A study on sequential comprehensibility annotations of non-native utterances based on listeners' shadowing, reading, and script-shadowing

(聴取者のシャドーイングと読み上げ音声を使った外国語音声の可解性を表す時系列アノテーションに関する研究)

林 振超

LIN Zhenchao

ID Number: 37-186986

Supervisor: Prof. Nobuaki Minematsu

Department of Electrical Engineering and Information Systems,

Graduate School of Engineering,

The University of Tokyo

Master Thesis

August 2020

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

LIN Zhenchao
August 2020

Acknowledgements

I would like to thank Prof. Nobuaki Minematsu, who gives me advice and support during two years on both academic and professional path. I would also thank to Lecture Daisuke Saito for maintaining infrastructure in the laboratory. All members in the laboratory, no matter who has graduated or not, actively help me collect experiment data and solve problems I encountered in finishing my master thesis. Additionally, I want to thank to Prof. Noriko Nakanishi of Kobe Gakuin University for giving me pratical advice to improve experiment result.

Besides, I would like to thank all my friends in Japan especially in UTokyo and Tokyo International Exchange Center (TIEC).

Abstract

Lack of annotation is a well-known problem for CALL (Computer-Aided Language Learning) studies [8] and it is not rare that researchers cannot find annotations or labels in the databases for specific purposes. However, DNN-based speech processing requires a large number of annotations on speech corpora. This study proposes a method to calculate sequential annotations of comprehensibility of learners' utterances based on listeners' shadowing, reading and script-shadowing. Experiments showed promising results. Further, acoustic features are proved to be sufficient enough for annotation to substitute phonological features derived by DNN-based ASR models. Finally, proposed sequential annotations are visualized for non-native utterances.

Table of contents

List of figures

List of tables

Chapter 1

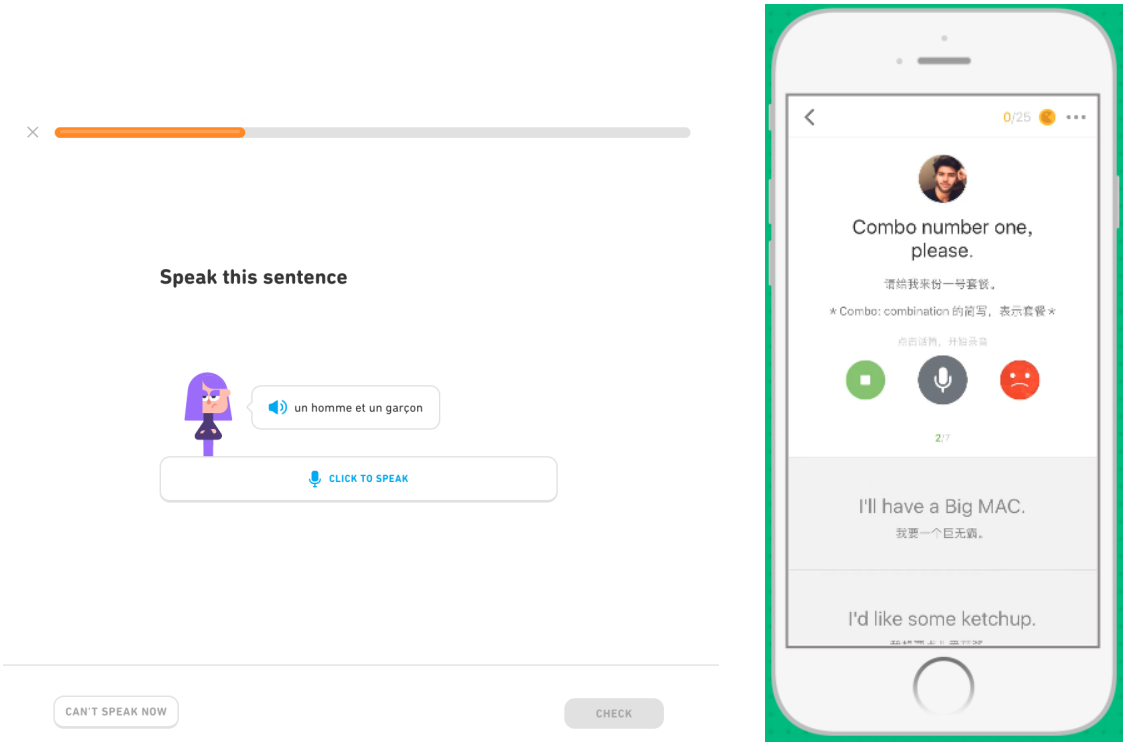
Introduction

This chapter begins with the background of the speech assessment. Based on existing problems in current speech assessment field, the objective of this study is explicitly defined: sequential comprehensibility annotations. Finally, the overall structure of this thesis is explained.

1.1 Background

Currently, labor shortage has been a huge problem for many countries and to mitigate this situation, some countries rely on introducing foreign labor. For example, Japan has actively introduced foreign talents into the workspace. However, unlike English, Japanese is not a global language and the language barrier may prevent the foreigners from meeting the needs of the local labor market. Therefore, how to objectively and specifically measure learners' language proficiency becomes a problem to solve.

When a language teacher is assessing a learner's utterance, traditionally, native-like accented utterance is always preferred. In the learner's utterance, it often contains some accents which may be transferred from learners' L1 [7]. Although this kind of utterances are often evaluated to be less proficient, the utterance may still be correctly and easily understood by native speakers or non-native speakers. From the perspective of information exchange, language is used for communicating with each other and understandable utterances are acceptable and enough. Moreover, nowadays, language teachers claim that the goal of verbal skill training should not be conducted for native-likeness but for intelligibility and comprehensibility [26].



(a) Duolingo[1] (b) Liulishuo[2]

Fig. 1.1 Interface of Duolingo and Liulishuo

With the development of machine learning and deep learning, automatic language assessment starts to appear frequently. Popular language learning applications like Duolingo [1] and Liulishuo [2] (which means fluently spearking in Chinese), whose interfaces are displayed in 1.1, both focus on learners' pronunciation and WER(word error rate) in recognized utterances. For beginners, with the feedback from the machine, they can improve their pronunciations steadily. However, for users who are more interested in making themselves easily understood, they need to be evaluated in terms of comprehensibility.

In order to automatically evaluate the utterance from the perspective of comprehensibility, the dataset including annotations, as known as corpus, is required for CALL (Computer-Aided Language Learning). However, the number of usable corpora is limited and they may not satisfy the following 2 requirements: 1) annotations shoule be directly related to comprehensibility. 2) the corpus is annotated at a fine-grained level. Normally, for most corpora, an overall score or descriptive categories are paired with the utterance or the speaker (learner). Learners' utterances may not always be bad or good in the whole sentence and learners may wonder how they perform on a single word or phone. Corpora annotated by sequential annotations based on comprehensibility are desired.

Based on the aforementioned two requirements, a reliable method for collecting comprehensibility-based sequential annotation is required for building the corpus. This study is aiming to provide an effective method to build this kind of corpus.

1.2 Objective

This research proposes a method to build corpora paired with sequential comprehensibility annotations given learners' reading utterances, which uses listeners' shadowing, reading and script-shadowing utterances. Through this method, the corpus and the process of building the corpus should be:

- The corpus should be annotated based on comprehensibility.
- The corpus should contain sequential annotations.
- The process of building the corpus should not rely on experts. Experts could annotate data in phonemic level but this kind of method cannot be adapted to scale up for a large corpus.
- The process of building the corpus should have a reasonable cost. High-cost methods such as using expensive devices to record brain activities are not pratical.

- The process of building the corpus can be compatible with any languages, including minority languages.

1.3 Structure of this thesis

Besides this chapter, the thesis is organized as following 5 chapters:

- Chapter 2 introduces basic knowledge about applied linguistics and previous works about language assessment.
- Chapter 3 introduces a method for annotating learners' utterances by comparing native reading and native shadowing.
- Chapter 4 annotates learners' utterances by comparing native script-shadowing and native shadowing.
- Chapter 5 diversifies shadowers' background range and explores the relationship of among different L1 listeners.
- Chapter 6 summarizes the whole thesis and introduces the future work to collect a large corpus which is called inter-learner shadowing.

Chapter 2

Basic Knowledge about Speech Assessment and Previous Works

This chapter begins with related technology and definitions of speech assessment. Dynamic Time Warping is a classic algorithm for comparing two sequences and deriving the best path between them. Speech assessment can be described through the point of view of intelligibility, comprehensibility and shadowability.

Previous works related to speech assessment are included in the second section. This section begins with pronunciation-based assessment, which prefers native-likeness, followed by intelligibility-based assessment. Finally, comprehensibility-based assessment is explained through some high-cost and low-cost examples.

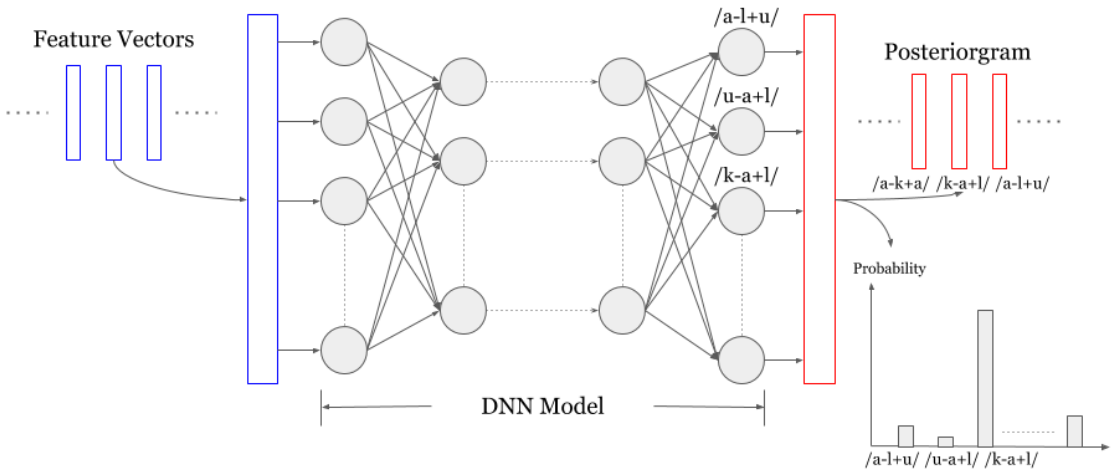


Fig. 2.1 How to calculate phonetic posteriorgram

2.1 Speech Assessment Basics

2.1.1 Phonetic Posteriorgram

Phonetic posteriorgram is a probability vector representing the posterior probabilities of a set of pre-defined phonetic classes for a speech frame [32]. The process of deriving phonetic posteriorgram is illustrated in 2.1. For one frame, its feature vector is used as the input of the DNN model and the output of the model is the probability vector of the posterior probabilities. The probability corresponding to the phonetic label of the frame is the maximum value of the vector of the posterior probabilities.

2.1.2 Dynamic Time Warping (DTW) and distance type

Dynamic Time Warping (DTW) is an algorithm to map one sequence to another sequence while maintaining minimal overall distance between two sequences. This algorithm is utilized in comparing stream data such as two speech segments. When comparing two speech segments, feature vectors such as MFCC (Mel-Frequency Cepstral Coefficients) and posteriorgram are usually used.

In Figure 2.2, for two sequences of feature vectors denoted by red and blue rectangles, their best matched pairs are denoted by black dots. Corresponding phones are noted near the

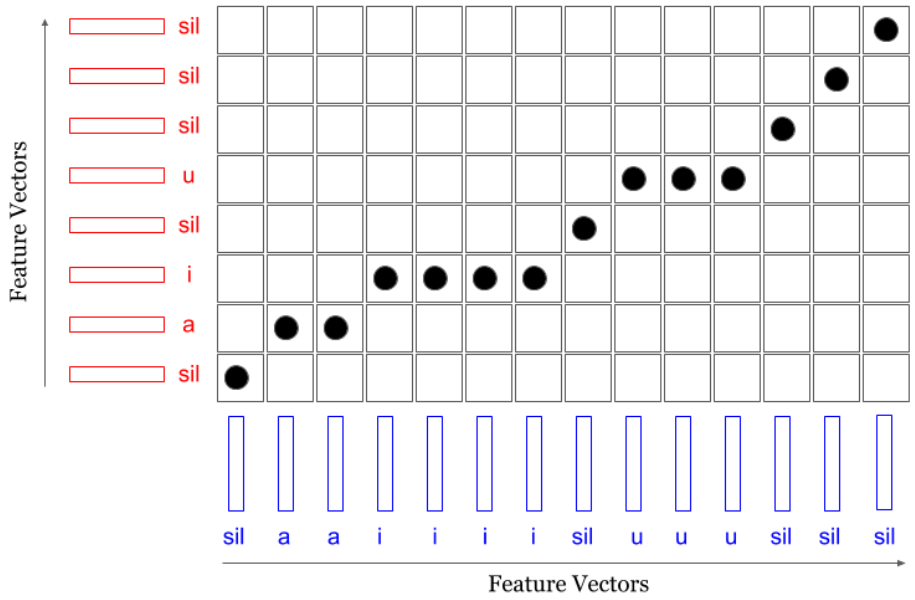


Fig. 2.2 DTW between two sequences of feature vectors

rectangle for each sequence of feature vectors, in which *sil* means silence. By applying DTW, best matched pairs are identified and form the best path between two sequences as connecting black dots in Figure 2.2.

DTW is implemented based on dynamic programming. Assume each movement can reach one step further in each direction, there may be at most three ways to reach one node in most cases. When using sequence index as i and j, to reach node (i, j), induction equation of the accumulated distance $f(i, j)$ can be represented as:

$$f(i,j) = \min \left\{ \begin{array}{c} f(i-1,j-1) + 2 * dist(i,j) \\ f(i-1,j) + dist(i,j) \\ f(i,j-1) + dist(i,j) \end{array} \right\} \tag{2.1}$$

Note that $dist(i, j)$ means distance between i-th vector in the first sequence and j-th vector in the second sequence. When calculating the distance between two vectors, some distance functions can be introduced. There are mainly 3 types of distance functions applied: Bhattacharyya Distance, Euclidean Distance and Cosine Distance.

- Bhattacharyya Distance

$$BD(x,y) = -\log(\sum_{i=1}^{N} \sqrt{x_i y_i}) \tag{2.2}$$

- Euclidean Distance

$$ED(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{2.3}$$

- Cosine Distance

$$CD(x,y) = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}} \tag{2.4}$$

Bhattacharyya Distance is used for measuring the distance between phonological feature vectors, which represent phonetic probability distributions, while Euclidean Distance and Cosline Distance are used for calculating the distance between acoutic feature vectors.

2.1.3 Intelligibility, Comprehensibility and Shadowability

In applied linguistics, learners' utterances are evaluated from the perspective of intelligibility and comprehensibility [20, 21, 26]. However, definitions of intelligibility and comprehensibility are difficult to define.

Intelligibility represents how accurately words or phones can be identified in the utterance. In [19], it is noted that intelligibility depends on the speaking skill of a speaker, the predictability of a content, and the language background of a listener. Intelligibility is measured through how many contents can be repeated and transcribed correctly in word-level.

Comprehensibility, indicates how smoothly and easily listeners can understand the utterance. Compared with intelligibility, evaluating comprehensibility of an utterance requires the listener to perform phonological analysis and syntactic analysis and this is normally collected by subjective questionnaire. In [16, 17], the definition of comprehensibility is considered to include the definition of intelligibility.

One more criterion of shadowability, which is referred to in this research afterwards, and comprehensibility are practically similar metrics, but theoretically and strictly speaking, both have a gap between them. In Chapter 3 and Chapter 4, the author is interested more in shadowability because the target is to predict smoothness of understanding via. observing listeners' shadowing behaviors.

2.2 Previous works

2.2.1 Pronunciation-based assessment

In previous studies aiming at automatic pronunciation assessment [29, 15, 31], comparison between learners' utterances and a native model of pronunciation was often made. When

evaluating learners' proficiency, learners listen to the audio prepared by native speakers and repeat audio contents in the accent they think approriate. Then recorded utterance and prompt text are used for measuring learners' language proficiency. In this process, GOP (Goodness of Pronunciation) is used as the feature to represent how correctly the learner has pronunciated.

GOP is an evaluation criterion for describing clarity of a speech [29]. Through calculating the posterior of each specific phoneme in the utterance and normalizing it by its duration, an overall score can be obtained technically. It is defined in [29] as:

$$\begin{aligned} GOP(p, O^{(p)}) &= \frac{1}{D_p} \log(P(p|O^{(p)})) \\ &= \frac{1}{D_p} \log(\frac{P(O^{(p)}|p)P(p)}{\sum_{q \in Q} P(O^{(p)}|q)P(q)}) \\ &\approx \frac{1}{D_p} \log(\frac{P(O^{(p)}|p)}{\max_{q \in Q} P(O^{(p)}|q)}) \end{aligned} \tag{2.5}$$

$O^{(p)}$ is a given utterance segment uttered as phone p. Q is the set of all phone models. D_p is the duration of corresponding utterance p.

The procedure of DNN-GOP is illustrated in Figure 2.3. For an utterance, feature vectors are extracted. Then based on provided text and calculated feature vectors, forced-alignment is applied to determine the phone of each frame. Again, using calculated feature vectors, as illustrated in 2.1, posteriorgram is derived by pre-trained DNN-model. Finally, for each frame, merge all posterior probabilities corresponding to the result of phone-level alignment and take average over all non-silence frames as GOP scores.

In [31], GOP scores calculated based on HMM(Hidden Markov Model) and DNN are compared by calculating the correlation to manual scores. In Figure 2.4, DNN-GOP can more precisely represent manual scores.

DNN-based GOP calculation can be simplified as following equation:

$$GOP(p, O^{(p)}) = \frac{1}{D_p} \log(P(p|O^{(p)})) \tag{2.6}$$

With the GOP score for each frame calculated, summing up posteriors probability and normalizing it by utterance or word duration. Then derived GOP scores can be combined into different patterns to obtain utterance-level or speaker-level GOP scores.

2.2.2 Intelligibility-based assessment

In Section 2.1.3, the definition of intelligibility is described as how accurately words and phonemes can be identified in the utterance. Intelligibility was measured objectively in [6, 19],

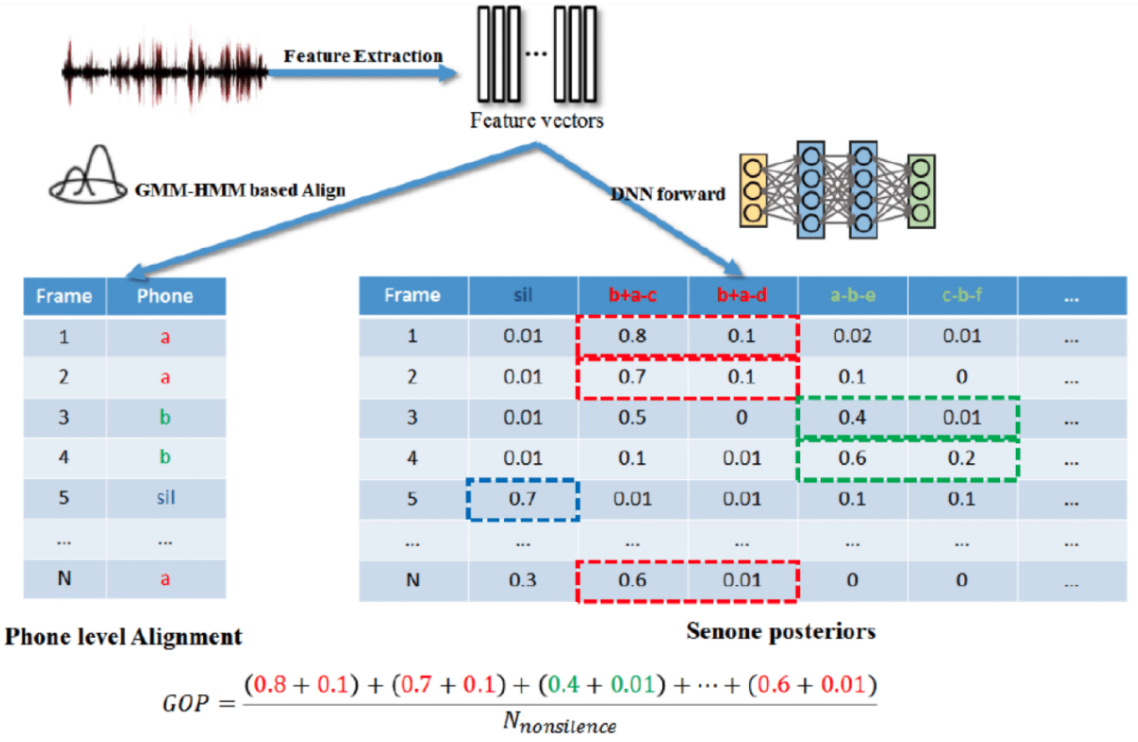


Fig. 2.3 DNN-GOP Overview [30]

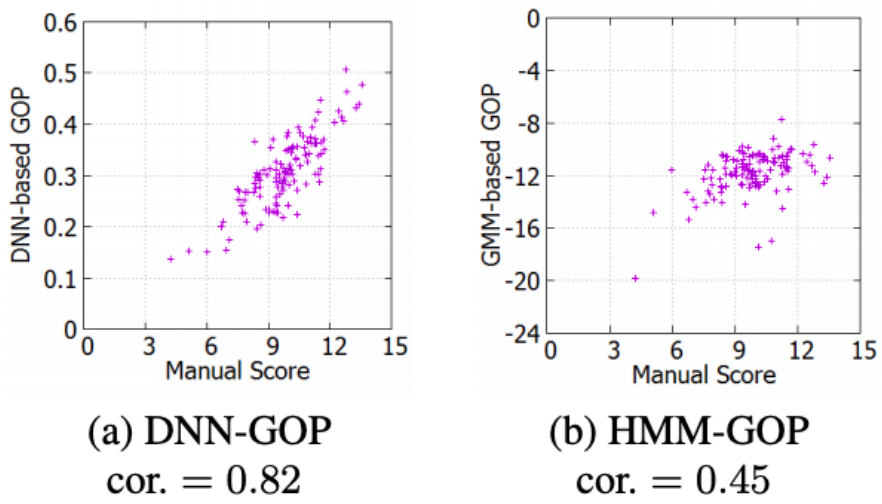


Fig. 2.4 Relationship between manual scores and DNN/HMM-based GOP

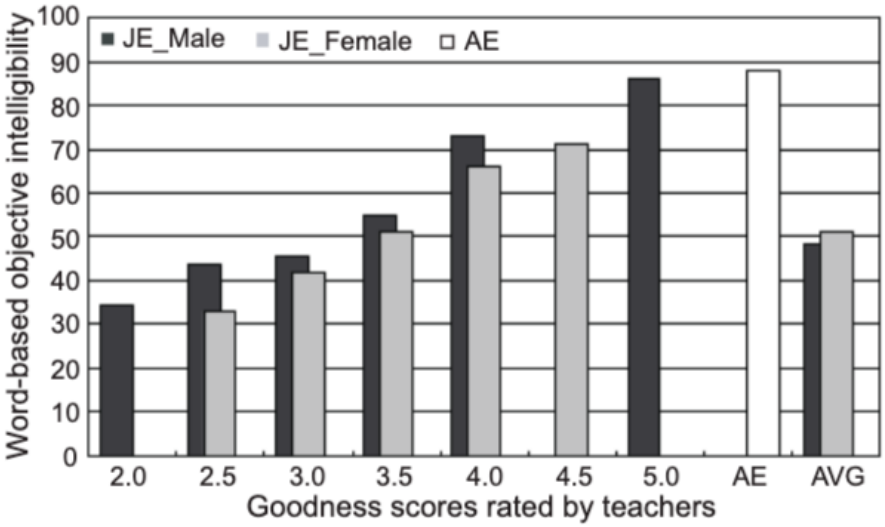


Fig. 2.5 Word-based intelligibility for different learner groups[19]

where English spoken by immigrants to the USA [6] and English spoken by Japanese college students [19] were presented to native American listeners on the telephone.

To validate whether word-based objective intelligibility scores are highly correlated with human scores, in [19], a listening test was conducted using the speech stimuli carefully selected from ERJ (English Read by Japanese) database [18]. During the experiment, American participants were asked to repeat Japanese accented Engligh immediately after hearing the stimulis. Then repeating utterances were transcribed manually and word identification rate was utilized as word-based intelligibility score. Manual transcriptions were used as intelligibility annotations representing which word segment in the non-native utterances is perceived by native listeners with how much accuracy [23, 24]. However, the authors consider that this approach lacks scalability because collection of transcriptions requires cost and time. Another disadvantage of this repeat-after-listen approach is offline transcription, which gives the chance to guess the word in transcribing. Therefore, this approach can be said not to be practical enough to increase the amount of annotations.

In Figure 2.5, Japanese accented English utterances were categorized by goodness scores rated by teachers and American English utterances were appended for comparison. For Japanese accented English utterances, clear linear relation was observed, while the group of 5.0 goodness score is comparable with American English speakers.

Fig. 2.6 Equipment for measuring brain activity

2.2.3 Comprehensibility-based assessment

In Seciont 2.1.3, the definition of comprehensibility is defined as how smoothly and easily listeners can understand the utterance. When listeners are listening to learners' utterances, if utterances are more comprehensible, the cognitive load in listeners' brain is low. Otherwise, listeners cannot easily understand contents with little cognitive load involved. Because comprehensibility should be measured by consecutively observing listeners' behavior or reaction, some online methods are required.

Cognitive load can be quantitatively measured by supportive equipment. In [9, 27, 11], EEG (electroencephalogram) recordings were made from listeners and listening efforts were discussed quantitatively. In [10], eye-trackers were used to measure the size of pupils to predict the magnitude of cognitive load when listening. However, due to the high cost for these high-tech equipments, deriving comprehensibility through these methods cannot be scalable to a large user groups.

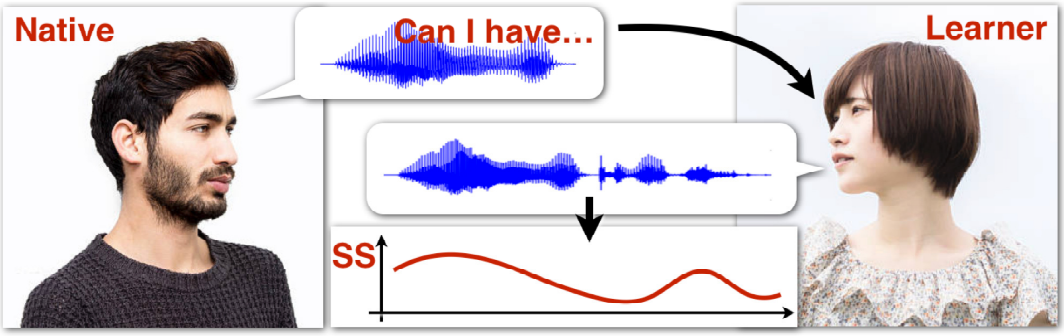
Another way to quantify comprehensibility is to use a simple equipment, a microphone. In [17, 16], unlike traditional pronunciation-based assessment which directly evaluated on learners' utterances, native speakers were asked to listen to learners' utterances and immediately repeat them in native accent, which is called shadowing. When learners' utterances are more comprehensible, native listeners can shadow them easily and smoothly. Based on the result of DNN-based ASR front-end, smoothness in native listeners' shadowing can represent how comprehensible corresponding learners' utterances are.

Chapter 3

L2 Speech Assessment based on Comparison between Native Reading and Native Shadowing

In this chapter, a novel method for evaluating learners' utterances through comparion between native shadowing and corresponding native reading is introduced. Compared with the method in previous works which required a DNN-based ASR front-end [16], this method can mitigate the effect that a DNN-based ASR front-end for a specific language is sometimes or often hard to train or find.

To prove the validity of this method, utterance-based evaluation is done and this method is proved to be more effective than traditional pronunciation-based evaluation and comprehensibility-based evaluation introduced in [16]. Furthermore, word-based evaluation is analyzed and annotations derived by comparing method, which involves native reading and native shadowing, are highly correlated with human-rated subjective shadowability scores. Finally, linear prediction is imported to explore the potential of each feature, which give researchers a hint about what next improvement should be focused on.



SS means smoothness of shadowing.

Fig. 3.1 Conventional form of shadowing

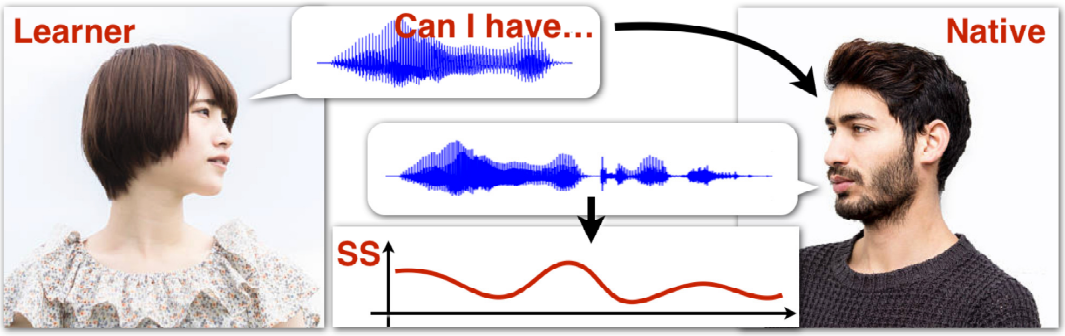


Fig. 3.2 Reverse form of shadowing

3.1 Conventional form of shadowing and reverse form of shadowing

Shadowing is a special type of listen-and-repeat practice, where a listener has to repeat a given utterance as simultaneously as possible, shown in Figure 3.1. Shadowing was originally introduced as a practising strategy for simultaneous interpreters since it includes not only speaking and listening but also understanding a given speech. Recently, researches and teachers have shown that shadowing is also effective for second language learning [12–14]. Conversation is generally a speech activity where three processes of speaking, listening, and understanding are overlapped. Practically speaking, conversation is a multi-task speech activity, and shadowing is used in classrooms as it can put learners effectively in this multi-task situation. In [31], the author proposed a DNN-based technique to predict smoothness of shadowing, SS for short in Figure 3.1.

In [16, 17], a novel method of predicting comprehensibility of an utterance was proposed, that is native listeners' reverse shadowing and it does not require any special device like EEG or eye-tracker. In the conventional form of shadowing, native utterances are presented to

learners, who are shadowers. In reverse form of shadowing, learners' utterances are presented to native shadowers, shown in Figure 3.2. Here, shadowers are asked not to imitate accented pronunciations but to reproduce what was said in their own native pronunciation. Since smooth shadowing always requires smooth understanding [28], smoothness or brokenness of natives' shadowing, which was acoustically measured, was examined and shown to be effective to predict comprehensibility subjectively rated by shadowers.

3.2 Problems in previous works and proposed method

The author considers that it is very natural to view native listeners' shadowing as a kind of annotation assigned to a given non-native utterance, which can characterize comprehensibility of that utterance quantitatively even in the form of sequential data, not a single score or some descriptions. As was shown in [16, 17], non-expert and ordinary native listeners can be adopted as shadowers, but the authors can point out two drawbacks when we simply use DNN-based GOP scores to represent comprehensibility dynamics.

Generally speaking, annotations or labels are given manually to speech samples. When some techniques are used for annotation, they should be stable and reliable techniques. DNN-based ASR can work more stably and reliably than HMM-based ASR, but DNN-based ASR still needs adaptation techniques with respect to speaker identity, speaking style, recording environment, etc. This fact implies that DNN-based GOP scores can be unstable and unreliable in some specific situations.

The other drawback is more crucial. In [16, 17], the target language of learning was Japanese, which is not an international language, and in this case, shadowers should be native speakers of Japanese. If we adopt English as target language, as it is used internationally, some learners want to know how comprehensible their utterances are to non-native speakers of English. In this case, we can ask non-native listeners to shadow. Even when their shadowing is very smooth, however, their utterances are often accented, which easily reduces GOP scores if the DNN-based acoustic models are trained with a native speech corpus. If non-native acoustic models are available separately for each of native languages, the above problem may be able to be solved, but this is very impractical. Further, even native shadowers may be rejected as shadower if their pronunciations are regionally accented. To use listeners' shadowing as spoken annotation effectively, a different method of calculating smoothness of shadowing is needed, which should be stabler.

In this section, the above problems are solved just by introducing another simple speech task to shadowers. When listeners shadow a given utterance, they do not refer visually to the sentence that the learner read aloud. Then, after they shadow, we present the sentence

Fig. 3.3 Karaoke-style Recording Interface[16]

visually and ask them to read it aloud. The author considers that reading is the most prepared speech and shadowing is probably the least prepared speech, or hastened speech. If smooth or quick understanding is possible enough while shadowing, the shadowing speech will become acoustically closer to the reading speech. DTW between the two types of speech gives us the optimal path, on which a sequence of local distances can be viewed as a sequence of comprehensibility. Further in this method, DTW is conducted within the same speaker, with the same microphone. The author can claim that this is the best recording condition for utterance comparison using DTW.

3.3 Data collection

3.3.1 Collection of Vietnamese Japanese readings

From 60 Vietnamese learners of Japanese, we collected two kinds of reading utterances, readings of sentences in Japanese textbooks and those of the learners' own essays or passages. The period of learning Japanese was one year for 27 learners and two to three years for the other 33 learners. For textbook reading, recording was done by a unit of phrase and every recording was so long as a phrase. For essay or passage reading, recording unit was not fixed and many learners recorded sentence by sentence, but some others recorded their whole essay or passage in a single recording session. A part of the recordings were used for collecting native listeners' shadowing utterances, which is explained below.

3.3.2 Collection of natives' shadowings and readings

Two native speakers participated in our experiments and each of them shadowed a different set of 800 utterances of Vietnamese-Japanese (VJ). After the entire shadowing experiment, we recorded the shadowers' reading utterances. For each shadower, 400 sentences that the

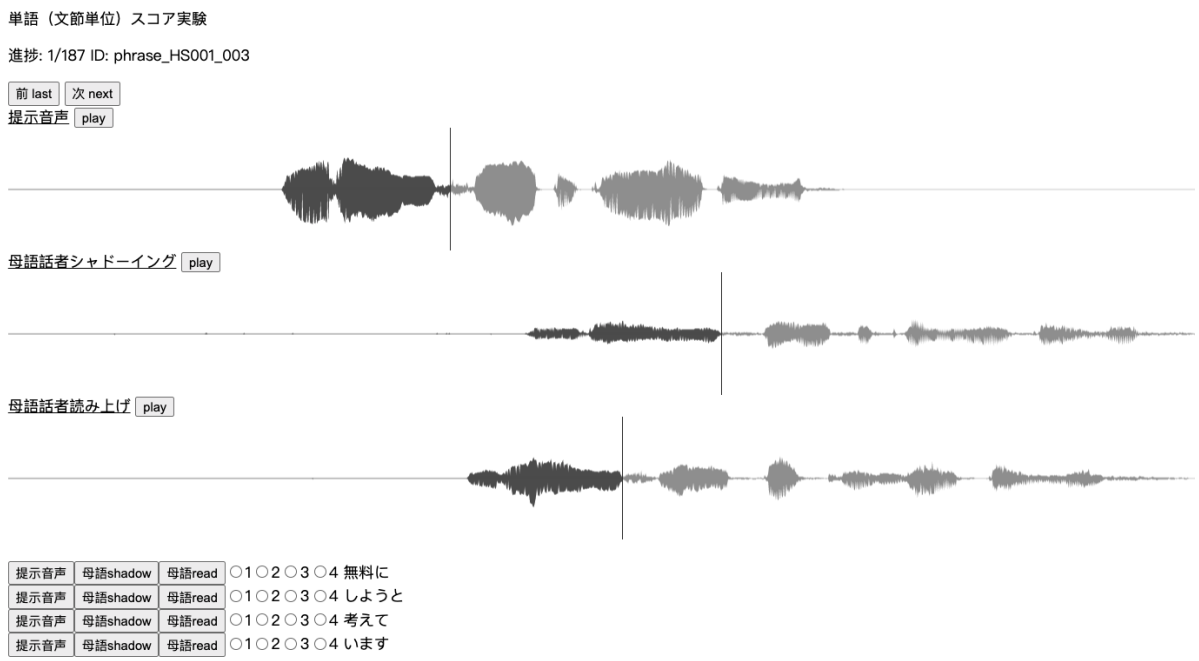


Fig. 3.4 Word-based scoring interface

Vietnamese learners used for recording, i.e. a half amount of data used for shadowing recording, were visually presented to the shadower, who was asked to read aloud the sentences.

3.3.3 Collection of utterance-based and word-based shadowability scores from native shadowers

After reading aloud each sentence, both of the reading utterance and its corresponding shadowing utterance were presented to the shadower through headphones so that s/he could rate degree of shadowability of the whole utterance in a seven-degree scale, i.e. how correctly or incorrectly articulation was performed in shadowing.

For word-based assessment, Figure 3.4 shows the interface used in word-based scoring. Through clicking buttons before the word, corresponding audio segment, whose index is obtained by forced alignment based on DNN-based ASR front-end, is played and shadowers can repeatedly play the whole utterance or the word segment. A four-level scale was used and the four levels indicate (1) totally broken, (2) broken, (3) partially broken, and (4) smoothly shadowed. Score assignment was done only once, but the shadowers were allowed to listen to the recordings repeatedly. These scores are used as word-based subjective scores of shadowability in the following sections. In the following section, we compare correlations of the GOP scores

of native listeners' shadowings to the subjective shadowability scores and those of the DTW scores between shadowing and reading to the subjective shadowability scores.

3.4 Experiment

3.4.1 Utterance-based evaluation

Detailed procedures of DTW

Following [31], posteriorgram was adopted as speech representation and any utterance was represented as a sequence of phoneme-posterior vectors. Here, the CSJ-based KALDI recipe [25] was used to train Japanese DNN acoustic models. For DTW between shadowings and readings, all the utterances were converted to their posteriograms with a DNN-based ASR front end [16]. The most problematic thing in the DTW is that readings and shadowings often have pauses at different positions within the utterances. In readings, pauses are intentionally inserted at punctuations or phrase boundaries in sentences, so that the utterances will become more natural and clear. Nevertheless, in shadowings, pauses are often found not at syntactic boundaries but at positions where listeners' understanding process did not work smoothly and they had to wait to continue shadowing. This irregular pausing is mainly due to low shadowability of original learners' utterances. To calculate shadowability scores objectively from the DTW path, pauses in mismatched positions between readings and shadowings have to be handled in a proper way.

Further, while some words are missing in shadowing utterances, words that are not in readings can sometimes be found in shadowings, e.g. repetitions and unintentionally produced words of surprise such as what or hmm. To handle these phenomena adequately, the following procedure was examined.

Comparison between a shadowing and its reading was done via. DTW and accumulated distances are calculated only on speech segments. We prepared two types of distances, D_{RS} and D_{SR}, shown in Figure 3.5. In D_{RS}, reading was used as reference and speech segments were detected from the reading, that are drawn in blue in Figure 3.5. The DTW paths for those speech segments were used to calculate the accumulated distance. In D_{SR}, shadowing was used as reference and speech segments were detected from the shadowing, that are drawn in red in the figure. The DTW path for those speech segments were used. In either case, the accumulated distance was normalized by the number of speech frames in reading or shadowing in utterance-based pattern.

Another question is how to detect pauses in a specific utterance. In this paper, two methods for pause detection are applied. One is based on the result of forced alignment and this

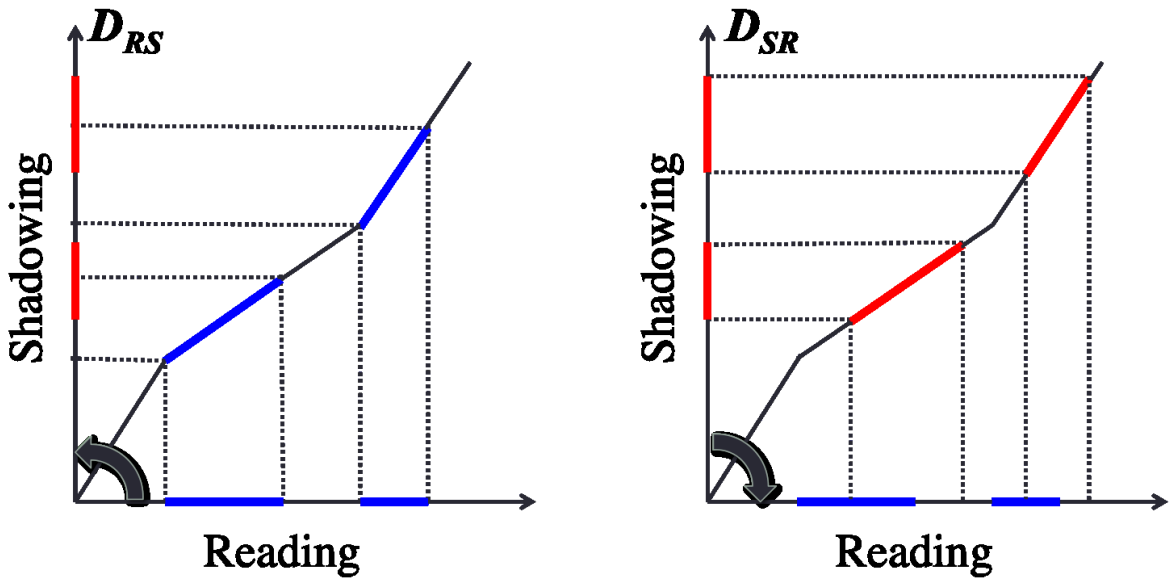


Fig. 3.5 Two kinds of DTW scores, D_{RS} and D_{SR}

method can be applied only when text of learners' speech is available. The other is based on posteriogram and this method can be applied even without text of learners' speech.

In Figure 3.6, an example of DTW of utterances of speaker HS001 is illustrated. Background color is painted to represent local distances, where darker red means larger distance while deeper blue means smaller. Speech segments found in the optimal path are painted by scattered white dots. The upper figure in Figure 3.6 is drawn based on D_{RS} and the lower is drawn with D_{SR}. In the upper, it is found that some speech frames in reading are missing in shadowing.

Results and discussion

For a pair of a shadowing and its reading, two DTW scores are linearly combined as $\alpha D_{RS} + (1-\alpha)D_{SR}$, where α varied from 0.0 to 1.0 with a step of 0.1. When α=1, the combined score is the same as D_{RS} and when it is 0, the score is the same as D_{SR}. Correlations of the combined scores to the shadowability scores are shown in Table 3.1 for each of the two native listeners, where the upper table shows the correlations with text and forced alignment and the lower shows those without text but with posteriogram. In each table, correlations of GOP scores calculated only on shadowings are also shown.

The tables show that D_{RS} tends to have higher correlations than D_{SR}, but in the case of shadower HS002 with text, D_{SR} shows a higher correlation than D_{RS}. From this table, we can say that 0.6 will be the best weight for α. It is clearly shown that the scores based on DTW

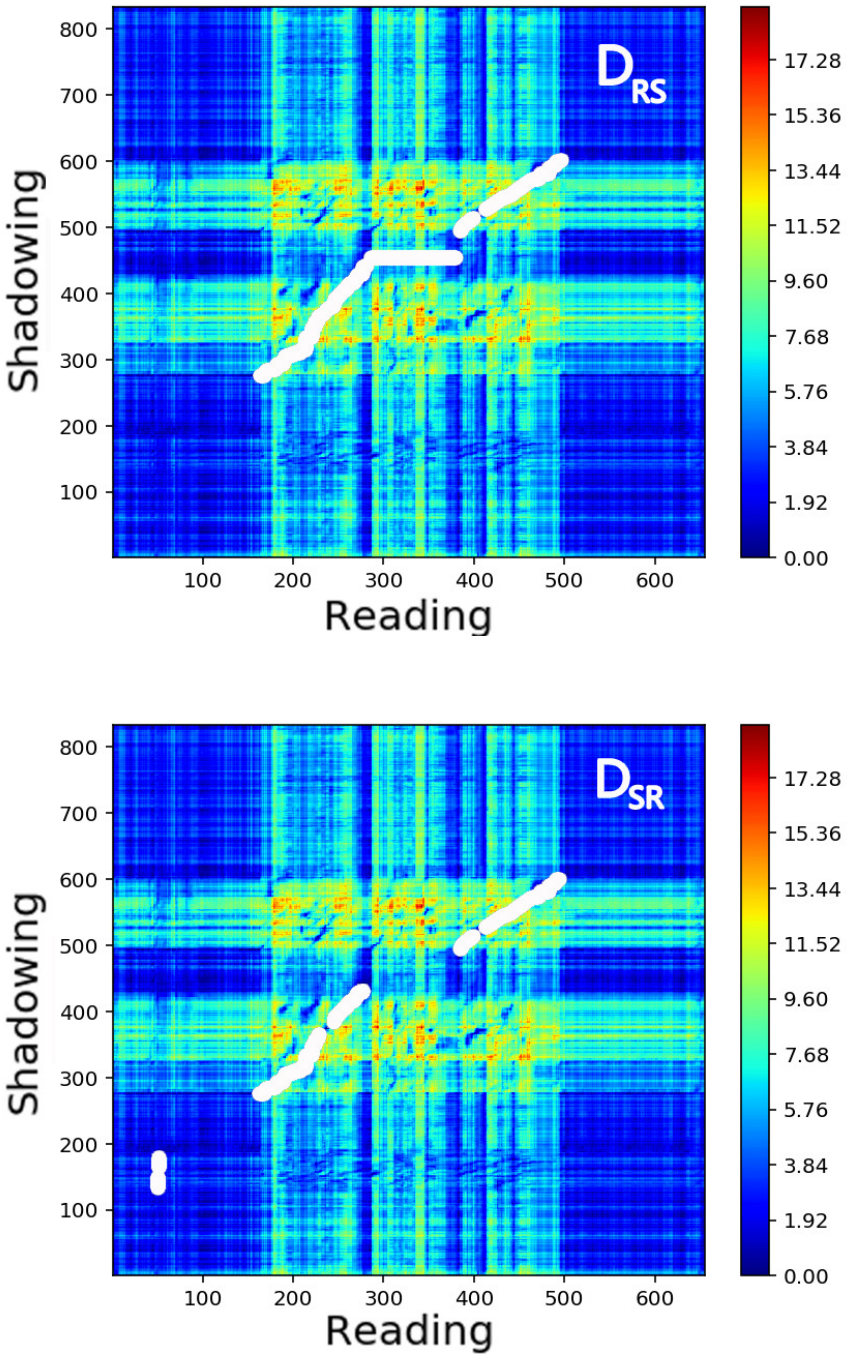


Fig. 3.6 Two kinds of DTW path, D_{RS} and D_{SR}

Table 3.1 Correlation of DTW and GOP to shadowability

1) with text and forced alignment

α	HS001	HS002
0.0(D_{SR})	-0.52	-0.61
0.1	-0.54	-0.61
0.2	-0.55	-0.62
0.3	-0.57	-0.61
0.4	-0.58	-0.61
0.5	-0.59	-0.60
0.6	-0.60	-0.59
0.7	-0.60	-0.58
0.8	-0.61	-0.57
0.9	-0.61	-0.56
1.0(D_{RS})	-0.61	-0.54
GOP	0.41	0.50

2) without text but with posteriogram

α	HS001	HS002
0.0(D_{SR})	-0.55	-0.60
0.1	-0.56	-0.62
0.2	-0.57	-0.63
0.3	-0.57	-0.63
0.4	-0.58	-0.64
0.5	-0.58	-0.64
0.6	-0.59	-0.64
0.7	-0.59	-0.64
0.8	-0.59	-0.63
0.9	-0.60	-0.63
1.0(D_{RS})	-0.60	-0.62
GOP	0.45	0.55

Table 3.2 Correlations between shadowability scores and automatically calculated scores

shadower	bGOP	pGOP	bDTW	pDTW	pDTW*
HS001	0.39	0.45	0.60	0.59	0.57
HS002	0.43	0.55	0.62	0.72	0.61
average	0.41	0.50	0.61	0.66	0.56

between shadowings and readings have higher correlations than the GOP score calculated only from shadowings but with DNN-based acoustic models. As discussed in Section 3.2, DTW-based scoring has much higher availability as it can be applied to non-native shadowers.

Scores of phoneme-based DTW and its variant

In the experiment mentioned above, the distance between native reading and native shadowing was accumulated frame-by-frame on the DTW path and every node on the path had the same weight. However, when shadowers prolonged or shortened their pronunciation for a single word or phone, the derived distances between native reading and native shadowing might vary. To eliminate this kind of effect, nodes on the DTW path can be first aggregated and averaged in phoneme-level. Then distances for each phoneme can be aggregated and averaged for deriving overall distance of the whole utterance. This kind of DTW score is called phoneme-base DTW score.

Table 3.2 shows correlations between subjective scores of shadowability and five kinds of automatically calculated scores of bGOP, pGOP, bDTW, pDTW, and pDTW*, where b and p mean baseline frame-based scores and phoneme-based scores. bDTW is the similiar method used in Table 3.1. pDTW, which is the abbreviation of phoneme-based DTW, represents the method to aggregate nodes by each phoneme and then aggregate the distance of each phoneme. pDTW* indicates the variant of pDTW, where functional words were ignored in calculating pDTW. It is clearly shown that DTW-based scores are more highly correlated to subjective scores than GOP-based scores, but that in DTW, pDTW* scores do not achieve higher correlations than pDTW. This indicates that subjective rating of shadowability is influenced not only by how shadowers shadowed content words in learners' utterances but also by how they shadowed functional words.

3.4.2 Word-based evaluation

How to verity the sequential annotation with a finer granularity

In Section 3.4.1, a sequence of local DTW distances were calculated along the DTW path obtained between the two utterances from the shadower. Although this sequential annotation is surely related to the learner's utterance presented to the shadower, we still don't know what kind scores should be assigned to individual words, syllables, phonemes, even frames in the learner's utterance. The sequential annotation obtained in Section 3.4.1 need to be further processed for annotation.

Frame-based assignment will be technically possible, but probably impossible by humans. To what kind of small units, can human raters assign scores reliably? In the experiments, because we compare machine scores and human scores for assessment, and because non-expert native shadowers will have difficulty in assessing even syllable-based units, we decided to conduct experiments using word[1] as basic unit for annotation.

Procedure of deriving word-based annotations

There are two types of shadowability annotations that can be derived word by word from the DTW alignment between native reading (NR) and native shadowing (NS). One is a sequence of the DTW local distances on the DTW path, which corresponds to brokenness of articulation in shadowing. The other is amount of time required to shadow each word in learner reading (LR), which is calculated by comparing the length of a word in LR and that of its corresponding word in NS. In good and synchronous shadowing, the two lengths are similar but if a word in LR is difficult to understand, then, the length of the corresponding word in NS becomes longer. To detect word boundaries, forced alignment was applied both to LR and NS.

Unlike [31], in this study, DTW is always conducted within the same speaker, i.e. between NS and NR. Taking this condition into account, acoustic representation of MFCC was also tested. The local distance between frames was calculated as the Bhattacharyya Distance with posteriorgram and as the Euclid Distance or the Cosine Distance with MFCC.

In Section 3.4.1, DTW was conducted between an overall reading and an overall shadowing from the same native speaker, where pauses were removed in advance to reduce alignment errors. In this experiment, however, since a subjective score was assigned to each word not to an entire utterance, DTW was also conducted separately for each word as in Figure 3.7, where three paths are drawn for three word segments. For detecting word boundaries, forced alignment

[1] Strictly speaking, the adopted unit was *bunsetsu*, a word concatenated with a post-positional word of 1-mora or 2-mora length.

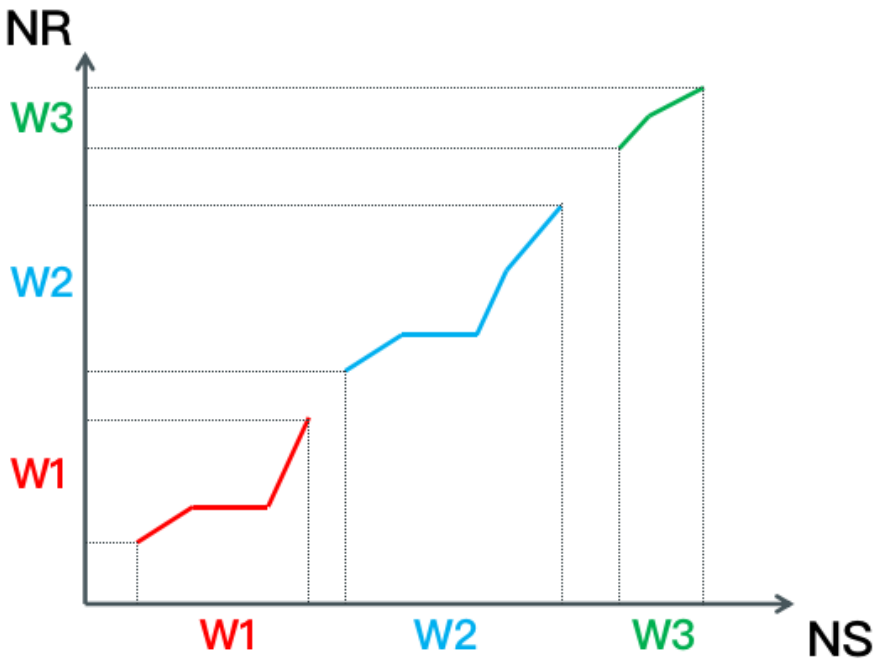


Fig. 3.7 Word-based DTW between native shadowing (NS) and native reading (NR)

was performed both on NR and NS. The average of the local distances within each word was defined as objective score for that word, which should be assigned to the corresponding word in LR.

Temporal lengthening or shortening in shadowing is represented by the ratio of the length of each word in LR to the length of the corresponding word in NS, shown in Figure 3.8. For easy comparison, the starting time of LR and that of NS are positioned at the same time index. The first word in NS is longer, while the other two words are similar to those in LR in length.

Results and discussion

Table 3.3 shows correlations calculated between the subjective scores and each kind of the objective scores separately for each of the two shadowers, HS001 and HS002. Correlations calculated by dealing with the two shadowers together are also shown. In the table, posteriorgram shows the highest correlations while MFCC unexpectedly shows low correlations. After the experiments, we found that recording of NSs and NRs were not done in a single session [5]. All the NSs were recorded on a day and all the NRs were recorded on another day. The recording equipment was shared in both recordings, but the recording room was different. This

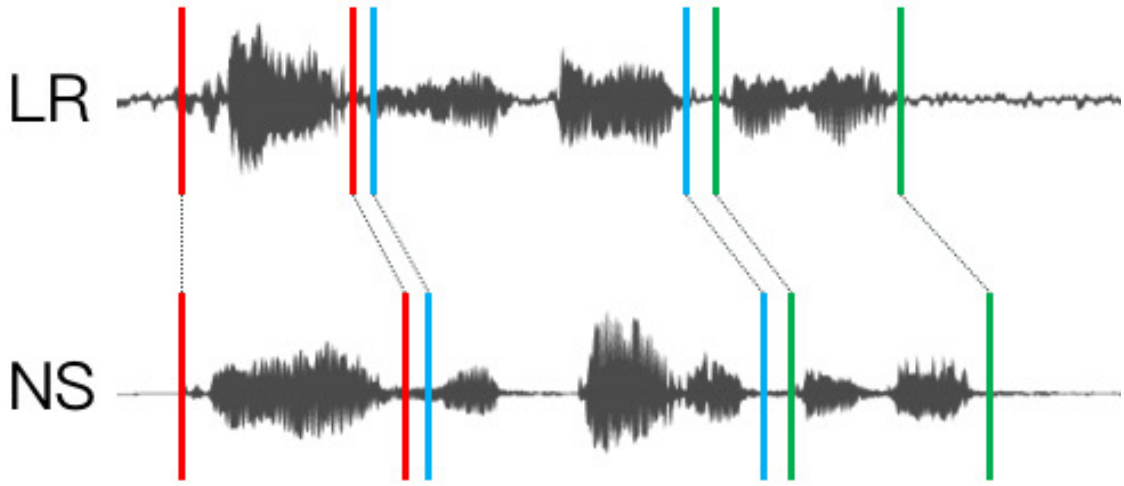


Fig. 3.8 Word-based temporal lengthening between LR and NS

conditional gap may have influenced the experimental results. As for consecutive recording of NS and NR, its technical and pedagogical validity will be discussed later, where effectiveness of MFCC is examined again.

Generally speaking, correlations between subjective scores and objective scores tend to be high when a score is assigned holistically to each learner by using his/her utterances together, i.e. learner-based annotation. As the unit of annotation becomes smaller such as one sentence, one phrase, one word, one syllable, and one phoneme, correlations tend to be smaller [22]. Even in these cases, correlations can become larger when averaged scores are used over multiple raters. This is because inevitable deviations in subjective assessment can be reduced. In the experiment here, we took what is supposed to be the minimum unit for non-expert native speakers, i.e. word. Further, when listener-based diversity is taken into account, averaging operations over listeners may not be adequate. This is why we did not calculate the averaged scores over the shadowers but calculated correlations by using their scores as independent data. In spite of these difficult conditions, posteriorgram shows very high correlations in Table 3.3.

It should be much noted that Posteriorgram's correlations in Table 3.3 are higher than those obtained in Section 3.4.1, where posteriorgram-based DTW was conducted between NS and NR for utterance-based annotation. The reason of higher correlations for shorter units is considered to be due to uniqueness of native shadowing. In shadowing an L2 utterance, it is likely that many words are shadowed smoothly while a few others are not. In this case, brokenness of articulation is suddenly raised. Distinction between these words is impossible for utterance-level annotation but easy for word-based annotation.

Table 3.3 Correlations of word-based subjective scores and objective scores for each of the two shadowers

Feature	HS001	HS002	both
MFCC_E	0.454	0.446	0.457
MFCC_C	0.169	0.128	0.144
Lengthening	0.430	0.329	0.375
Posteriorgram	**0.709**	**0.797**	**0.734**
pGOP (NS)	0.579	0.528	0.572
pGOP (LR)	0.108	0.205	0.123

In Table 3.3, the averaged score of phoneme-based GOPs (pGOPs) calculated for each NS word and that for each LR word are also used to calculate their correlations to word-based subjective scores. Their correlations are by far lower than posteriograms' correlations. In [16], the averaged score of pGOPs was calculated for each NS utterance and it showed a high correlation (0.73) to subjective scores. In [16], an objective score was obtained as the averaged score among 27 shadowers' shadowings and a subjective score was also obtained as the averaged score among 27 shadowers' judgments. In [5], the same data was re-analyzed without averaging operations, and the resulting correlation was found to be low (0.50). The averaging operation is powerful to increase correlations, but posteriorgram's correlations in Table 3.3 were obtained without averaging operations over raters.

3.4.3 Word-based evaluation predicted by linear regression

Features to generate the annotation

Similarly to Section 3.4.2, there are two types of annotations that can be derived. One is the DTW-based distances between native reading and native shadowing, and the other one is amount of time required to shadowe each word in learner reading.

In this experiment, since shadowing and reading are performed by the same speaker, DTW is done based on posteriorgram and MFCC. When calculating the local distance between two corresponding frames, Bhattacharyya Distance is applied to posteriorgram features while Euclid Distance and Cosine Distance are applied to MFCC features. The length of a word is obtained by forced alignment.

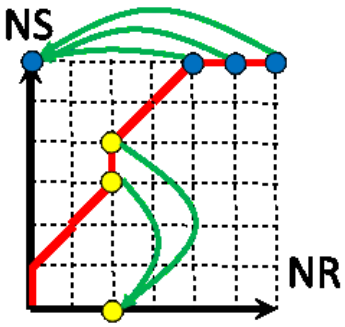


Fig. 3.9 How to map the DTW path to native shadowing (NS) and native reading (NR)

DTW between native shadowing and native reading

In 3.4.1, DTW was applied on overall reading and shadowing utterance and pause removal was done to reduce errors caused by silent frames. In this experiment, since a subjective score was given to each word segment but not to the entire utterance, DTW was done separately within each word. For this, text and forced alignment is required to obtain word boundaries. By averaging the local distances within a word frame by frame, its objective score is derived and assigned to that word in the learner's utterance.

After applying DTW between native shadowing and native reading, the derived path and its distances can be viewed as annotation for the learner's reading. Next step is to map these distances to the learner's utterance.

Before processing the learner's utterance, nodes on the DTW path are assigned separately to each frame in native shadowing and to that in native reading, illustrated in Figure 3.9. Here, we have a DTW path p between native shadowing and native reading, in which each node contains the frame index for native reading and that for native shadowing. The score of the i-th frame in native reading and that in native shadowing is computed as

$$\text{score}_{NR}[i] = \text{avg}(\text{Dist}(i, j)),\ j \in \{x | p[x][0] = i\} \tag{3.1}$$

$$\text{score}_{NS}[i] = \text{avg}(\text{Dist}(i, j)),\ j \in \{x | p[x][1] = i\}. \tag{3.2}$$

This process is illustrated in Fig. 3.9. Blue dots represent how path nodes sharing the same vertical position are merged into one frame in native shadowing. Yellow dots represent how path nodes sharing the same horizontal position are merged into one frame in native reading.

The above process enables us to have sequential annotations for each frame of native reading and native shadowing.

DTW between learner reading and native utterance (reading or shadowing)

Finally, these annotations shall be mapped on the learner's utterance. Again, DTW is applied between the learner's utterance and the native's utterance, where it should be noted that the native utterance is either reading or shadowing. This DTW enables us to have sequential annotations for each frame of the learner's utterance. Here, p is the DTW path between the learner's reading and the native's utterance. The score of the i-th frame in the learner's reading is computed similarly as

$$\text{score}_{LR}[i] = \text{avg}(\text{score}_{NS}[p[j][1]]), j \in \{x|p[x][0] = i\} \tag{3.3}$$

or

$$\text{score}_{LR}[i] = \text{avg}(\text{score}_{NR}[p[j][1]]), j \in \{x|p[x][0] = i\}. \tag{3.4}$$

It should be noted that DTW between native shadowing and native reading, and DTW between a learner and a native shadower were conducted with the same spectrum features.

Prediction of subjective scores using multiple word-based objective scores

Through deriving scores on learners' utterances using the above method, for each utterance, there are 7 types of features available in the prediction. For one frame, its score can be aggregated from native reading and native shadowing, and the distance can be calculated by 1) MFCC feature vectors with Euclidean distance, 2) MFCC feature vectors with Cosine distance and 3) posterior feature vectors with Bhattacharyya distance. Thus, there are 6 types of scores from different utterance sources and distance types. Additionally, delay between learner reading and native shadowing was measured.

As shown in Table 3.3, each word in a learner's reading has four scores and in this section, those scores are used to predict their corresponding subjective score. Here, the simplest model of prediction, linear regression, is used. 70% of learners' words are used for training and the remaining are used for testing. Selection of the training data was done repeatedly and differently, and testing was done 1,000 times. Table 3.4 shows the performance of our regression model with different combinations of features.

Correlations of the machine scores to the human scores are shown in Fig. 3.10 and Figure 3.11, where all the four features are used.

From Table 3.4, although linear regression on MFCC and posterior features can effectively predict word-based machine scores, feature combination includes posteriorgram can always achieve a relatively higher correlation while others are not satisfying. Moreover, DTW between

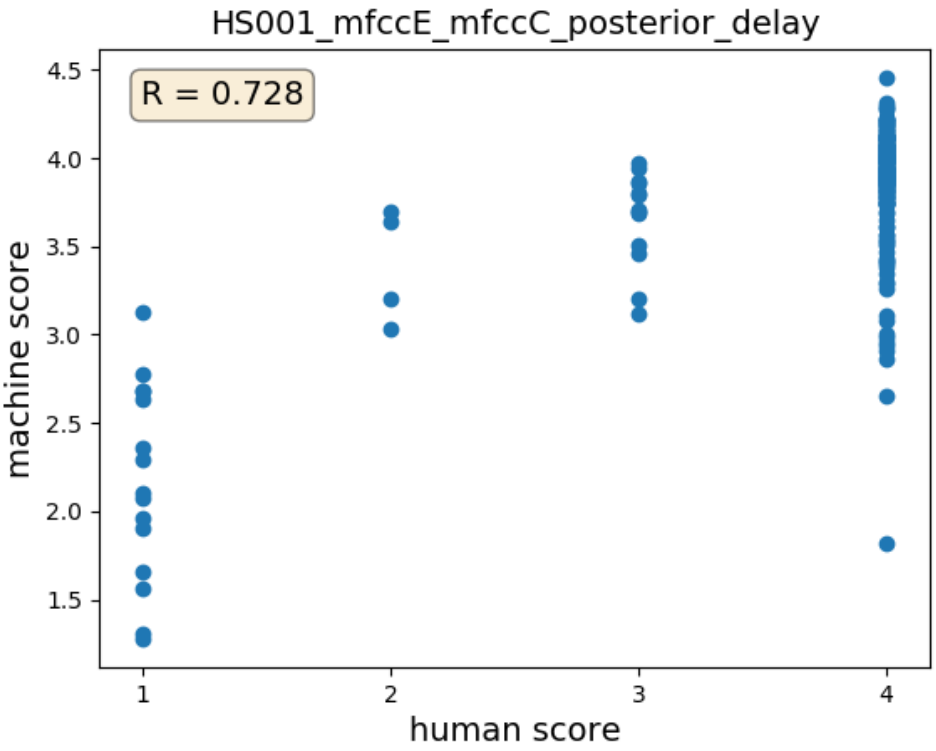


Fig. 3.10 Distribution of regression results (HS001)

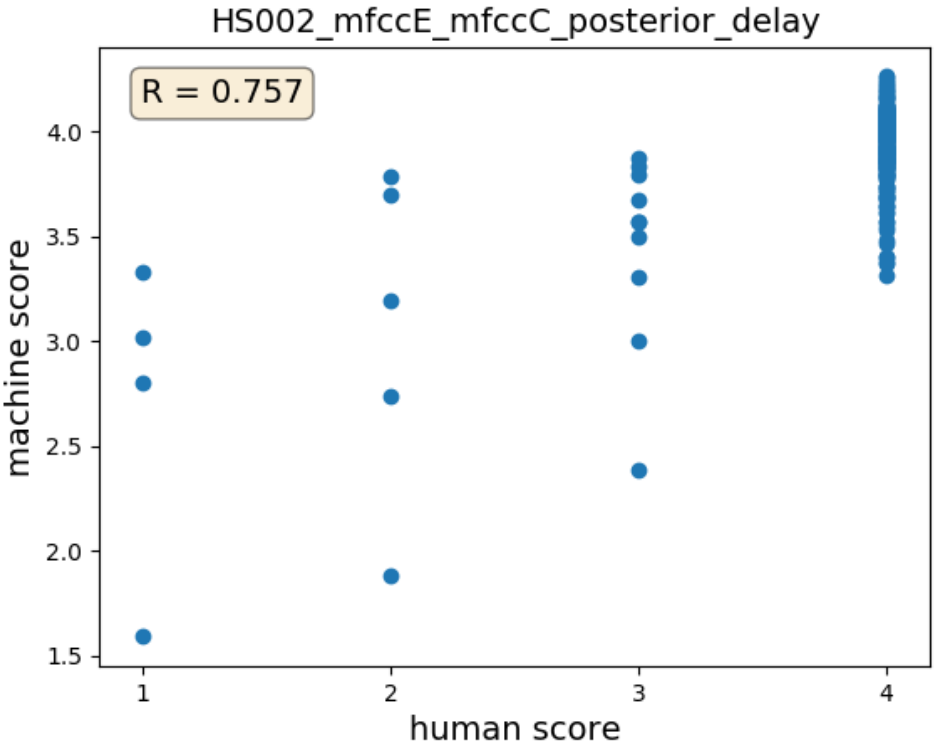


Fig. 3.11 Distribution of regression results (HS002)

Feature	HS001	HS002
MFCC_E	0.503	0.409
MFCC_C	0.230	0.151
Posteriorgram	0.718	**0.761**
MFCC_E + Posteriorgram	0.720	0.758
MFCC_C + Posteriorgram	0.725	**0.761**
MFCC_E + Delay	0.564	0.455
MFCC_C + Delay	0.467	0.354
Posteriorgram + Delay	0.714	0.758
MFCC_E + Posteriorgram + Delay	0.721	0.756
MFCC_C + Posteriorgram + Delay	0.724	0.760
MFCC_E + MFCC_C	0.529	0.419
MFCC_E + MFCC_C + Posteriorgram + Delay	**0.728**	0.757

Table 3.4 Performances of the regression models with different features for each of the two shadowers

native utterances and learner utterances, which are different speakers, is not reliable especially based on MFCC features. Native shadowing and native reading can be an effective method for deriving annotations on learners' utterance when DNN-based ASR frond-end is available. This requires the author to find a way to mitigate this problem and this is introduced in next chapter.

Chapter 4

L2 Speech Assessment based on Comparison between Native Script-Shadowing and Native Shadowing

In this chapter, a new form of native reading is introduced, which is called native script-shadowing, and this mitigates inevitable problems in Chapter 3 because of different speaking rate control and various reading styles.

Through comparing native shadowing and native script-shadowing, more reliable annotations can be derived and annotations are verified by ordinary Euclidean distances and weighted Euclidean distances. The correlation derived by weighted Euclidean distances can adequately replace posterior-based phonemic distances.

Then inter-shadower comparion is introduced. Different shadowers may have different language background and apply different shadowing strategies. How these factors can affect the reliability of the annotation is explained and analyzed in the last of this chapter.

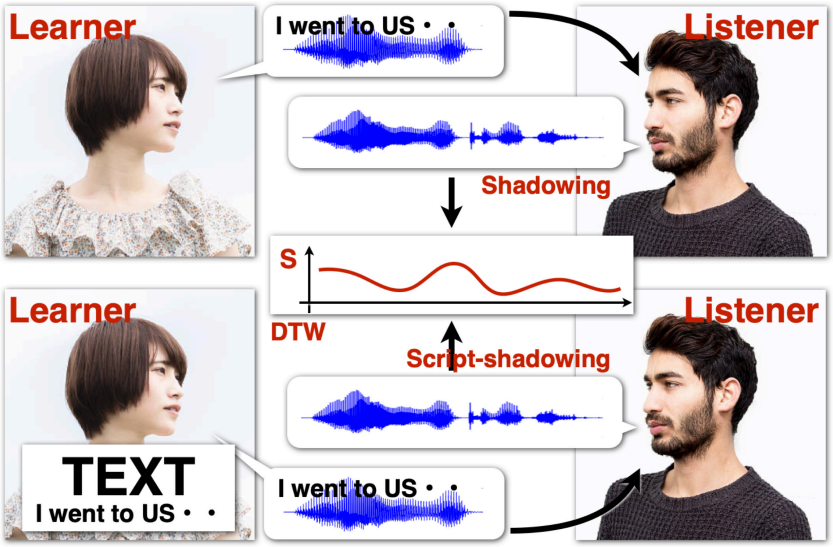


Fig. 4.1 Comparison bet. shadowing and script-shadowing

4.1 Three inevitable problems and a simple solution

The revserse-shadow-and-read method has been proven to be effective with finer granularity. Here, a shadowing is treated as the least prepared speech and a reading is as the most prepared speech. If both utterances are similar, easiness and quickness of understanding should be high. It is a simple principle.

In this method, however, we found three inevitable problems. 1) Different shadowers adopt different shadowing strategies. Some shadowers try to minimize delay of shadowing, paying less attention to articulation, and others try to maximize articulation, paying less attention to delay. When both types of shadowers read in a similar reading style, even in the case that DTW-distances between NS and NR are different between the two types of shadowers, they sometimes assign similar subjective scores. 2) Reading styles can vary among shadowers. L2 utterances are often slow, and when native shadowers read the text, some of them read it quickly. Quick phonation often results in inarticulate phonation even when natives read. This causes a bias when calculating shadowability. 3) Recording NS and NR is done independently. When the two utterances are presented as waveforms to teachers and learners, it is difficult to interpret acoustic gaps between the two utterances.

In this chapter, we introduce a small change of the data collection protocol into the method examined in the previous chapter. Although our solution is not a technical solution, its effectiveness is very high. In the previous section, a shadowing and a reading were viewed as the least prepared speech and the most prepared speech, respectively. To solve the above problems simultenously, we realized that comparison should be made not between a shadowing

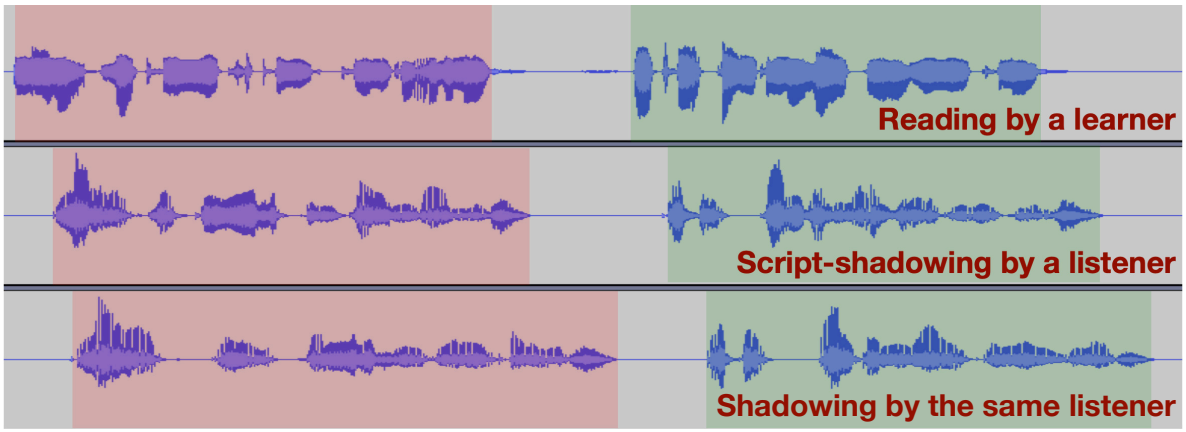


Fig. 4.2 An example triplet of LR, NSS, and NS

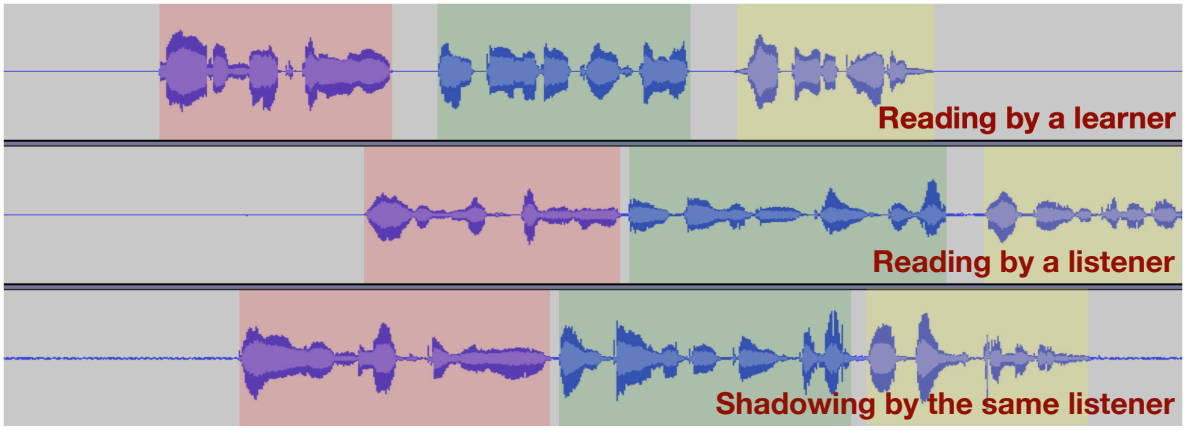


Fig. 4.3 An example triplet of LR, NR, and NS

and a reading, but between a shadowing and the best shadowing. The best shadowing can be obtained by asking a shadower to shadow repeatedly or asking a shadower to shadow with transcript given. Here, we took the second option and, in applied linguistics, this form of shadowing is often called script-shadowing, shown in Figure 4.1.

To analyze utterances of shadowing and script-shadowing, two datasets were prepared. One contains the two types of shadowing utterances from native listeners of English, to whom 30 Japanese English utterances were presented. The other contains the two types of utterances from native listeners of Japanese, to whom 30 Chinese Japanese utterances were presented.

4.2 Easy-to-understand presentation of utterances

Figure 4.2 shows a typical example of a reading from a learner (LR), a script-shadowing from a native shadower (NSS), and a shadowing from the same shadower (NS). Since all the three

Table 4.1 Ratios of phrase lengths in NR, NSS, and LR

	S1	S2	S3	S4	S5	S6	S7
NR/LR	1.40	0.76	0.75	0.82	0.94	1.18	0.80
NSS/LR	1.20	0.81	1.01	0.98	0.99	1.02	0.99

Table 4.2 Correlations of posterior-based phonemic distances and purely acoustic distances

HS1	HS2	both	CR
0.661	0.627	0.658	0.822

CR = Consecutive Recording of NS and NSS.

utterances share the same time axis, the temporal structure of these utterances can be directly compared. For example, as NSS can be viewed as the best shadowing, delay in NSS is short but in NS, it becomes longer. Phrase boundaries are manually visualized with different colors. Without those illustrations, however, even learners can understand that the first phrase become longer in NS compared to that in LR and NSS. This implies that some parts in the first phrase in LR are difficult to understand quickly, and by listening to NS, learners can get to know which parts reduced comprehensibility. Figure 4.3 shows an example of LR, NR, and NS. LR and NS share the time axis but NR was recorded independently of LR and NS. Then, NR cannot be compared to LR and NS visually and directly. Pedagogically speaking, Figure 4.2 is by far more informative than Figure 4.3.

4.2.1 Speaking rate control realized in script-shadowing

In Figure 4.2, the length of each phrase of NSS tends to be similar to that in LR because NSS is basically synchronous reading with LR. Before collecting NS and NSS for Japanese English utterances, NR was also recorded independently. Table 4.1 shows the ratio of utterance-based NR/LR and that of NSS/LR for each of the native shadowers. For every shadower, NSS/LR becomes closer to 1.0. By asking native shadowers to script-shadow, we can obtain utterances temporally aligned to LR.

4.2.2 Acoustic comparison between consecutive recordings

In Chapter 3, NR and NS were recorded in different rooms. and in this chapter, NSS and NS were recorded consecutively in the same room. Posterior-based DTW was conducted for pairs of NR and NS and for pairs of NSS and NS. At each node on the obtained DTW paths, the MFCC

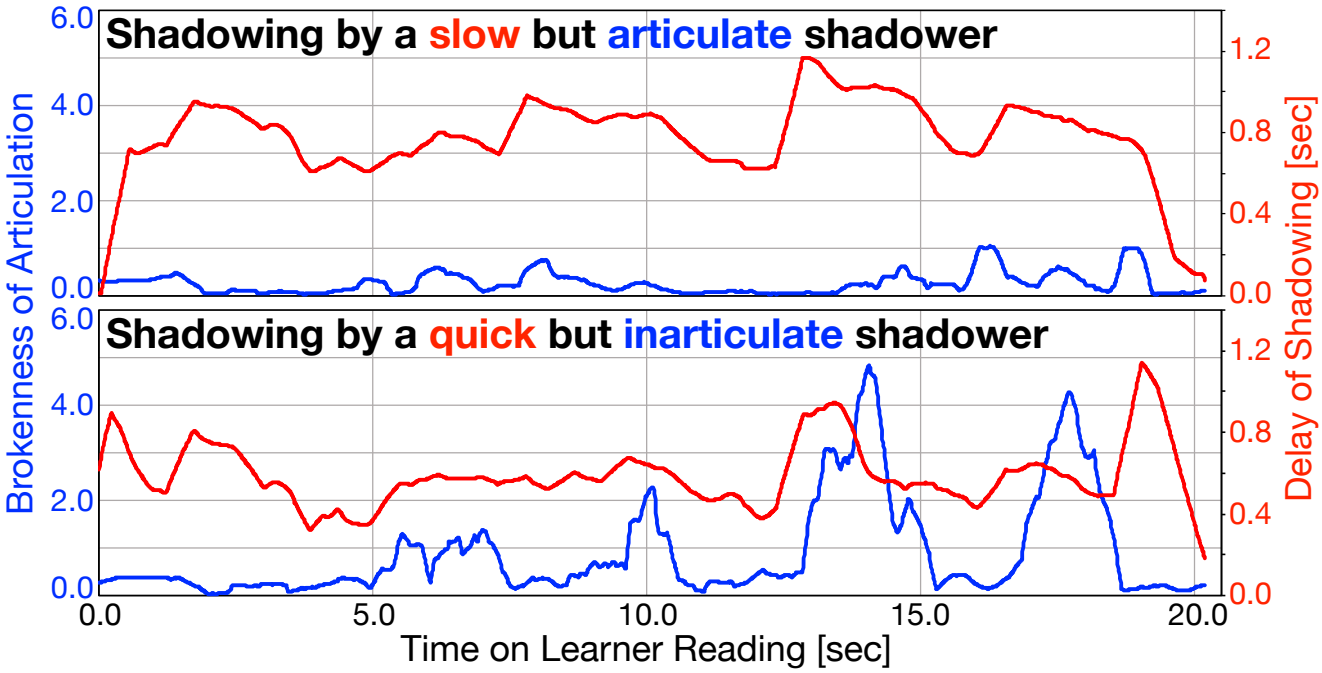


Fig. 4.4 Frame-based annotation of shadowability

distance between the two corresponding frames was calculated. Here, the MFCC distance was calculated as weighted Euclidean distance. Table 4.2 shows word-based correlations between posterior-based distances and MFCC distances. By consecutive recording, posterior-based DTW can be replaced by MFCC-based DTW. The former requires DNN models and therefore this approach is difficult to be applied to minority languages. Consecutive recording guarantees the effectiveness of our approach to any language.

4.2.3 Toward frame-based shadowability annotation

Frame-based annotation of shadowability is calculated and visualized tentatively. Here, moving average is conducted on every 0.5 sec segment with 10 msec shift for the posterior-based DTW scores. Frame-based delays between LR and NS are also averaged in a similar way. Figure 4.4 shows shadowability graphs for a single LR, shadowed by two shadowers. One shadower shadows smoothly but with long delays, while the other shadows not so smoothly but with small delays. In the figure, a strategic difference of shadowing is observed between shadowers.

Table 4.3 Correlations of posterior-based distances and MFCC-based distances among different shadowers

J001	J002	J003	J004	J005	J006
0.62	0.67	0.77	0.81	0.77	0.65

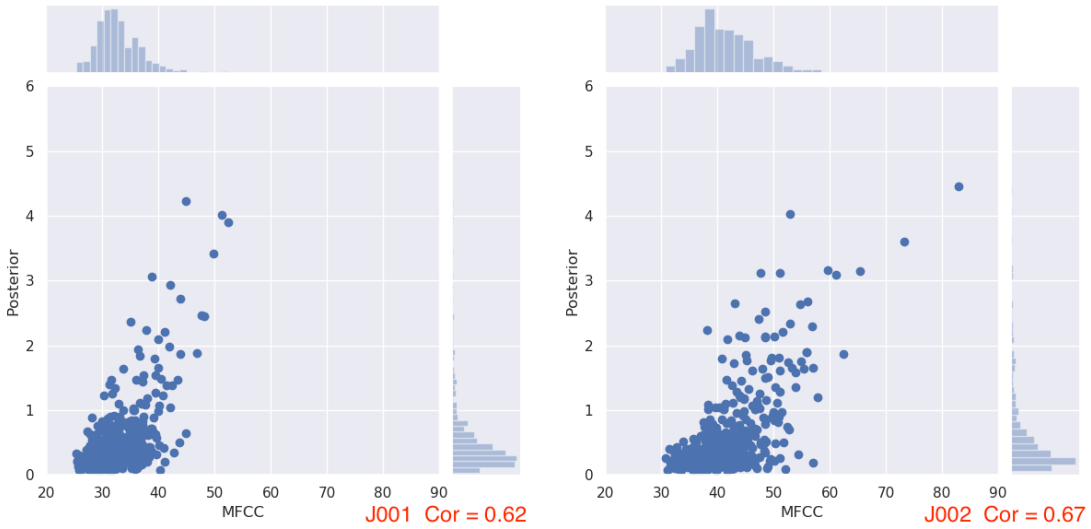
4.2.4 Inter-shadower comparison between consecutive recordings

In Section 4.2.2, the MFCC distance, which was calculated as weighted Euclidean distance, was shown to be adequate to replace posterior-based DTW distance. When the shadower is taken into consideration, shadowers' shadowing may provide us with different annotations when they apply various shadowing strategies. After finishing recording tasks, a native Japanese speaker was asked to pick one good shadower and one bad shadower among these 6 native Japanese shadowers. In this experiment, the native Japanese shadower, whose index was J001, was labelled as a good shadower while J004 was labelled as a bad shadower.

In this section, 30 Chinese Japanese utterances were presented to 6 native Japanese shadowers. Native Japanese shadowers were required to shadow and script-shadow every Chinese Japanese utterance. Each pair of native shadowing and native script-shadowing was compared by posterior-based DTW and MFCC-based DTW. The correlation of posterior-based DTW distances and MFCC-based DTW distances were calculated and listed in Table 4.3. Distributions of the posterior-based DTW distance and corresponding MFCC-based DTW distance were plotted in Figure 4.5.

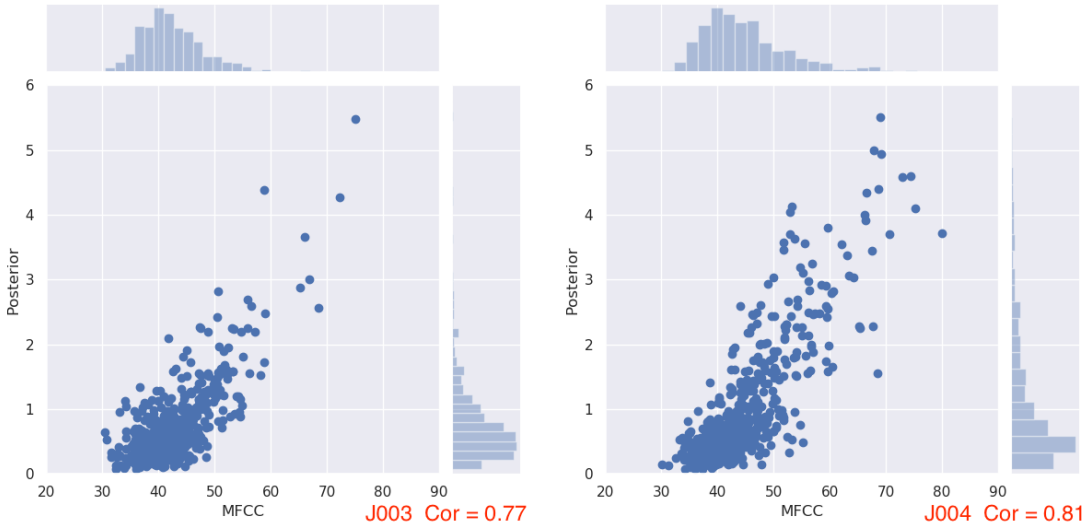
In the table, it can be obviously pointed out that the correlation of posterior-based distances and MFCC-based distances of the good shadower is the lowest one while the correlation of posterior-based distances and MFCC-based distances of the bad shadower is the highest one. After looking into the distribution of the posterior-based distance with corresponding MFCC-based distance, from the perspective of posterior-based distance of MFCC-based distance, J001 has a relatively narrow range while J004 has a wider distribution range.

The relationship of the distribution range and the correlation observed above can be explained by the difference of language background and shadowing strategy. J001 is a good shadower and this may due to his/her past exposure to Chinese language, thus he/she could relatively smoothly shadow the utterance most of the times. As for shadowing strategy, Figure 4.4 actually was plotted based on the shadowing and the script-shadowing from J001 and J004 based on the same learner's utterance. J001 shadowed in a slow but articulate manner, which made him/her a good shadower and covered learners' inexpert language skills. J004 shadowed in a quick but inarticulate manner, which made him/her a bad shadower and his/her performance of the shadowing adequately reflected learners' language skills.



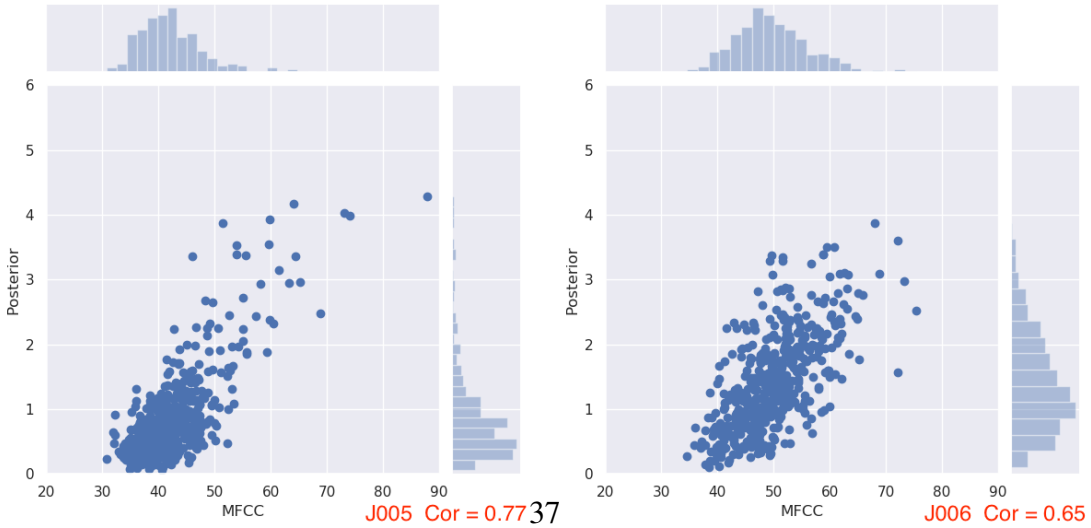
(a) J001

(b) J002



(c) J003

(d) J004



(e) J005

(f) J006

Fig. 4.5 Distributions of posterior-based distances and MFCC-based distances among different shadowers

Even within the range of native shadowers, the results can vary and the reason behind this phenomenon is deserved to be explored. Therefore, in the next chapter, shadowers from different backgrounds are recruited to explore the relationship between the shadower's background and reliability of corresponding sequential comprehensibility annotations derived by acoustic features.

Chapter 5

L2 Speech Assessment based on Comparison among Listeners' Script-Shadowing, First Shadowing and Second Shadowing

In this chapter, unlike Chapter 3 and 4, shadowers are not only native shadowers of the target language. Invited shadowers can be categorized into 3 types: native shadowers of the target language, native shadowers of learners' L1, and other native shadowers whose L1 is not the target language nor learners' L1. Moreover, the target language changes from Japanese to English to validate shadowing-and-reading(script-shadowing) method can generalize to other languages.

5.1 First shadowing and second shadowing

In Chapter 3 and Chapter 4, target language was Japanese and only native Japanese speakers were used as shadowers. In Chapter 4, to replace posterior phonemic features, linear regression on acoustic features was applied and proven to be more effective than ordinary Euclidean distance. Compared with Euclidean distances between acoustic feature vectors, weighted Euclidean distances have higher correlations to posterior phonemic distances. However, correlations among different shadowers vary and it may due to shadowers' various language backgrounds (L1 and exposure to learners' L1).

For the corpus collected in this experiment, English is used as the target language and learners' L1 is Japanese. As for shadowers' reading utterances, shadowers' reading utterances are replaced by shadowers' script-shadowing utterances same as Chapter 4. To normalize the effect caused by shadowers' language background and shadowing proficiency, an addtional task is required for shadowers.

As for shadowers' shadowing utterances, an additional shadowing utterance is required after collecting the first shadowing utterance and this shadowing utterance is called second shadowing. Based on subjective observation, second shadowings, especially provided by shadowers who has the same L1 as learners, demonstrate far more proficient than first shadowings.

5.2 Corpus collection

In this experiment, the target language is English and 30 university-level Japanese students recorded readings of their essaies. These 30 reading utterances were used as learner reading utterances.

As for shadowers, three types of shadowers were invited:

- Native Japanese shadowers who are proficient in English. Some of them are professional interpreters and are proficient in shadowing tasks.
- Native English shadowers, who do not have any knowledge of Japanese.
- Non-native speakers of English or Japanese but very fluent in speaking English. Their L1s and levels of Japanese Language Proficiency Test (JLPT) were also recorded for reference. JLPT has 5 levels: N1, N2, N3, N4 and N5. The easiest level is N5 and the most difficult level is N1 [3].

There are 9 native Japanese shadowers, 1 native English shadower and 6 native shadowers whose L1 is other languages. Shadowers' profiles are listed in Table 5.1.

Table 5.1 Shadower profile

Index	L1	JLPT Level
A001	Japanese	
A002	Japanese	
A003	Japanese	
A004	Japanese	
A005	Japanese	
A006	Japanese	
A007	Japanese	
A008	Japanese	
A009	Japanese	
B001	English	
C001	Vietnamese	L2
C002	Bulgarian/Russian	
C003	Chinese	
C004	Chinese	
C005	Vietnamese	
C006	Chinese	

For each shadower, he/she was required to do following 3 tasks: shadow learners' utterances for the first time, shadow learner's utterances for the second time and perform script-shadowing. Among the 3 tasks, first shadowing and second shadowing could only be done for one time while shadowers could repeat record their script-shadowings for several times if they wanted.

5.3 Procedure of data processing

Unlike Japanese, English sentences can not be divided into bunsetsu-like segments. In this experiment, annotations are calculated in utterance-based patterns. To fully explore the correlations of MFCC-based DTW distance and posterior-based DTW distance, word boundary detection is not applied.

As introduced above, first shadowing (NS1), second shadowing(NS2) and script-shadowing(NSS) were collected for each learner's utterance and shadower pair. MFCC-based DTW and posterior-based DTW are done between two pairs, which means $DTW(NS1, NSS)$ and $DTW(NS2, NSS)$. After calculating distances, two types of correlations are calculated as:

$$Cor1 = Correlation(MFCCDTW(NS1, NSS), PosteriorDTW(NS1, NSS)) \tag{5.1}$$

$$Cor2 = Correlation(MFCCDTW(NS2, NSS), PosteriorDTW(NS2, NSS)) \tag{5.2}$$

For MFCC-based DTW distances, Euclidean distance and weighted Euclidean distance are both tested. Similiar to Chapter 4, same linear regression model is introduced in measuring weighted Euclidean distance. However, utterances in this English read by Japanese corpus collected in this chapter are not used as training data. All utterances in the Japanese read by Chinese corpus mentioned and tried in Chapter 4 are used as the only source of model training data. If weighted Euclidean distance can achieve higher correlations than ordinary Euclidean distance, posterior-based DTW distances of other minor languages can also be predicted effectively by linear regression models trained by major languages with abundant corpus.

5.4 Results and discussions

Table 5.2 shows correlations calculated by different shadowing utterances or Euclidean distance types. The correlation of most shadowers improves from first shadowing to second shadowing, espacially for native shadowers whose L1 is not English nor Japanese. Native shadowers of other languages have a relatively higher correlations after second shadowing. This is due to

Table 5.2 Correlations of posterior-based distances and MFCC-based distances between script-shadowing and shadowing

Index	Euclidean MFCC		Weighted Euclidean MFCC	
	Cor1	Cor2	Cor1	Cor2
A001	0.717	0.750	0.791	0.771
A002	0.714	0.744	0.736	0.764
A003	0.857	0.860	0.773	0.896
A004	0.857	0.852	0.775	0.821
A005	0.746	0.720	0.755	0.711
A006	0.790	0.811	0.873	0.861
A007	0.761	0.760	0.827	0.847
A008	0.852	0.859	0.868	0.861
A009	0.820	0.835	0.889	0.908
B001	0.888	0.848	0.926	0.878
C001	0.746	0.784	0.862	0.884
C002	0.700	0.842	0.814	0.884
C003	0.711	0.811	0.808	0.835
C004	0.811	0.868	0.778	0.917
C005	0.864	0.852	0.889	0.880
C006	0.961	0.978	0.976	0.970

native shadowers' unfamiliarity with Japanese-accented English. Native shadowers of other languages, for example, Chinese and Vietnamese might not have listened to Japanese-accented English and they performed awkwardly in shadowing experiment. This kind of unfamiliarity makes their MFCC-based distances and posterior-based distances to be unstable and sparse on one-dimension Gaussian distribution (variance is larger). Broad range can lead to higher correlations between shadowing utterances and script-shadowing utterances. To be rather the reverse, native shadowers who are familiar with Japanese-accented English or have enough exposure to Japanese English utterances (native Japanese shadowers and English shadowers) could produce stable shadowing utterances and distribution range can be narrow.

As for weighted Euclidean distances, its model is fully trained by a different language. The correlations improve after using weighted Euclidean distances for both first shadowing and second shadowing utterances. This proves that weighted Euclidean distance can replace ordinary Euclidean distance even the language for training corpus and testing corpus are different.

5.5 Content analysis along with acoustic analysis

In [33], text content was analyzed and compared with acoutic analysis using the same corpus in this chapter. When learners' reading scripts and manual transcripts of shadowers' (first or second) shadowing are optimally aligned, some types of errors can be quantified. Assume there are N words in the learner's reading script and this script is viewed as standard answer. After optimally aligning corresponding manual transcript of the shadower's shadowing utterance, there might be some types of errors such as substitution errors(S), deletion errors(D) and insertion errors(I). Percentage of correct words and accuracy are defined as:

$$PercentCorrect = \frac{N-D-S}{N} * 100\% \tag{5.3}$$

$$PercentAccuracy = \frac{N-D-S-I}{N} * 100\% \tag{5.4}$$

In Table 5.3, correlations of posterior-based DTW distances and percentage of accuracy or correct words are calculated. Index A, B, and C mean native Japanese shadower, native English shadower and native shadower of other languages. Index 1 and 2 represent first shadowing or second shadowing. It can be observed that posterior-based DTW distances, which have been proved to be highly correlated with subjective shadowability scores in Chapter 3, are also highly correlated with the percentage of word accuracy and correct rate.

Table 5.3 Correlations of DTW distances and percentage of accuracy/correct words[33]

Index	Correlation(DTW, Accuracy)	Correlation(DTW, Correct)
A1	-0.865	-0.868
B1	-0.790	-0.730
C1	-0.834	-0.843
A2	-0.813	-0.828
B2	-0.665	-0.694
C2	-0.829	-0.805

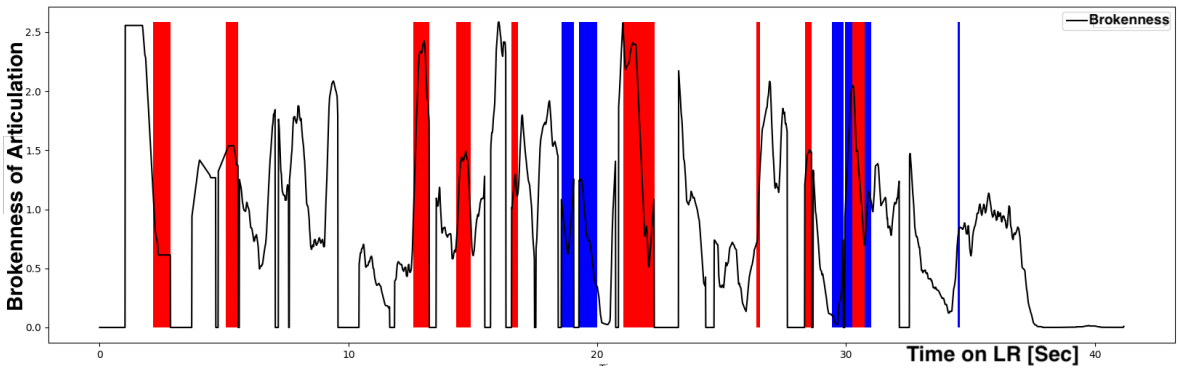


Fig. 5.1 Frame-based annotations of shadowability with altered word range

Figure 5.1 is an example of sequential shadowability annotations derived by the method in 3.4.3 but without linear regression for one learner's utterance and colored substitution/deletion intervals are explicitly rendered. The rendered curve represents aggregated DTW distances corresponding to that frame of the learner's utterance. Red intervals represent durations of words which is changed to be another word in the shadower's first shadowing utterance while blue intervals represent durations of words deleted in first shadowing utterance. Forced-alignment is applied to obtain word boundaries and shadowability annotations of silent frames are reset to be zero.

Most colored intervals in the figure have relatively larger aggregated distances (shadowability annotations). Some other peaks on the curve which are not colored may due to hard to understand but still correct word segment. High correlations of DTw distances to accuracy or correct rate can be reflected in Figure 5.1. Recall the previous work about intelligibility-based evaluation in [19]. Shadow-and-script-shadow method can effectively estimate reliable word accuracy and correct rate through calculating DTW distances which doesn't require human labor to transcribe the text and can eliminate negative guessing effect due to offline transcription.

Chapter 6

Conclusion

6.1 Conclusion

In this study, an effective method for deriving sequential comprehensibility annotations of non-native utterances is introduced and verified. Based on the experiment results in the study, annotations can satisfy following requirements:

- Evaluate the utterance based on comprehensibility.
- Annotate the utterance sequentially.
- Don't rely on experts to manually annotate the utterance.
- The cost should be reasonable.
- Don't require pretrained DNN-based ASR front-end.

From previous works, native listeners' reverse form of shadowings can adequately annotate learners' utterances based on comprehensibility with DNN-based ASR front-end used. To make this method compatible with minority languages, native reading(script-shadowing) is introduced. The distance between the reverse form of shadowing, which is also called native shadowing, and native reading can represent comprehensibility of learners' utterances.

6.2 Future work

In this study, a novel method for effectively collecting sequential annotations is proposed. However, for real applications, a large corpus should be prepared and comprehensibility-based language assessment model should be trained and tested. This requires the researcher to collect large amount of data using the proposed method. How can researchers efficiently gather data? A pratical way called inter-learner shadowing in illustrated in Figure 6.1.

In the figure, every learner has their own L1 and is learning another language. The Vietnamese learner is learning Japanese but can't speaker English. The Japanese learner is learning English but can't speak Vietnamese. The American learner is learning Vietnamese but can't speak Japanese. Vietnamese learner's Japanese utterance is shadowed by native Japanese speaker. Japanese speaker's English utterance is shadowed by native English shadower. English speaker's Vietname utterance can be shadowed by native Vietnamese speaker.

This process can be viewed as speech version of *Lang*-8 [4]. Every learner can help others through its own L1 and receive feedback from the other speakers who speaks his/her target language as L1. If this kind of infrastructure can be constructed, data collection issues can be solved and training a comprehensibility-based language assessment model is possible.

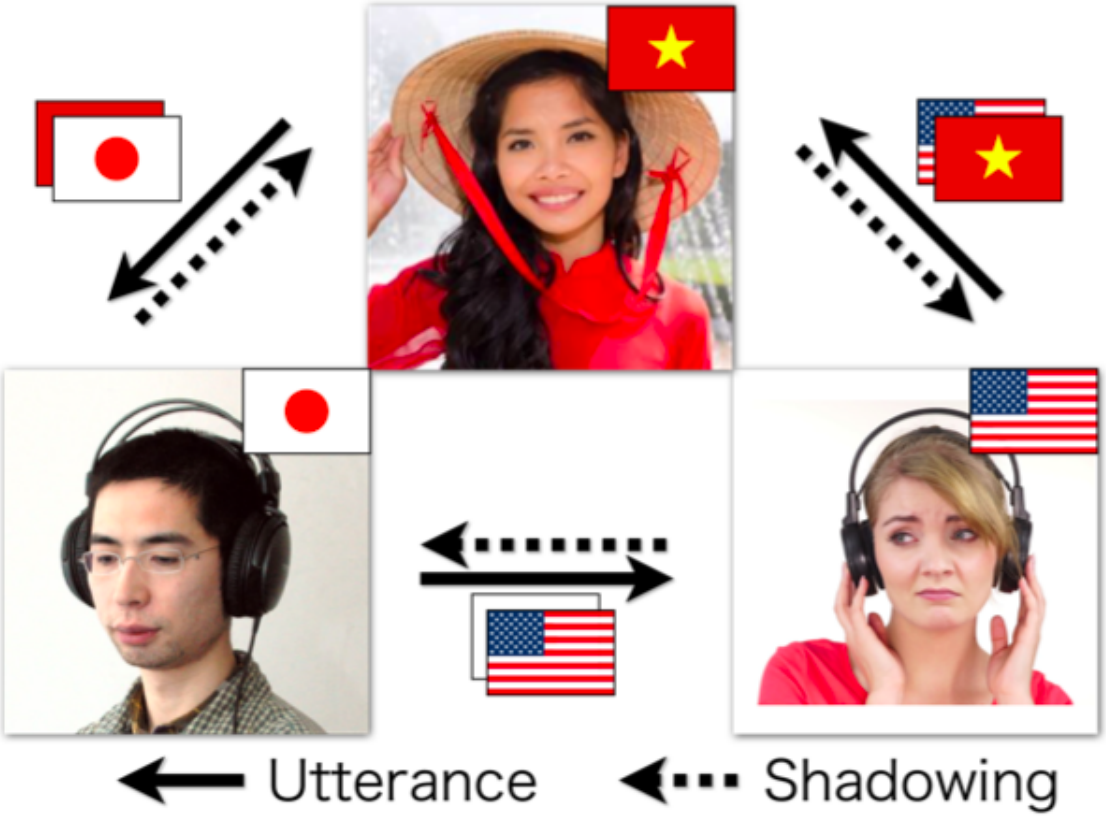


Fig. 6.1 Inter-learner shadowing

After collecting enough data, training a model representing a virtual shadower may be possible. Learners don't need to rely on real speakers to obtain feedbacks on their L2 utterances.

References

[1] Duolingo. URL https://www.duolingo.com/courses.

[2] Liulishuo. URL https://m.liulishuo.com/en/liulishuo.html.

[3] Japanese Language Proficiency Test (JLPT). URL https://www.jlpt.jp/.

[4] Lang-8. URL https://lang-8.com/.

[5] S. Ando, Z. Lin, T. Trisitichoke, Y. Inoue, F. Yoshizawa, D. Saito, and N. Minematsu. A large collection of sentences read aloud by vietnamese learners of japanese and native speaker's reverse shadowings. In *Proc. O-COCOSDA*, pages 1–6, 2019.

[6] J. Bernstein. Objective measurement of intelligibility. In *Proc. ICPhS*, pages 1581–1584, 2003.

[7] David Birdsong. Nativelikeness and non-nativelikeness in L2A research. *International Review of Applied Linguistics in Language Teaching*, 43(4):319–328, 2005.

[8] Maxine Eskenazi. An overview of spoken language technology for education. *Speech Communication*, 51(10):832 – 844, 2009. ISSN 0167-6393. Spoken Language Technology for Education.

[9] Jeremy Goslin, Hester Duffy, and Caroline Floccia. An erp investigation of regional and foreign accent processing. *Brain and language*, 122:92–102, 06 2012.

[10] Avashna Govender and Simon King. Using pupillometry to measure the cognitive load of synthetic speech. pages 2838–2842, 09 2018.

[11] Anja Hahne. What's different in second-language processing? evidence from event-related potentials. *Journal of psycholinguistic research*, 30:251–66, 06 2001.

[12] Yo Hamada. The effectiveness of pre-and post-shadowing in improving listening comprehension skills. *The Language Teacher*, 38(1):3–10, 2014.

[13] Yo Hamada. Shadowing: Who benefits and how? uncovering a booming efl teaching technique for listening comprehension. *Language Teaching Research*, 20(1):35–52, 2016.

[14] Kun Ting Hsieh, Da Hui Dong, and Li Yi Wang. A preliminary study of applying shadowing technique to english intonation instruction. *Taiwan Journal of Linguistics*, 11 (2):43–65, 2013.

[15] W. Hu, Y. Qian, F. K. Soong, and Y. Wang. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154–166, 2015.

[16] Yusuke Inoue, Suguru Kabashima, Daisuke Saito, Nobuaki Minematsu, Kumi Kanamura, and Yutaka Yamauchi. A study of objective measurement of comprehensibility through native speakers' shadowing of learners' utterances. In *Proc. Interspeech 2018*, pages 1651–1655, 2018.

[17] Suguru Kabashima, Yuusuke Inoue, Daisuke Saito, and Nobuaki Minematsu. Dnn-based scoring of language learners' proficiency using learners' shadowings and native listeners' responsive shadowings. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 971–978. IEEE, 2018.

[18] Nobuaki Minematsu, Yoshihiro Tomiyama, Kei Yoshimoto, Katsumasa Shimizu, Seiji Nakagawa, Masatake Dantsuji, and Shozo Makino. Development of english speech database read by japanese to support call research. *ICA*, 01 2004.

[19] Nobuaki Minematsu, Koji Okabe, Keisuke Ogaki, and Keikichi Hirose. Measurement of objective intelligibility of japanese accented english using erj (english read by japanese) database. pages 1481–1484, 01 2011.

[20] Murray J. Munro and Tracey M. Derwing. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1):73–97, 1995.

[21] Murray J. Munro and Tracey M. Derwing. The functional load principle in esl pronunciation instruction: An exploratory study. *System*, 34(4):520 – 531, 2006. ISSN 0346-251X.

[22] Seiichi Nakagawa, Kazumasa Mori, and Naoki Nakamura. A statistical method of evaluating pronunciation proficiency for english words spoken by japanese. In *Proc. INTERSPEECH*, pages 3193–3196, 2003.

[23] T. Pongkittiphan, N. Minematsu, T. Makino, and K. Hirose. Automatic detection of the words that will become unintelligible through japanese accented pronunciation of english. In *Proc. SLaTE*, pages 109–111, 2013.

[24] T. Pongkittiphan, N. Minematsu, T. Makino, D. Saito, and K. Hirose. Automatic prediction of intelligibility of english words spoken with japanese accents – comparative study of features and models used for prediction –. In *Proc. SLaTE*, pages 19–22, 2015.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. B. Glembek, N. Goel, M. Hannemann, P. Motlíček, Qian Y., Schwarz P., J. Silovský, G. Stemmer, and K. Veselý. The kaldi speech recognition toolkit. In *Proc. ASRU*, 2011.

[26] Jane Setter. T. M. Derwing and M. J. Munro. Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research (Language Learning amp; Language Teaching Vol. 42). *Applied Linguistics*, 38(3):430–433, 12 2016. ISSN 0142-6001.

[27] Jieun Song and Paul Iverson. Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents. *Cognition*, 179:163–170, 06 2018.

[28] Tasavat Trisitichoke, Shintaro Ando, Yusuke Inoue, Daisuke Saito, and Nobuaki Minematsu. Influence of content variations on smoothness of native speakers' reverse shadowing. In *Proc. ICPhS*, 2019.

[29] Silke Witt and Steve Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95–108, 02 2000.

[30] Junwei Yue. DNN-based automatic assessment of shadowing speech. *Master's Thesis, The University of Tokyo*, 2017.

[31] Junwei Yue, Fumiya Shiozawa, Shohei Toyama, Yutaka Yamauchi, Kayoko Ito, Daisuke Saito, and Nobuaki Minematsu. Automatic scoring of shadowing speech based on dnn posteriors and their dtw. In *Proc. Interspeech*, pages 1422–1426, 08 2017.

[32] Y. Zhang and J. R. Glass. Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 398–403, 2009.

[33] Chuanbo Zhu, Zhenchao Lin, Nobuaki Minematsu, and Noriko Nakanishi. Analyses on instantaneous perception of japanese english by listeners with various language profiles. *Phonetic Society of Japan (submitted)*, 2020.

Appendix A

Publications

Domestic Conferences and Meetings

- Z. Lin, Y. Inoue, T. Trisitichoke, S. Ando, D. Saito, and N. Minematsu, "Native Listeners' Shadowing of Non-native Utterances as Spoken Annotation Representing Comprehensibility of Non-native Utterances," in Proc. Autumn Meeting of Acoustical Society of Japan, 2019, 2-4-6.
- Zhenchao Lin, Yusuke Inoue, Shintaro Ando, Daisuke Saito and Nobuaki Minematsu, "Native Listeners' Shadowing of Non-native Utterances and Reading of Text Toward Comprehensibility-based Annotation of the Utterances," in Spoken Language Processing, 2019-SLP-130(9), 1-6 (2019-11-29), 2188-8663
- Z. Lin, D. Saito, and N. Minematsu, "An experimental study of sequential annotation for comprehensibility based on native listeners' shadowing and reading," in Proc. Spring Meeting of Acoustical Society of Japan, 2020, 3-P-46.
- A. Yasukagawa, S. Ando, E. Konno, Z. Lin, Y. Inoue, D. Saito, N. Minematsu, K. Saito, "An experimental study of automatic scoring of fluency of spontaneous English utterances by Japanese learners," in Proc. Spring Meeting of Acoustical Society of Japan, 2020, 3-P-44.
- Z. Lin, R. Takashima, D. Saito, N. Minematsu, and N. Nakanishi, "Frame-based shadowability annotation using shadowing and script-shadowing L2 utterances," in Proc. Autumn Meeting of Acoustical Society of Japan, 2020, 3-T3-14.

- Chuanbo Zhu, Zhenchao Lin, Nobuaki Minematsu, and Noriko Nakanishi. "Analyses on instantaneous perception of Japanese English by listeners with various language profiles," in Phonetic Society of Japan (submitted), 2020.

International Conferences and Meetings

- Zhenchao Lin, Yusuke Inoue, Tasavat Trisitichoke, Shintaro Ando, Daisuke Saito, Nobuaki Minematsu. (2019). "Native Listeners' Shadowing of Non-native Utterances as Spoken Annotation Representing Comprehensibility of the Utterances," in Proc. SLaTE, 2019, 43-47.
- Shintaro Ando, Zhenchao Lin, Tasavat Trisitichoke, Yusuke Inoue, Fuki Yoshizawa, Daisuke Saito, Nobuaki Minematsu. (2019). "A Large Collection of Sentences Read Aloud by Vietnamese Learners of Japanese and Native Speaker's Reverse Shadowings," in Proc. O-COCOSDA, 2019, 1-6.
- Zhenchao Lin, Ryo Takashima, Daisuke Saito, Nobuaki Minematsu, Noriko Nakanishi, "Shadowability Annotation with Fine Granularity on L2 Utterances and Its Improvement with Native Listeners' Script-shadowing," in Proc. INTERSPEECH, 2020 (to appear).