# A Method for Translation of Paralinguistic Information

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

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

## Abstract

This paper is concerned with speech-to-speech translation that is sensitive to paralinguistic information. 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 a regression model that maps source language duration and power information into the target language. We evaluate the proposed method on a digit translation task and show that paralinguistic information in input speech appears in output speech, and that this information can be used by target language speakers to detect emphasis.

## 1. Introduction

In human communication, speakers use many different varieties of information to convey their thoughts and emotions. For example, great speakers enthrall their listeners by not only the contents of the speech but also their zealous voice and confident looks. This paralinguistic information is not a factor in written communication, but in spoken communication it has great importance. These acoustic and visual cues transmit additional information that cannot be expressed in words. Even if the context is the same, if the intonation and facial expression are different an utterance can take an entirely different meaning [1, 2].

However, the most commonly used speech translation model is the cascaded approach, which treats Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-to-Speech (TTS) as black boxes, and uses words as the basic unit for information sharing between these three components. There are several major limitations of this approach.

For example, it is widely known that errors in the ASR stage can propagate throughout the translation process, and considering several hypotheses during the MT stage can improve accuracy of the system as a whole [3]. Another less noted limitation, which is the focus of this paper, is that the input of ASR contains rich prosody information, but the words output by ASR have lost all prosody information. 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).

In our research we explore a speech-to-speech translation system that not only translates linguistic information, but also paralinguistic speech information between source and target utterances. Our final goal is to allow the user to speak a foreign language like a native speaker by recognizing the input acoustic features (F0, duration, power, spectrum etc.) so that we can adequately reconstruct these details in the target language.

From the many different possible paralinguistic features to handle, in this paper we chose duration and power. We propose a method that can translate these paralinguistic features from the input speech to the output speech in continuous space. In this method, 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. We perform experiments that use this technique to translate paralinguistic features and reconstruct the input speech's paralinguistic information, particularly emphasis, in output speech.

We evaluate the proposed method by recording parallel emphasized utterances and using this corpus to train and test our paralinguistic translation model. We measure the emphasis recognition rate and intensity by objective and subjective assessment, and find that the proposed paralinguistic translation method is effective in translating this paralinguistic information.

## 2. Conventional Speech-to-Speech Translation

Conventionally, speech to speech translation is composed of ASR, MT, and TTS. First, ASR finds the best source language sentence $\mathbf{E}$ given the speech signal $S$,

$$\hat{\mathbf{E}} = \arg\max_{\mathbf{E}} P(\mathbf{E}|S). \tag{1}$$

Second, MT finds the best target language sentence $\mathbf{J}$ given the sentence $\mathbf{E}$,

$$\hat{\mathbf{J}} = \arg\max_{\mathbf{J}} P(\mathbf{J}|\hat{\mathbf{E}}). \tag{2}$$

Finally, TTS finds finds the best target language speech parameter vector sequence $\mathbf{C}$ given the sentence $\hat{\mathbf{J}}$,

$$\hat{\mathbf{C}} = \arg\max_{\mathbf{C}} P(\mathbf{O}|\hat{\mathbf{J}}) \tag{3}$$

$$\text{subject to } \mathbf{O} = \mathbf{MC}, \tag{4}$$

where $\mathbf{O}$ is a joint static and dynamic feature vector sequence of the target speech parameters and $\mathbf{M}$ is a transformation matrix from the static feature vector sequence into the joint static and dynamic feature vector sequence.

It should be noted that in the ASR step here we are translating speech S, which is full of rich acoustic and prosodic cues, into a simple discrete string of words $\mathbf{E}$. As a result, in conventional systems all of the acoustic features of speech are lost during recognition, as shown in Figure 1. These features include the gender of the speaker, emotion, emphasis, and rhythm. In the TTS stage, acoustic parameters are generated from the target sentence and training speech only, which indicates that they will reflect no feature of the input speech.

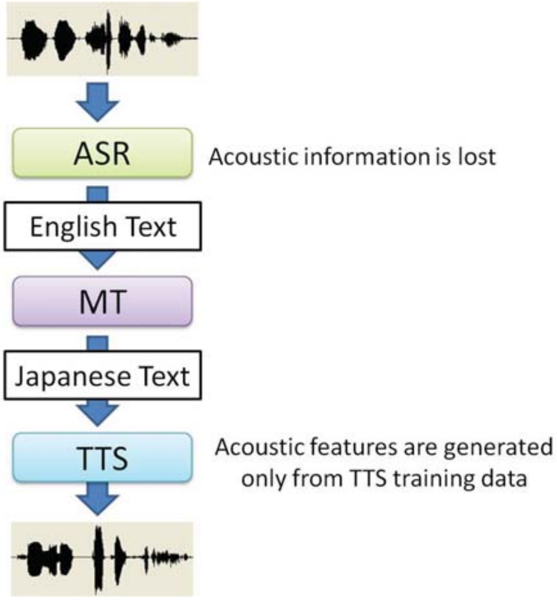
ASR
Acoustic information is lost
English Text
MT
Japanese Text
TTS
Acoustic features are generated only from TTS training data

Figure 1: Conventional speech to speech translation model

## 3. Acoustic Feature Translation Model

In order to resolve this problem of lost acoustic information, we propose a method to translate paralinguistic features of the source speech into the target language. Our proposed method consists of three parts: word recognition and feature extraction with ASR, lexical and paralinguistic translation with MT and linear regression respectively, and speech synthesis with TTS. While this is the same general architecture as traditional speech translation systems, we add an additional model to translate not only lexical information but also two types of paralinguistic information: duration and power. In this paper, in order to focus specifically on paralinguistic translation we chose a simple, small-vocabulary lexical MT task: number-to-number 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

$$\hat{\mathbf{E}}, \hat{\mathbf{X}} = \arg\max_{\mathbf{E},\mathbf{X}} P(\mathbf{E}, \mathbf{X}|S), \tag{5}$$

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 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 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 and power vector $\boldsymbol{x_i}$ of each word $e_i$. The duration component of the vector is chosen based on the time spent in each state of the HMM acoustic model in the path found by the Viterbi algorithm. For example, if word $e_i$ is represented by the acoustic model $A$, the duration component will be a vector with length equal to the number of HMM states representing $e_i$ in $A$, with each element being an integer representing the number of frames emitted by each state. The power component of the vector is chosen in the same way, and we take the mean 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.

### 3.2. Lexical Translation

Lexical translation is defined as finding the best translation $\mathbf{J}$ of sentence $\mathbf{E}$.

$$\hat{\mathbf{J}} = \arg\max_{\mathbf{J}} P(\mathbf{J}|\mathbf{E}), \tag{6}$$

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 tool like Moses [4], to obtain this translation in standard translation tasks. However in this paper we have chosen a simple number-to-number translation task so we can simply write one-to-one lexical translation rules with no loss in accuracy.

### 3.3. Paralinguistic Translation

Paralinguistic translation converts the source-side duration and mean power vector $\mathbf{X}$ into the target-side duration and mean power vector $\mathbf{Y}$ according to the following equation

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X}). \tag{7}$$

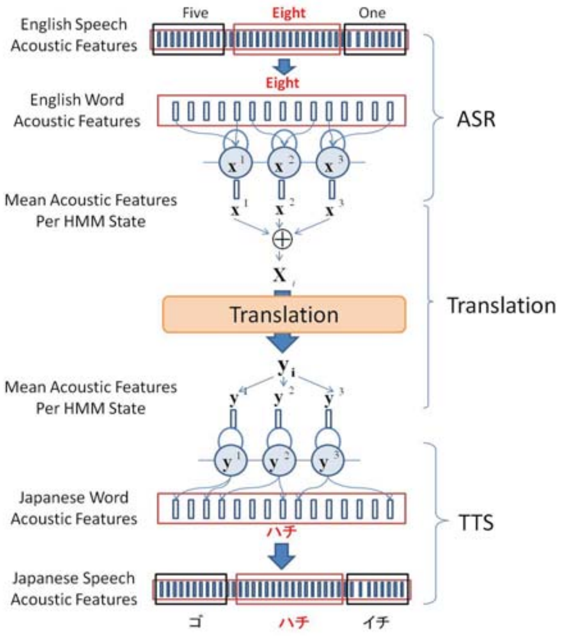
English Speech Acoustic Features
Five
Eight
One
Eight
English Word Acoustic Features
ASR
Mean Acoustic Features Per HMM State
Translation
Translation
Mean Acoustic Features Per HMM State
Japanese Word Acoustic Features
ハチ
TTS
Japanese Speech Acoustic Features
ゴ
ハチ
イチ

Figure 2: Overview of paralinguistic translation

In particular, we control duration and power of each word using a source-side duration and power super vector $\boldsymbol{x_i} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_{N_x}}]^\top$ and a target-side duration and power super vector $\boldsymbol{y_i} = \left[\boldsymbol{y_1}, \cdots, \boldsymbol{y_{N_y}}\right]^\top$. 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. $^\top$ indicates transposition. The sentence duration and power vector consists of the concatenation of the word duration and power vectors such that $\mathbf{Y} = [\boldsymbol{y_1}, \cdots, \boldsymbol{y_i}, \cdots, \boldsymbol{y_I}]$ where $I$ is the length of the sentence. In this work, to simplify our translation task, we assume that duration and power translation of each word pair is independent from that of other words, allowing us to find the optimal Y using the following equation:

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}} \prod_i P(\boldsymbol{y_i}|\boldsymbol{x_i}). \tag{8}$$

The word-to-word acoustic translation probability $P(\boldsymbol{y_i}|\boldsymbol{x_i})$ can be defined with any function, but in this work we choose to use linear regression, which indicates that $\boldsymbol{y_i}$ is distributed according to a normal distribution

$$P(\boldsymbol{y_i}|\boldsymbol{x_i}) = N(\boldsymbol{y_i}; \mathbf{W}_{e_i,j_i}\boldsymbol{x'_i}, S) \tag{9}$$

where $x'$ is $\left[1 x^\top\right]^\top$ 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 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 squared error (RMSE) with a regularization term

$$\hat{\mathbf{W}}_{e,j} = \arg\min_{\mathbf{W}_{e_i,j_i}} \sum_{n=1}^{N} ||\boldsymbol{y^*}_n - \boldsymbol{y}_n||^2 + \alpha||\mathbf{W}_{e_i,j_i}||^2, \tag{10}$$

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

### 3.4. Speech Synthesis

In the TTS part of the system we use an HMM-based speech synthesis system [5], and reflect the duration and power information of the target word paralinguistic information vector onto the output speech. The output speech parameter vector sequence $\mathbf{C} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_T]^\top$ 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:

$$\hat{\mathbf{C}} = \arg\max_{\mathbf{C}} P(\mathbf{O}|\hat{\mathbf{J}}, \hat{\mathbf{Y}}) \tag{11}$$

$$\text{subject to } \mathbf{O} = \mathbf{MC}, \tag{12}$$

where $\mathbf{O}$ is a joint static and dynamic feature vector sequence of the target 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.

## 4. Evaluation

### 4.1. Experimental Setting

We examine the effectiveness of the proposed method through English-Japanese speech-to-speech translation experiments. 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 accidentally makes a mistake when listening to the ticket number, the speaker re-speaks, emphasizing the place where the listener has made the mistake. 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. This information will not be present with linguistic information only.

[1] We chose $\alpha$ to be 10 based on preliminary tests but the value had little effect on subjective results.

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 lexical content to be spoken was 500 sentences from the AURORA2 data set, chosen to be word balanced by greedy search [6]. The training set is 445 utterances and the test set is 55 utterances, graded by 3 evaluators. We plan to make this data freely available by the publication of this paper.

Before the experiments, we analyzed the recorded speech's emphasis. We found several inclinations of emphasized segments such as shifts in duration and power. For example there are often long silences before or after emphasized words, and the emphasized word itself becomes longer and louder.

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 [7]. For the translation model we use 445 utterances of speech from our recorded corpus for training and hold out the remainder for testing. As the recognition and translation tasks are simple are simple , the ASR and MT models achieved 100% accuracy on every sentence in the test set. 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 baseline and two proposed models shown below:

Baseline: traditional lexical translation model only

Duration: Paralinguistic translation of duration only

Duration + Power: Paralinguistic translation of duration and power

The word translation result is the same between both models, but the proposed model has more information than the baseline model with regards to duration and power. In addition, we use naturally spoken speech as an oracle output. We evaluate both varieties of output speech with respect to how well they represent emphasis.

### 4.2. Experimental Results

We first perform an objective assessment of the translation accuracy of duration and power, the results of which are found in Figure 3 and Figure 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 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.

| | |
|---|---|
| Training sentences | 8440 |
| Word error rate | 0 |
| HMM states | 16 |

Table 1: Setting of ASR

| | |
|---|---|
| Training utterances | 445 |
| Test utterances | 55 |
| Regularization term | 10 |

Table 2: Setting of paralinguistic translation

| | |
|---|---|
| Training utterances | 445 |
| HMM states | 16 |

Table 3: Setting of TTS

As a subjective evaluation we asked native speakers of Japanese to evaluate how well emphasis was translated into the target language. 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 5. We can see that both of the proposed systems show a clear improvement in the emphasis recognition rate over the baseline. Subjectively the evaluators found that there is a clear difference in the duration and power of the words. In the proposed model where only duration was translated, many testers said emphasis was possible to recognize, but sometime it was not so clear and they were confused. When we also translate power, emphasis became more clear and some examples of emphasis that only depended on power were also able to be recognized. When we examined the remaining errors, we noticed that even when mistakes were made, mistakenly recognized positions tended to be directly before or after the correct word, instead of being in an entirely different part of the utterance.

The second experiment asked the evaluators to subjectively judge the strength of emphasis, graded 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 6. This figure shows that there

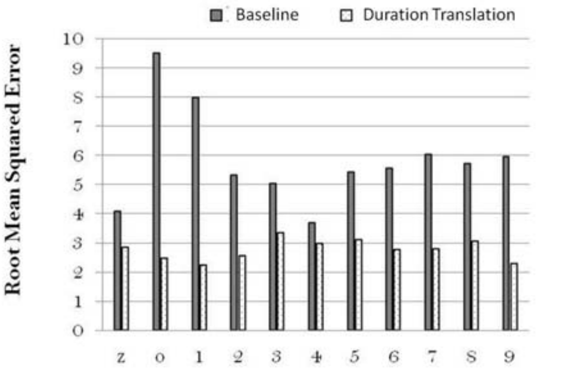

Figure 3: Root mean squared error rate (RMSE) between the reference target duration and the system output for each digit

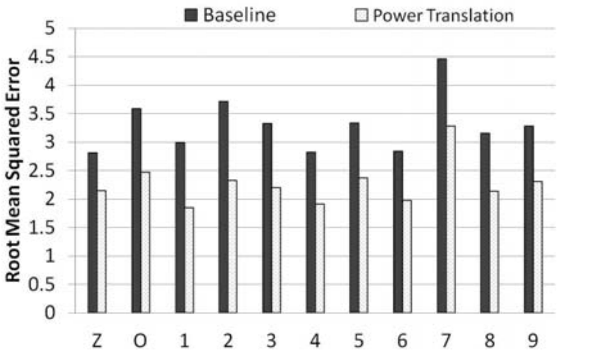

Figure 4: Root mean squared error rate (RMSE) between the reference target power and the system output for each digit

is a significant improvement in the subjective perception of strength of emphasis as well. Particularly, when we analyzed the result we found two interesting trends between duration translation and duration and power translation. Particularly, the former method was often labeled with a score of 2 indicating that the duration is not sufficient to represent emphasis clearly. However, duration+power almost always scored 3 and can be recognized as the position of emphasis. This means that in English-Japanese speech translation, speech's power is an important factor to convey emphasis.

## 5. Related Works

There have been several studies demonstrating improved speech translation performance by utilizing paralinguistic information of source side speech. For example, [8] focuses on using the input speech's acoustic information to improve translation accuracy. They try to explore a tight coupling of ASR and MT for speech translation, sharing information on the phone level to boost translation accuracy as measured by BLEU score. Other related works focus on using speech intonation to reduce translation ambiguity on the target side

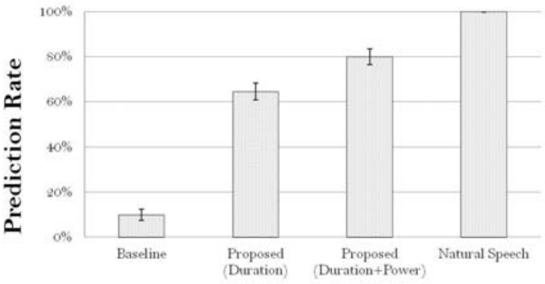

Figure 5: Prediction rate

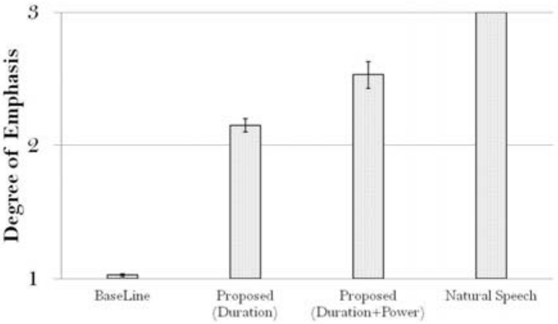

Figure 6: Degree of emphasis

[9, 10].

While the above methods consider paralinguistic information to boost translation accuracy, as we mentioned before, there is more to speech translation than just the accuracy of the target sentence. It is also necessary to consider other features such as the speaker's facial and prosodic expressions to fully convey all of the information included in natural speech. 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 mouth shape and uses this information to translate speaker emotion from source to target [1, 11]. On the other hand, [2] focus on the input speech's prosody, extracting F0 from the source speech at the sentence level and clustering accent groups. These are then translated into target side accent groups. V. Kumar et al consider the prosody in encoded as factors in the Moses translation engine to convey prosody from source to target [12].

In our work, we also focus on source speech paralinguistic features, but unlike previous work we extract them and translate to target paralinguistic features directly and in continuous space. In this framework, we need two translation models. One for word-to-word lexical translation, and another for paralinguistic translation. We train a paralinguistic translation model with linear regression for each word pair. This allows for relatively simple, language-independent implementation and is more appropriate for continuous features such as duration and power.

## 6. Conclusion

In this paper we proposed a method to translate duration and power information for speech-to-speech translation. Experimental results showed that duration and power information in input speech appears in output speech, and that this information can be used by target language speakers to detect emphasis.

In future work we plan to expand beyond the easy lexical translation task in the current paper to a more general translation task. Our next step is to expand our method to work with phrase-based machine translation. Phrase-based SMT handles non-monotonicity, insertions, and deletions naturally, and we are currently in the process devising methods to deal with the expand vocabulary in paralinguistic translation. In addition, traditional speech-to-speech translation, the ASR and TTS systems generally use phoneme-based HMM acoustic models. And it will be necessary to change our word-based ASR and TTS to phoneme-based systems to improve their performance on open-domain tasks. Finally, while we limited our study to duration and power, we plan to expand to other acoustic features such as F0, which play an important part in other language pairs, and also paralinguistic features other than emphasis.

## 7. Acknowledgment

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

## 8. References

[1] S. Ogata, T. Misawa, S. Nakamura, and S. Morishima, "Multi-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, 2001.

[2] P. D. Agero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *In Proceedings of ICASSP*. ICASSP, 2006.

[3] H. Ney, "Speech translation: coupling of recognition and translation," in *Proceedings of Acoustics, Speech, and Signal Processing.* IEEE Int. Conf, 1999.

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of ACL*, 2007.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis." Speech Communication, 2009.

[6] J. Zhang and 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*. ICPhS, 2003.

[7] H. G. Hirsh and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, 2000.

[8] J. Jiang, Z. Ahmed, J. Carson-Berndsen, P. Cahill, and A. Way, "Phonetic representation- based speech translation," in *Proceedings of Machine Translation Summit 13*, 2011.

[9] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX," in *In Proceedings of the 5th International Conference on Spoken Language Processing*. ICSLP, 1998.

[10] 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, 2001.

[11] S. Morishima and S. Nakamura, "Multimodal translation system using texture mapped Lip-Sync images for video mail and automatic dubbing applications." EURASIP, 2004.

[12] V. Kumar, S. Bangalore, and S. Narayanan, "Enriching machine-mediated speech-to-speech translation using contextual information," 2011.