

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 help people communicate across different languages. The most commonly used speech translation model is composed by Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-To-Speech synthesis (TTS) components, in which they are sharing information only in text level. However, spoken communication is different from written communication, as we use rich acoustic cues in order to transmit more information. This paper is concerned with speech-to-speech translation that is sensitive to paralinguistic information. Our long-term goal is to made a system that allows user to speak a foreign language with the same expressiveness as if they were speaking in their own language by reconstructing input acoustic features (F0, duration, spectrum etc.) 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 including regression and Neural Network (NN) model. We evaluate the proposed method and show that paralinguistic information in input speech of source language can appears in output speech of target language.

Keywords: Paralinguistic Information, Speech to speech translation, Emotion

1. Introduction

We speak with 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. It may not be a critical
5 factor in written communication, but in spoken communication it has great importance. Because even if the content of words is the same, if the intonation and facial expression are different an utterance can take an entirely different meaning. Therefore it is necessary to take into account paralinguistic factor in any systems that are constructed to augment human-to-human communication.

10 Speech-to-speech translation system is one of technologies that help people communicate across different languages. 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, but the words output by ASR is in written
15 form that have lost all prosody information. The words output by TTS will then be given the canonical prosody for the input text, not reflecting these traits. Thus, information sharing between the ASR, MT, and TTS modules is weak, and after ASR source-side acoustic details are lost (for example: speech rhythm, emphasis, or emotion).

20 This paper is concerned with speech-to-speech translation that is sensitive to paralinguistic information. Our long-term goal is to made a system that allows user to speak a foreign language with the same expressiveness as if they were speaking in their own language by reconstructing input acoustic features (F0, duration, spectrum etc.) in the target language. From the many different
25 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.

First, we extract features at the level of Hidden Markov Model (HMM) states, and use linear regression to translate them to the duration and power of HMM states of the output speech. Furthermore, we also expand the paralinguistic translation model to adapt to more general tasks by training a single model that is applicable to all words using neural networks. There are two merits to using neural networks. First, neural network possess sufficient power to express difficult regression problems such as translation of acoustic features for multiple words. Second, neural network can be expanded with features expressing additional information such as the input word and translated word, the position of both words, parts of speech, and so on. We perform experiments that use this technique to translate paralinguistic features and reconstruct the input speech's paralinguistic information, particularly emphasis, in output speech.

2. Conventional Speech-to-Speech Translation

In Conventional Speech-to-Speech, ASR module decode text of utterance from input speech. Now acoustic feature represent as $\mathbf{X} = [x_1, x_2 \dots x_T]$ and spoken word represent as $\mathbf{E} = [f_1 f_2 \dots f_N]$ then the probability is $P(\mathbf{E} | \mathbf{X})$. ASR system decode \mathbf{E} that maximize $P(\mathbf{E} | \mathbf{X})$. $P(\mathbf{E} | \mathbf{X})$ can convert by Bayes' theorem as below

$$P(\mathbf{E} | \mathbf{X}) = \frac{P(\mathbf{X} | \mathbf{E})P(\mathbf{E})}{P(\mathbf{X})} \quad (1)$$

From point of \mathbf{E} view $P(\mathbf{X})$ is a constant value. We can covert equation as

$$P(\mathbf{E} | \mathbf{X}) \propto P(\mathbf{X} | \mathbf{E})P(\mathbf{E}) \quad (2)$$

Then $P(\mathbf{X} | \mathbf{E})$ is Acoustic Model(AM) and $P(\mathbf{E})$ is Language Model(LM).

MT module decode target words sequence \mathbf{J} that maximize probability $P(\mathbf{J}|\mathbf{E})$ given \mathbf{E} .

$$\hat{\mathbf{J}} = \operatorname{argmax} P(\mathbf{J}|\mathbf{E}) \quad (3)$$

As same as ASR we can convert $P(\mathbf{J}|\mathbf{E})$ as below.

$$\hat{\mathbf{J}} \propto \operatorname{argmax} \frac{P(\mathbf{E}|\mathbf{J})P(\mathbf{J})}{P(\mathbf{E})} \quad (4)$$

Then $P(\mathbf{E}|\mathbf{J})$ is a translation model.

TTS module generate speech parameter $\mathbf{O} = o_1, o_2, \dots o_T$, T is the length of \mathbf{O} , given HMM AM $\lambda = [\lambda_1, \lambda_2 \dots \lambda_N]$ that represent J . The out put $\mathbf{O} = o_1, o_2, \dots o_T$ can be represent by

$$\hat{\mathbf{O}} = \operatorname{argmax}_{\mathbf{O}} P(\mathbf{O} | \lambda, T) \quad (5)$$

In these three module, they share information as \mathbf{E} or \mathbf{J} , so that the input non-lexical information of \mathbf{E} lost through ASR.

3. Speech translation considering paralinguistic information

For achieve paralinguistic information translation, we need consider how to handle paralinguistic features such as \mathbf{X} . In ASR and TTS module, phoneme is a smallest lexical unit that represent speech. And in MT module, a word is a smallest unit of system. From point of speech processing phoneme is a good segment to handle paralinguistic features but in human speaking we usually speak emotionally such as emphasis, surprise and sadness in word, phrase and sentence level. So we consider word is better segment to learn these prosody mapping rule between source to target speech. We make the word AM for each word and perform ASR and TTS. Then we extract the acoustic features \mathbf{X} belong to each words and translate \mathbf{X} through ASR and translate acoustic feature from source to target directory by regression model in MT part. Finally we use translated acoustic features to generate output speech's in TTS part.

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

$$\hat{\mathbf{E}}, \hat{\mathbf{X}} = \operatorname{argmax}_{\mathbf{E}, \mathbf{X}} P(\mathbf{E}, \mathbf{X} | S), \quad (6)$$

where S indicates the input speech, \mathbf{E} indicates the words included in the utterance and \mathbf{X} indicates paralinguistic features of the words in \mathbf{E} . In order to

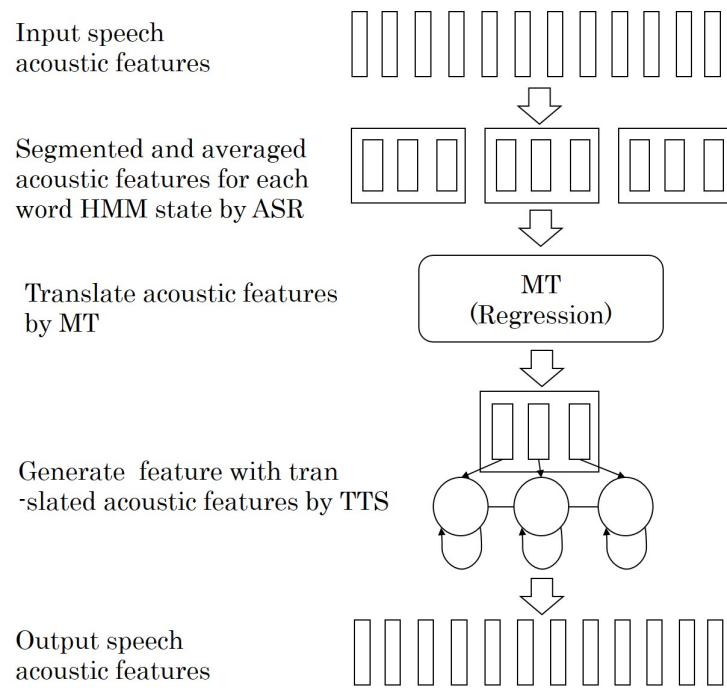


Figure 1: Overview of proposed method

recognize this information, we construct a word-based HMM acoustic model. The acoustic model is trained with audio recordings of speech and the corresponding transcriptions \mathbf{E} using the standard Baum-Welch algorithm. Once we have created our model, we perform simple speech recognition using the HMM
65 acoustic model and a language model that assigns a uniform probability to all digits. Viterbi decoding can be used to find \mathbf{E} . Finally we can decide the duration vector \mathbf{x}_i of each word e_i based on the time spent in each state of the HMM acoustic model 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
70 power value of each feature over frames that are aligned to the same state of the acoustic model. 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. In ASR part, we don't need labeling the prosody of speech we just segment each words and extract observed acoustic feature.

75 3.2. Lexical Translation

Lexical translation finds the best translation \mathbf{J} of sentence \mathbf{E} .

$$\hat{\mathbf{J}} = \underset{\mathbf{J}}{\operatorname{argmax}} P(\mathbf{J}|\mathbf{E}), \quad (7)$$

where \mathbf{J} indicates the target language sentence and \mathbf{E} indicates the recognized source language sentence. Generally we can use a 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.

80 3.3. Paralinguistic Translation

Paralinguistic translation converts the source-side acoustic features vector \mathbf{X} into the target-side acoustic features vector \mathbf{Y} according to the following equation

$$\hat{\mathbf{Y}} = \underset{\mathbf{Y}}{\operatorname{argmax}} P(\mathbf{Y}|\mathbf{X}). \quad (8)$$

In particular, we control duration and power of each word using a source-side duration and power super vector $\mathbf{x}_i = [\mathbf{x}_1, \dots, \mathbf{x}_{N_x}]^\top$ and a target-side duration

and power super vector $\mathbf{y}_i = [\mathbf{y}_1, \dots, \mathbf{y}_{N_y}]^\top$. 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 $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_i, \dots, \mathbf{y}_I]$ where I is the length of the sentence. 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 \mathbf{Y} using the following equation:

$$\hat{\mathbf{Y}} = \underset{\mathbf{Y}}{\operatorname{argmax}} \prod_i P(\mathbf{y}_i | \mathbf{x}_i). \quad (9)$$

The word-to-word acoustic translation probability $P(\mathbf{y}_i | \mathbf{x}_i)$ is defined according to linear regression matrix that indicates that \mathbf{y}_i is distributed according to a normal distribution

$$P(\mathbf{y}_i | \mathbf{x}_i) = N(\mathbf{y}_i; \mathbf{W}_{e_i, j_i} \mathbf{x}'_i, S) \quad (10)$$

where \mathbf{x}' is $[1x^\top]^\top$ and \mathbf{W}_{e_i, j_i} is a regression matrix (including a 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 word pairs $\langle e, j \rangle$ we want to translate. In order to do so, we optimize each regression matrix on the translation model training data for $\langle e, j \rangle$ by minimize root mean squared error (RMSE) with a regularization term

$$\hat{\mathbf{W}}_{e,j} = \underset{\mathbf{W}_{e,j}}{\operatorname{argmin}} \sum_{n=1}^N \|\mathbf{y}_n^* - \mathbf{y}_n\|^2 + \alpha \|\mathbf{W}_{e,j}\|^2, \quad (11)$$

where N is the number of training samples for the word pair, n is the ID of each training sample, \mathbf{y}^* is target language reference word duration and power vector, and α is a hyper-parameter for the regularization term to prevent over-fitting.¹ This maximization can be solved in closed form using simple matrix operations.

¹We chose α to be 10 based on preliminary tests but the value had little effect on subjective results.

3.4. Speech Synthesis

In the TTS part of the system we use an HMM-based speech synthesis system [1], and reflect the duration and power information of the target word
 100 paralinguistic information vector onto the output speech. The output speech parameter vector sequence $\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_T]^\top$ is determined by maximizing the target HMM likelihood function given the target word duration and power vector $\hat{\mathbf{Y}}$ and the target language sentence $\hat{\mathbf{J}}$ as follows:

$$\hat{\mathbf{C}} = \underset{\mathbf{C}}{\operatorname{argmax}} P(\mathbf{O}|\hat{\mathbf{J}}, \hat{\mathbf{Y}}) \quad (12)$$

$$\text{subject to } \mathbf{O} = \mathbf{M}\mathbf{C}, \quad (13)$$

where \mathbf{O} is a joint static and dynamic feature vector sequence of the target
 105 speech parameters and \mathbf{M} is a transformation matrix from the static feature vector sequence into the joint static and dynamic feature vector sequence. While TTS generally uses phoneme-based HMM models, we instead used a word based HMM to maintain the consistency of feature extraction and translation. In this task the vocabulary is small, so we construct an independent context model.

110 4. End to End models of Paralinguistic feature translation methods

In this section we describe two ways to translate paralinguistic features of the source words to target words. One is simple linear regression another is Neural Network with word embed vector.

4.1. Linear regression models

Paralinguistic translation converts the source-side paralinguistic feature \mathbf{X} into the target-side paralinguistic feature \mathbf{Y} following Voice Conversion ideas [2, 3, 4]

$$\hat{\mathbf{Y}} = \underset{\mathbf{Y}}{\operatorname{argmax}} P(\mathbf{Y}|\mathbf{X}) \quad (14)$$

115 In particular, we control duration and power using source-side word feature vector $\mathbf{x}_i = [x_1, \dots, x_{N_x}]^\top$ and target-side word feature vector $\mathbf{y}_i = [y_1, \dots, y_{N_y}]^\top$. In these vectors N_x represents the number of HMM states in source side and

N_y represents the number of HMM states in target side. The sentence feature vector consists of the concatenation of the word duration and power vectors such
 120 as $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_i, \dots, \mathbf{y}_I]$ 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:

$$\hat{\mathbf{Y}} = \underset{\mathbf{Y}}{\operatorname{argmax}} \prod_i P(\mathbf{y}_i | \mathbf{x}_i). \quad (15)$$

This can be defined with any function, but we choose to use linear regression, which indicates that \mathbf{y}_i is distributed according to a normal distribution

$$P(\mathbf{y}_i | \mathbf{x}_i) = N(\mathbf{y}_i; \mathbf{W}_{e_i, j_i} \mathbf{x}'_i, S) \quad (16)$$

125 where, \mathbf{x}' is $[1 \mathbf{x}^\top]^\top$ and \mathbf{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 root mean
 130 squared error (RMSE) with a regularization term

$$\hat{\mathbf{W}}_{e,j} = \underset{\mathbf{W}_{e_i, j_i}}{\operatorname{argmax}} \sum_{n=1}^N \|\mathbf{y}^*_n - \mathbf{y}_n\|^2 + \alpha \|\mathbf{W}_{e_i, j_i}\|^2, \quad (17)$$

where N is the number of training samples, n is an id of a training sample, \mathbf{y}^* is target language reference word duration and power vector, and α is a hyper-parameter for the regularization term.²

Linear regression model need train a regression matrix for each word pair
 135 $\langle e, j \rangle$. 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 $\mathbf{W}_{e,j}$ into a single global regression matrix \mathbf{W} and training the matrix over all samples in

²We chose α to be 10 based on preliminary tests.

the corpus. However, this model can be expected to not be expressive enough
 140 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.

145 4.2. Global Neural Network Models

As a solution to the problem of the lack of expressibility in linear regression,
 we 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 the task. In addition,
 150 they allow for adding features for many different types of information following
 ASR, MT and TTS’s common practice, such 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 [1],
 so we can likely improve estimation of the output duration and power vector in
 155 translation as well.

In this research, we prepare a feed forward neural network that proposes the
 best output word acoustic feature vector $\hat{\mathbf{Y}}$ given input word acoustic feature
 vector \mathbf{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,
 160 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 ι_i equal to the
 input vector value:

$$\iota_i = x_i. \quad (18)$$

The hidden units π_j are calculated according to the input-hidden unit weight

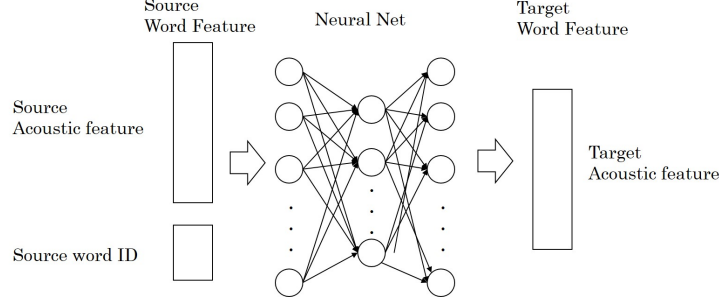


Figure 2: Neural Network for acoustic feature translation

matrix \mathbf{W}^h :

$$\pi_j = \frac{1}{1 + \exp(-\alpha \sum_i w_{ij}^h t_i)}, \quad (19)$$

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

$$\psi_k = \sum_j w_{jk}^o \pi_j \quad (20)$$

$$y_k = \psi_k, \quad (21)$$

where \mathbf{W}^o is hidden-output unit weight matrix. As an optimization criterion we use minimization of RMSE, which is achieved through simple back propagation.

5. Evaluation

5.1. Experimental Setting

We examine the effectiveness of the proposed method through English-Japanese speech-to-speech translation experiments, summarized in Table 1. 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 number, the speaker re-speaks, emphasizing the mistaken number. 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

ASR	
Tool	HTK
Training sentences	8440
HMM states	16
MT	
Tool	Moses
Training utterances	445
Test utterances	55
Neural net structure	29/25/16
TTS	
Tool	HTS
Training utterances	445
HMM states	16

Table 1: Experimental Settings

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 lexical content to be spoken was 500 sentences from the AURORA2 data set, chosen to be word balanced by greedy search [5]. The training set is 445 utterances and the test set is 55 utterances.³

We further used this data to build an English-Japanese speech translation system that include our proposed paralinguistic translation model. We used the AURORA2 8440 utterance bilingual speech corpus to train the ASR module. Speech signals were sampled at 8kHz with utterances from 55 males and 55 females. We set the number of HMM states per word in the ASR acoustic model to 16, the shift length to 5ms, and other various settings for ASR to follow [6][7]. The original Aurora2 has 8440 utterances for training and 4004 utterances for testing. Here we don't use the original testing part in our experiments.

³Freely available at <http://www.phontron.com/pcbue>

Base line V.S. Linear Regression	
None	No translation of paralinguistic information
EachLR	Linear regression with a model for each word

Table 2: Compare baseline against prosed linear regression for each words

Proposed LR V.S. NN model	
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

Table 3: Compare proposed Linear Regression against Neural Network model for all words

To simplify the problems, the experiments were done where the ASR has no
 190 error. Therefore we selected 500 balanced sentence from the 8440 utterances of
 training data, and divide the utterances into 445 utts for training set and 55
 utts of testing for the paralinguistic translation. So the achieve 100% accuracy
 is because it is a closed test set. For TTS, we use the same 445 utterances
 for training an independent context synthesis model. In this case, the speech
 195 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 two baselines Table 2 and pro-
 posed three global models of paralinguistic translation each other Table 3.

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

200 5.2. Objective Evaluation

We first perform an objective assessment of the translation accuracy of du-
 ration and power, the results of which are found in Figure 3, 4, 5 and Figure
 6. We compared the difference between the system duration and power and the
 reference speech duration and power in terms of RMSE.

205 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

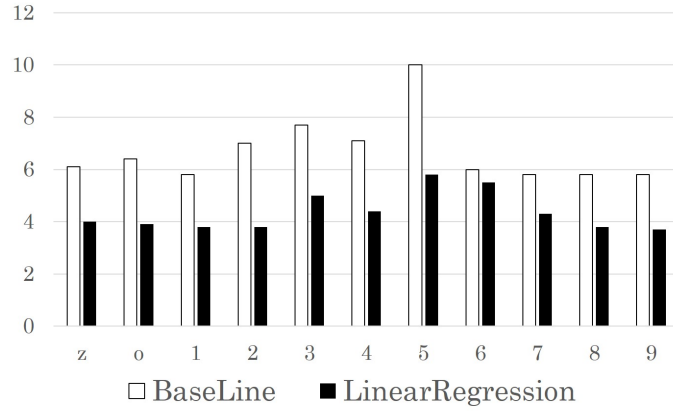


Figure 3: Root mean squared error rate (RMSE) between the reference target duration and the system output for each digit

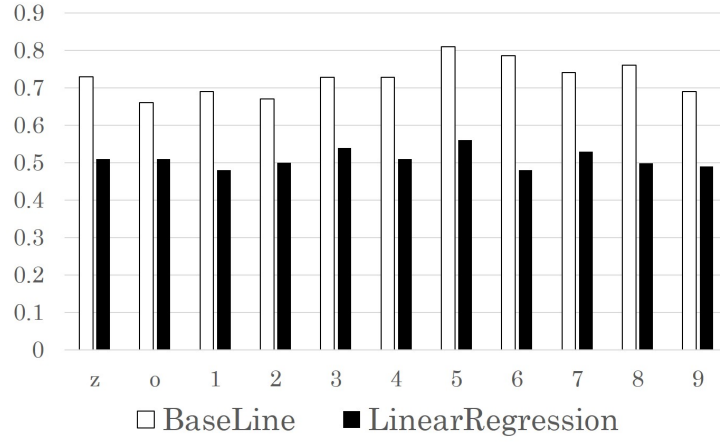


Figure 4: Root mean squared error rate (RMSE) between the reference target power and the system output for each digit

baseline. The AllNN model without linguistic information does slightly better but still falls well short of the EachNN baseline. Finally, we can see that AllNN is able to effectively model translation of paralinguistic information, although accuracy of power lags slightly behind that of duration.

210

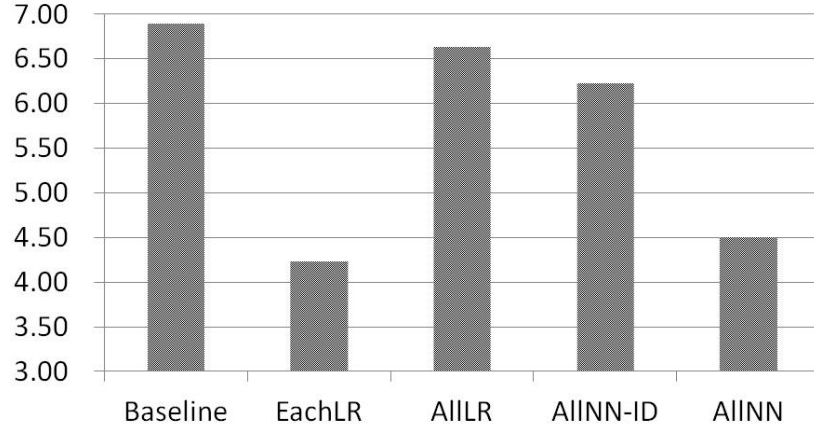


Figure 5: RMSE between the reference and system duration

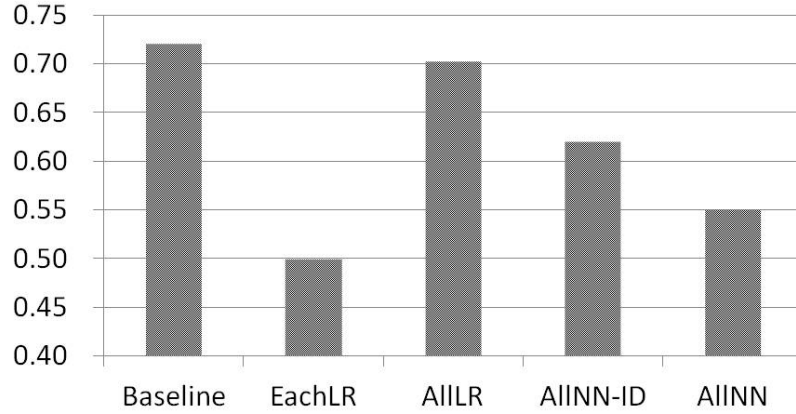


Figure 6: RMSE between the reference and system power

We also show the relationship between the number of NN hidden units and RMSE of duration in 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

215 EachLR model ($17 \times 16 \times 11 = 2992$) and the AllNN model with 25 hidden units ($28 \times 25 + 25 \times 16 = 1100$), it can be seen that we were able to significantly decrease

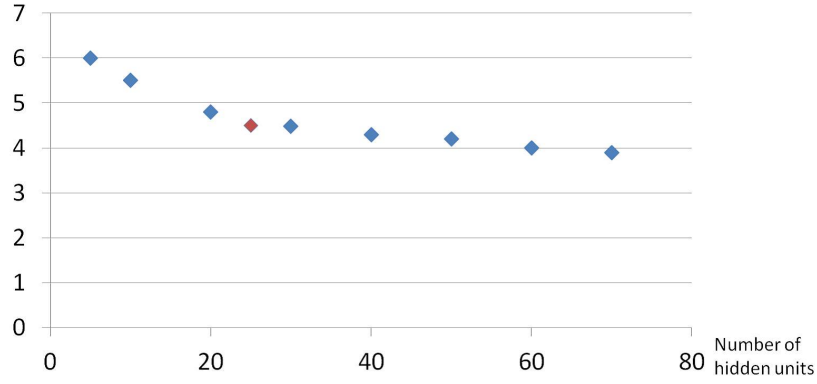


Figure 7: RMSE of duration for each number of NN hidden units

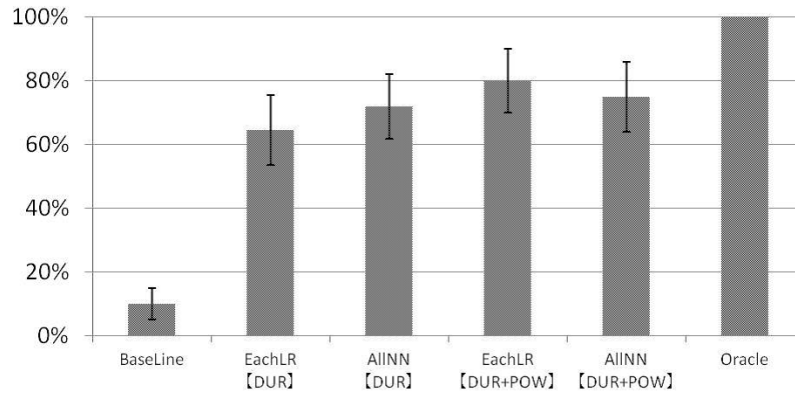


Figure 8: Prediction rate

the number of parameters with little change in accuracy.

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 base-line, oracle, and EachLR and AllNN models when translating duration or duration+power.

The first experiment asked the evaluators to attempt to recognize the identi-

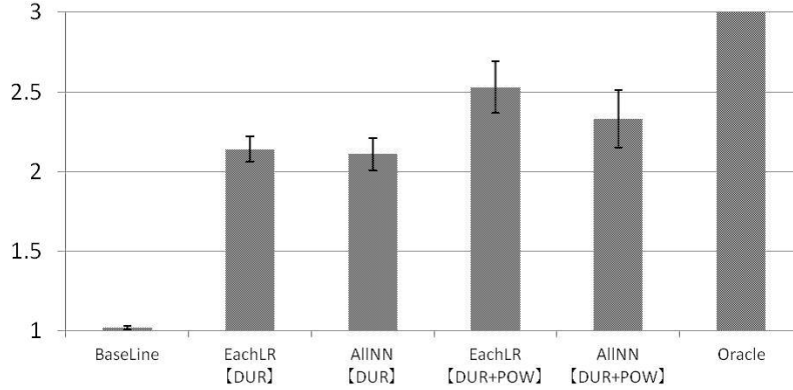


Figure 9: Subjective degree of emphasis

ties 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 Linear Regression and Neural Network model, in-
dicating 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, there seems
to be a slight subjective preference towards EachLR when power is considered,
reflecting the slightly larger RMSE found in the automatic evaluation. We

also performed emphasis translation only use power, but generated speech's naturalness is quite low. There are the value of speech volume drastic changes in short time. Because of our proposed method process we extract power feature for each frames that given by duration information so that the power extraction
 245 has high dependency on duration. In this method, if we try to handle other acoustic feature(example F0) then we need to modeling duration together.

6. Related Works

There have been several studies demonstrating improved speech translation performance by translation source side speech non-lexical information to target
 250 side speech non-lexical information. In previous work, [8, 9, 10] focus for input speech's acoustic information and try to explore a tight coupling of ASR and MT for speech translation with sharing information, they boost translation quality as measured by BLEU score. Other related works focus on speech intonation recognizing using that to reduce translation ambiguity on target side[11, 12].
 255 These methods consider non-lexical information to boost translation accuracy. However as we mentioned before, there is more to speech translation than just accuracy. We should consider other features such as the speakers facial and speech expressions. There is some research that considers translating these expressions and improves speech translation quality in other ways that cannot
 260 be measured by BLEU. For example some work focuses on mouth shape and tries to translate speaker emotion from source to target.[13, 14]. On the other hand, [15, 16] focus input speech's prosody, extracting F0 from source speech at the sentence level and clustering accent groups. These are then translated into target side accent groups. consider the prosody encoded as factors in a factored
 265 translation model [17] 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 frame work, we need two translation models. One for word to word translation model, another for acoustic translation model. We made acoustic translation

270 model with linear regression for each translation pair. In this work our proposed method can translate acoustic feature with consider the similarity and continuously. This method is very simple and we can translate acoustic feature(duration) without decline BLEU score.

After this work, there are published related work that modeling emphasis by HMM acoustic model and calculate emphasis level and translation the emphasis level in word level[18, 19]. And this method can handle power and duration value continuously but for make emphasis acoustic model they need to annotate speech. We often have a different impression even we hear same speech. So that annotate for various prosody and emotion sometime not stable because it depends human feeling, In these related works, when they apply their method for multi-prosody and emotion their will be faced para-speech data collection issue. But this work did not need any annotate for speech, in point of data preparation view we can easily adapt our method for multi-prosody and emotion.

7. Conclusion

285 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 modeling input speech emphasis more than baseline. But we also see if we failed to regression target speech feature then we generate worth quality of speech. This issue can be say machine translation error of paralinguistic information. We should care this issue by make threshold or modeling the naturalness of paralinguistic feature like lexical MT.

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 SMT. 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.

8. Acknowledge

300 Part of this research was supported by JSPS KAKENHI Grant Number
24240032 and 26870371.

References

- [1] H. Zen, K. Tokuda, A. W. Black, Statistical parametric speech synthesis,
Speech Communication.
- 305 [2] T.Toda, A. .W.Black, K.Tokuda, Voice conversion based on maximum
likelihood estimation of spectral parameter trajectory, IEEE Trans. ASLP
15 (8) (2007) 2222–2235.
- [3] M.Abe, S.Nakamura, K.Shikano, H.Kuwabara, Voice conversion through
vector quantization, J. Acoust. Soc. Jpn (E) 11 (2) (1990) 71–76.
- 310 [4] Y.Stylianou, O.Cappé, E.Moulines, Continuous probabilistic transform for
voice conversion, IEEE Trans. SAP 6 (2) (1998) 131–142.
- [5] J. Zhang, S. Nakamura, An efficient algorithm to search for a minimum
sentence set for collecting speech database, in: Proceedings of the 15th
International Congress of Phonetic Sciences, 2003.
- 315 [6] H. G. Hirsh, D. Pearce, The AURORA experimental framework for the per-
formance evaluations of speech recognition systems under noisy conditions,
in: ISCA ITRW ASR2000, 2000.
- [7] R. Leonard, A database for speaker independent digit recognition, in: Pro-
ceedings of ICASSP, 1984.
- 320 [8] J. Jiang, Z. Ahmed, J. Carson-Berndsen, P. Cahill, A. Way, Phonetic
representation- based speech translation, in: Machine Translation Summit,
2011.

- [9] M. Ohgushi, G. Neubig, S. Sakti, T. Toda, S. Nakamura, An empirical comparison of joint optimization techniques for speech translation, in: 14th Annual Conference of the International Speech Communication Association (InterSpeech 2013), Lyon, France, 2013, pp. 2619–2622.
- [10] M. Dreyer, Y. Dong, Apro: All-pairs ranking optimization for mt tuning, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 1018–1023.
URL <http://www.aclweb.org/anthology/N15-1106>
- [11] Y. S. N. C. H. I. F. S. A. Y. Toshiyuki Takezawa, Tsuyoshi Morimoto, S. Yamamoto, A japanese-to- english speech translation system: Atr-matrix, in: In Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP, Sydney,Australia, 1998, p. pp. 957960.
- [12] W. Wahlster, Robust translation of spontaneous speech: a multi-engine approach, in: IJCAI'01 Proceedings of the 17th international joint conference on Artificial intelligence - Vol 2, IJCAI, Washington,America, 2001, pp. p 1484–1493.
- [13] S. N. Shin Ogata, Takafumi Misawa, S. Morishima, Muti- modal translation system by using automatic facial image tracking and model- based lip synchronization, in: ACM SIGGRAPH2001 Conference Abstracts and Applications,Sketch and Applications, Siggraph, Los Angeles, 2001, p. p231.
- [14] S. Nakamura, Y. Sasaki, Toward common evaluation frameworks for speech recognition and speech translation technologies, The Journal of the Acoustical Society of Japan.
- [15] P. D. Agüero, J. Adell, A. Bonafonte, Prosody generation for speech-to-speech translation, in: Proceedings of ICASSP, 2006.

- 350 [16] V. Kumar, S. Bangalore, S. Narayanan, Enriching machine-mediated
speech-to-speech translation using contextual information, *Computer
Speech and Language*.
- [17] P. Koehn, H. Hoang, Factored translation models, in: *Proceedings of
EMNLP*, 2007, pp. 868–876.
- 355 [18] Q. T. Do, S. Sakti, G. Neubig, T. Toda, S. Nakamura, Improving trans-
lation of emphasis with pause prediction in speech-to-speech translation
systems, in: *12th International Workshop on Spoken Language Transla-
tion (IWSLT)*, Da Nang, Vietnam, 2015.
- [19] Q. T. Do, S. Takamichi, S. Sakti, G. Neubig, T. Toda, S. Nakamura, Pre-
360 serving word-level emphasis in speech-to-speech translation using linear re-
gression HSMs, in: *16th Annual Conference of the International Speech
Communication Association (InterSpeech 2015)*, Dresden, Germany, 2015.