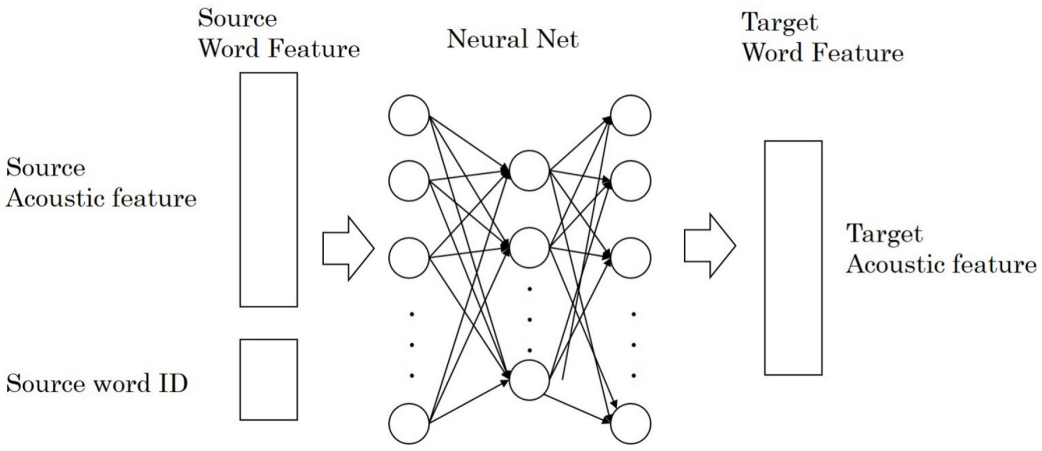

Figure 2: Neural Network for acoustic feature translation

as word ID vectors, word position, left and right words of input and target words, part of speech, the number of syllables, accent types, etc. This information is known to be useful in TTS [24], so we can likely improve estimation of the output duration and power vector in translation as well. In this research, we use a feed forward neural network that proposes the best output word acoustic feature vector given input word acoustic feature vector $\boldsymbol{X}$. As additional features, we also add a binary vector with the ID of the present word set to 1, and the position of the output word. In this work, because the task is simple we just use this simple feature set, but this could be expanded easily more for complicated tasks. For the sake of simplicity in this formulation we show an example with the word acoustic feature vector only. First, we set each input unit x_i equal to the input vector value: $l_i = x_i$ The hidden units h_j are calculated according to the input-hidden unit weight matrix W_h:

$$\pi_j = \frac{1}{1 + \exp(-\alpha \sum_i w^h_{i,j} l_i)} \tag{16}$$

where α is gradient of sigmoid function. The output units ψ_k and final acoustic feature output y_k are set as

$$\psi_k = \sum_j w^o_{j,k} \pi_j . y_k = \psi_k \tag{17}$$

where W_o is the hidden-output unit weight matrix. As an optimization criterion we use minimization of RMSE, which is achieved through simple back

propagation and weight update, as is standard practice in neural network models.

5. Evaluation

5.1. Experimental Setting

We examine the effectiveness of the proposed method through English-Japanese speech-to-speech translation experiments. We use the "AURORA-2" data set. The "AURORA-2" data are based on a version of the original TIDigits down-sampled at 8 kHz from 55 male and 55 female speakers. Different noise signals have been artificially added to clean speech data.

As mentioned previously, in these experiments we assume the use of speech-to-speech translation in a situation where the speaker is attempting to reserve a ticket by phone in a different language. When the listener makes a mistake when listening to the ticket digit, the speaker re-speaks, emphasizing the mistaken digit. In this situation, if we can translate the paralinguistic information, particularly emphasis, this will provide useful information to the listener about where the mistake is. In order to simulate this situation, we recorded a bilingual speech corpus where an English-Japanese bilingual speaker emphasizes one word during speech in a string of digits. The content spoken was 500 sentences from the AURORA-2 test set, chosen to be word balanced by greedy search [25] This was further split into a training set of 445 utterances and the test set is 55 utterances.

To train the ASR model, we use 8440 utterances of clean and noisy speech from the training set of the AURORA-2 dataset and train with the HTK toolkit. In the ASR module we trained an HMM AM, where each word has 16 HMM states, and for silence we allocate 3 states. The lexical translation is performed by Moses [13].We further used the 445 utterances of training data to build an English-Japanese speech translation system that includes our proposed paralinguistic translation model. We set the number of HMM states per word in the ASR AM to 16, the shift length to 5ms, and other various settings to follow [17, 14]. To simplify the problem, experiments were done where ASR has no errors. For TTS, we use the same 445 utterances for training an independent context synthesis model. In this case, the speech signals were sampled at 16kHz. The shift length and HMM states are identical to the setting for ASR.

In the evaluation, we compare the following systems

- **Baseline**: No translation of paralinguistic information
- **EachLR**: Linear regression with a model for each word
- **AllLR**: A single linear regression model trained on all words
- **AllNN**: A single neural network model trained on all words
- **AllNN-ID**: The AllNN model without additional features

In addition, we use naturally spoken speech as an oracle output.

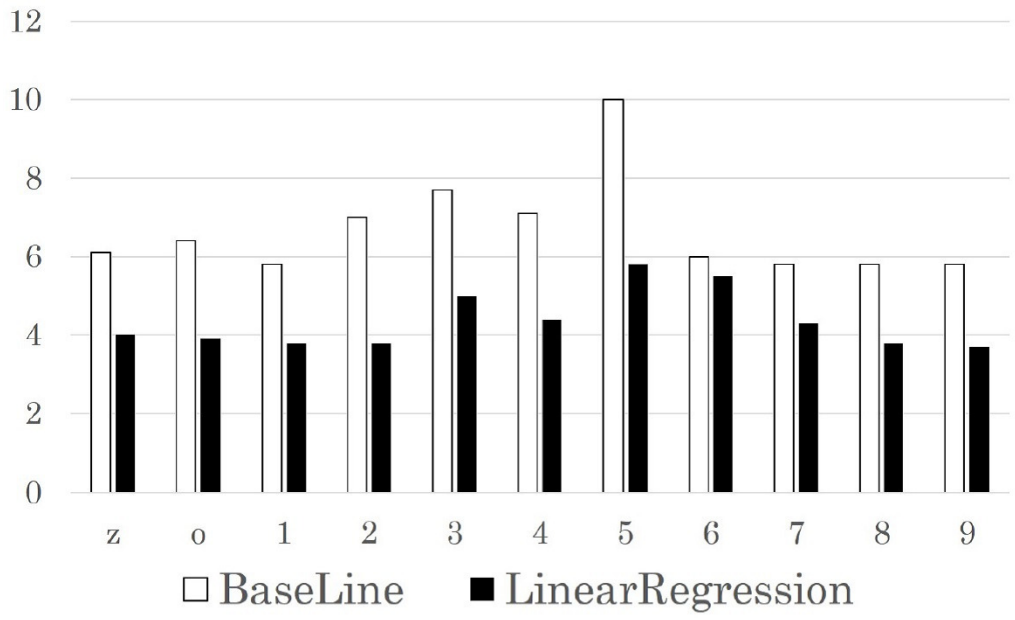

Figure 3: Root mean squared error rate (RMSE) between the reference target duration and the system output for each digit

5.2. Objective Evaluation

We first perform an objective assessment of the translation accuracy of duration and power, the results of which are found in Figure 3 and 4. For each of the nine digits plus "oh" and "zero," we compared the difference between the proposed and baseline duration and power and the reference speech duration and power in terms of RMSE. From these results, we can see that

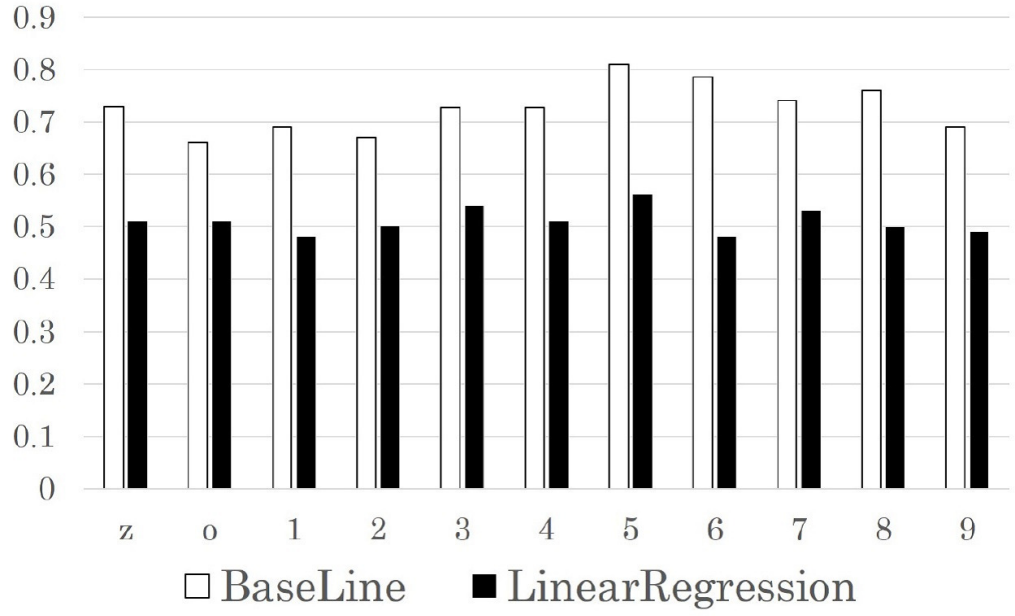

Figure 4: Root mean squared error rate (RMSE) between the reference target power and the system output for each digit

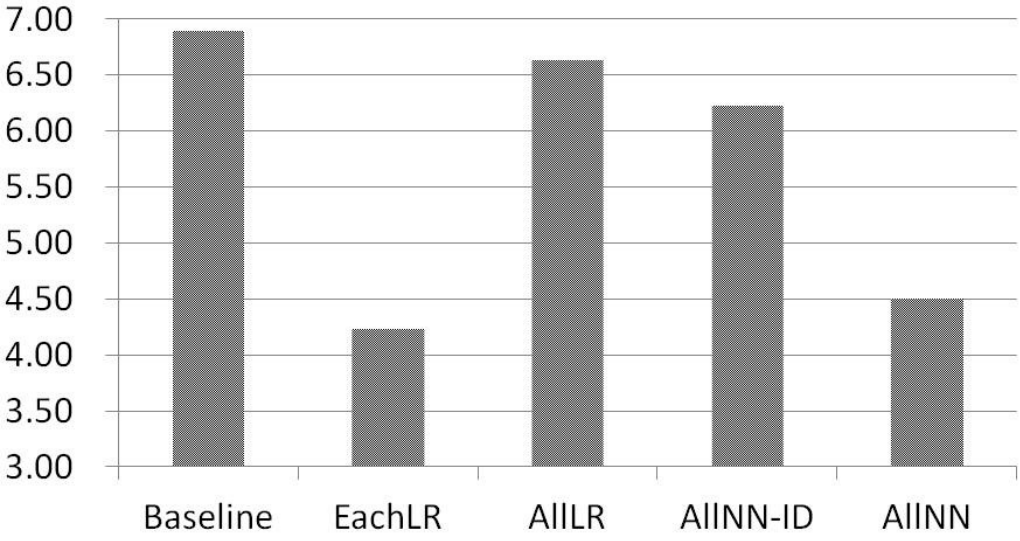

Figure 5: RMSE between the reference and system duration

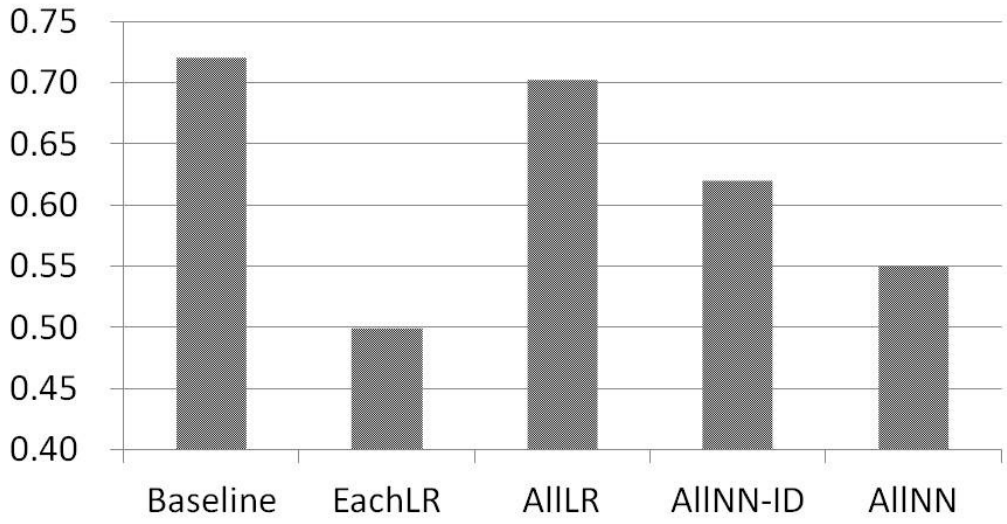

Figure 6: RMSE between the reference and system power

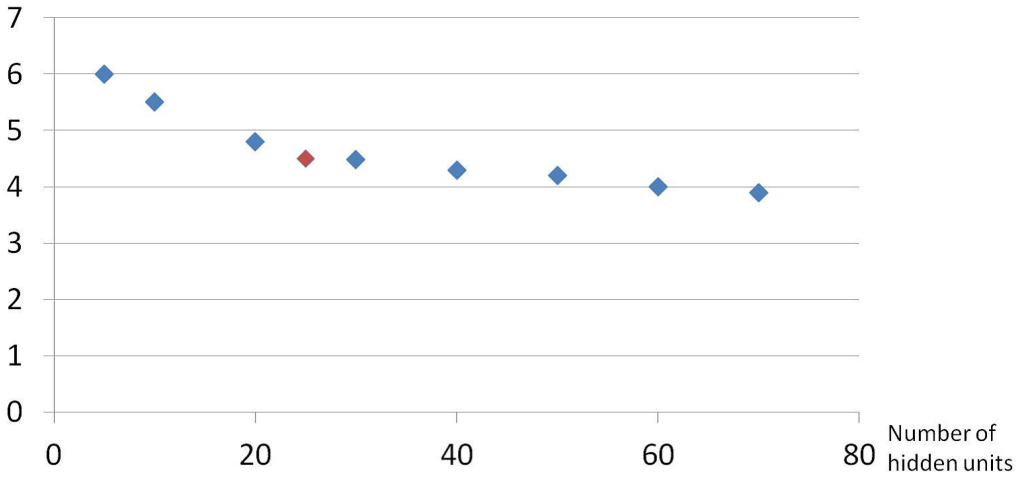

Figure 7: RMSE of duration for each number of NN hidden units

the target speech duration and power output by the proposed method is more similar to the reference than the baseline over all eleven categories, indicating the proposed method is objectively more accurate in translating duration and power. Second we compare the proposed linear regression against the neural network model in Figure 5 and 6. We compared the difference be-

tween the system duration and power and the reference speech duration and power in terms of RMSE. From these results, we can see that the AllLR model is not effective at mapping duration and power information, achieving results largely equal to the baseline. The AllNN model without linguistic information does slightly better but still falls well short of the EachNN baseline. Finally, we can see that our proposed methods outperform baseline and AllNN is able to effectively model translation of paralinguistic information, although accuracy of power lags slightly behind that of duration.

We also show the relationship between the number of NN hidden units and RMSE of duration in Figure 7 (the graph for power was similar). It can be seen that RMSE continues to decrease as we add more units, but with diminishing returns after 25 hidden units. When comparing the number of free parameters in the EachLR model (17*16*11=2992) and the AllNN model with 25 hidden units (28*25+25*16=1100), it can be seen that we were able to significantly decrease the number of parameters as well.

5.3. Subjective Evaluation

As a subjective evaluation we asked native speakers of Japanese to evaluate how well emphasis was translated into the target language for the baseline, oracle, and EachLR and AllNN models when translating duration or duration+power. The first experiment asked the evaluators to attempt to recognize the identities and positions of the emphasized words in the output speech. The overview of the result for the word and emphasis recognition rates is shown in Figure 8. We can see that all of the paralinguistic translation systems show a clear improvement in the emphasis recognition rate over the baseline. There is no significant difference between the linear regression and neural network models, indicating that the neural network learned a paralinguistic information mapping that allows listeners to identify emphasis effectively. The second experiment asked the evaluators to subjectively judge the strength of emphasis with the following three degrees:

- **1**: not emphasized
- **2**: slightly emphasized
- **3**: emphasized

The overview of the experiment regarding the strength of emphasis is shown in Figure 9. This figure shows that all systems show a significant improvement in the subjective perception of strength of emphasis. In this case,

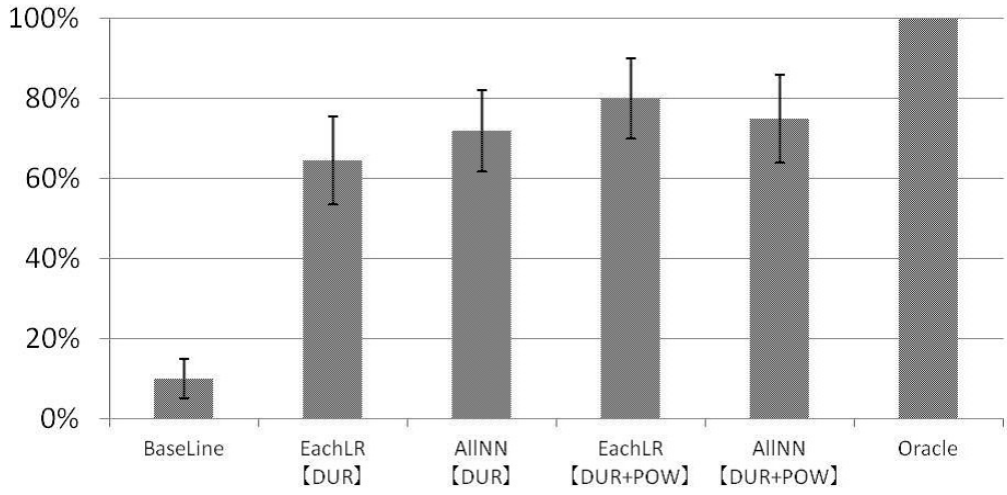

Figure 8: Prediction rate

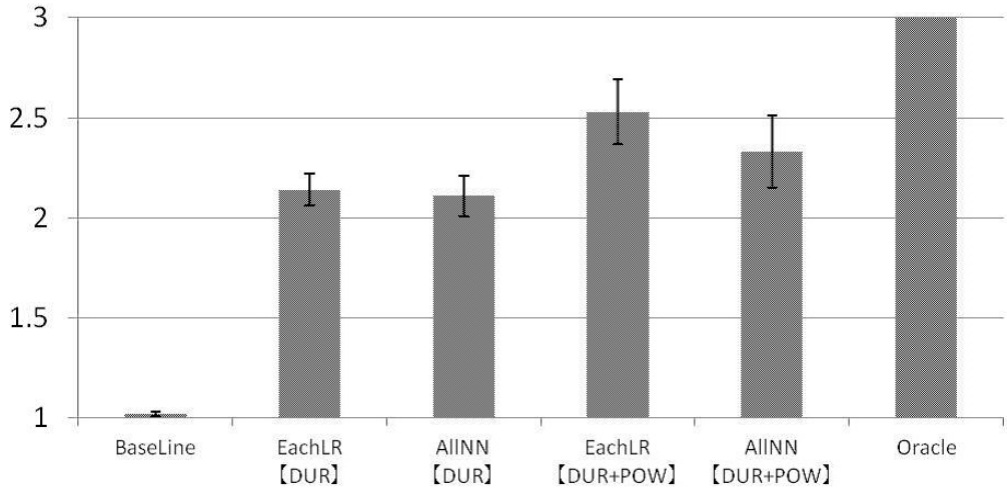

Figure 9: Prediction strength of emphasis

there seems to be a slight subjective preference towards EachLR when power is considered, reflecting the slightly smaller RMSE found in the automatic evaluation. We also performed emphasis translation that only used power, but the generated speech's naturalness was quite low. This resulted in drastic speech volume changes in a short time. Because our proposed method

extracts power features for each frame given by duration information, the power extraction has a high dependency on duration. In this method, if we try to handle other acoustic features (e.g. F0) then we also suspect that we will need to model duration together with these features as well.

6. Related Work

There have been several studies demonstrating improved speech translation performance by translating source side speech non-lexical information to target side speech non-lexical information. Some previous work [9, 16, 7] has focused on the input speech information (for example, phoneme similarity, number of fillers, and ASR parameters) and tried to explore a tight coupling of ASR and MT for speech translation, boosting translation quality as measured by BLEU score. Other related works focus on recognizing speech intonation to reduce translation ambiguity on the target side [20, 22]. These methods consider non-lexical information to boost translation accuracy. However as we mentioned before, there is more to speech translation than just accuracy, and we should consider other features such as the speaker's facial and prosodic expressions.

There is some research that considers translating these expressions and improves speech translation quality in other ways that cannot be measured by BLEU. For example some work focuses on facial information and tries to translate speaker emotion from source to target [19, 15]. On the other hand, [2, 18, 3] focus on the input speechs prosody, extracting F0 from source speech at the sentence level and clustering accent groups. These are then translated into target side accent groups, considering the prosody encoded as factors in a factored translation model [12] to convey prosody from source to target.

In our work, we focus on source speech acoustic features and extract them and translate to target acoustic features directly and continuously. In this framework, we need two translation models. One for the word-to-word translation, and another for acoustic translation. We made acoustic translation models with linear regression for each translation pair. This method is simple, and we can translate acoustic features without having an adverse affect on BLEU score. After this work was originally performed, several related works have modeled emphasis by HMM AMs and calculated emphasis levels and translated the emphasis at the word level [5, 6]. These works expand our work to large vocabulary translation tasks. The major difference of this word and our work is the paralinguistic extraction method. In their work they

handle emphasis as a level between 0-1 that calculates similarity between an HMM AM for emphasized speech and another HMM AM for normal speech. Each word has one emphasis level feature and maps these emphasis levels between input and target sequences. In their work, they need to annotate a paralinguistic label for each type of paralinguistic information they want to handle, and thus if they expand to other varieties of paralinguistic information (e.g. emotion or voice quality) they would need annotated training data to do so. On the other hand, in our work we perform normal ASR to obtain alignments and extract observed features, and do not need to specify specific linguistic labels.

State-of-the-art work on speech translation [4] translates input speech to target words directly with sequential attentional model. In this work they only focus linguistic features on target side and evaluate according to BLEU score. There is also work that focuses on direct speech-to-text translation using sequential attentional models [8, 23]. In this work, any paralinguistic features that exist on the source side may be reflected in the lexical content of the target translations, but paralinguistic information will not be reflected in the target speech.

7. Conclusion

In this paper we proposed a generalized model to translate duration and power information for speech-to-speech translation. Experimental results showed proposed method can model input speech emphasis more effectively than baseline methods. In future work we plan to expand beyond the digit translation task in the current paper to a more general translation task using phrase-based or attention-based neural MT. The difficulty here is the procurement of parallel corpora with similar paralinguistic information for large-vocabulary translation tasks. We are currently considering possibilities including simultaneous interpretation corpora and movie dubs. Another avenue for future work is to expand to other acoustic features such as F0, which play an important part in other language pairs.

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 655–658, Apr 1988.

[2] Pablo Daniel Agüero, Jordi Adell, and Antonio Bonafonte. Prosody generation for speech-to-speech translation. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 557–560, 2006.

[3] Gopala Krishna Anumanchipalli, Luís C. Oliveira, and Alan W. Black. Intent transfer in speech-to-speech machine translation. In *2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, December 2-5, 2012*, pages 153–158, 2012.

[4] Quoc Truong Do, Sakriani Sakti, and Satoshi Nakamura. Toward expressive speech translation: A unified sequence-to-sequence lstms approach for translating words and emphasis. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2640–2644, 2017.

[5] Quoc Truong Do, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Improving translation of emphasis with pause prediction in speech-to-speech translation systems. In *12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, December.

[6] Quoc Truong Do, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Preserving word-level emphasis in speech-to-speech translation using linear regression hsmms. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 3665–3669, 2015.

[7] Markus Dreyer and Yuanzhe Dong. APRO: all-pairs ranking optimization for MT tuning. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1018–1023, 2015.

[8] Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. An attentional model for speech translation without transcription. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 949–959, 2016.

[9] Jie Jiang, Zeeshan Ahmed, Julie Carson-Berndsen, Peter Cahill, and Andy Way. Phonetic representation-based speech translation.

[10] Takatomo Kano, Sakriani Sakti, Shinnosuke Takamichi, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. A method for translation of paralinguistic information. In *International Workshop on Spoken Language Translation*, pages 158–163.

[11] Takatomo Kano, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Generalizing continuous-space translation of paralinguistic information. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 2614–2618, 2013.

[12] Philipp Koehn and Hieu Hoang. Factored translation models. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 868–876, 2007.

[13] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*, 2007.

[14] R. G. Leonard. A database for speaker-independent digit recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '84, San Diego, California, USA, March 19-21, 1984*, pages 328–331, 1984.

[15] Shigeo Morishima and Satoshi Nakamura. Multi-modal translation system and its evaluation. In *4th IEEE International Conference on Multimodal Interfaces (ICMI 2002), 14-16 October 2002, Pittsburgh, PA, USA*, pages 241–246, 2002.

[16] Graham Neubig, Kevin Duh, Masaya Ogushi, Takatomo Kano, Tetsuo Kiso, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. The NAIST machine translation system for IWSLT2012. In *2012 International Workshop on Spoken Language Translation, IWSLT 2012, Hong Kong, December 6-7, 2012*, pages 54–60, 2012.

[17] David Pearce and Hans-Günter Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000*, pages 29–32, 2000.

[18] Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth Narayanan. Enriching machine-mediated speech-to-speech translation using contextual information. *Computer Speech & Language*, 27(2):492–508, 2013.

[19] Éva Székely, Ingmar Steiner, Zeeshan Ahmed, and Julie Carson-Berndsen. Facial expression-based affective speech translation. *J. Multimodal User Interfaces*, 8(1):87–96, 2014.

[20] Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida, Fumiaki Sugaya, Akio Yokoo, and Seiichi Yamamoto. A japanese-to-english speech translation system: ATR-MATRIX. In *The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney Convention Centre, Sydney, Australia, 30th November - 4th December 1998*, 1998.

[21] Tomoki Toda, Alan W. Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio, Speech & Language Processing*, 15(8):2222–2235, 2007.

[22] Wolfgang Wahlster. Robust translation of spontaneous speech: A multi-engine approach. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*, pages 1484–1493, 2001.

[23] Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly transcribe foreign speech. *CoRR*, abs/1703.08581, 2017.

[24] Heiga Zen, Keiichi Tokuda, and Alan W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

[25] J Zhang and Satoshi Nakamura. An efficient algorithm to search for a minimum sentence set for collecting speech database. In *International Congress of Phonetic Sciences*, pages 3145–3148, 2003.