DNN-based automatic assessment of shadowing speech

(DNNに基づくシャドーイング音声の自動評価**)**

楽　俊偉

YUE Junwei

ID Number: 37-155023

Supervisor: Prof. Nobuaki Minematsu

Department of Electrical Engineering and Information Systems,
Graduate School of Engineering,
The University of Tokyo

This thesis is submitted for the degree of
Master Thesis

September 2017

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.

YUE Junwei
September 2017

Acknowledgements

I would like to dedicate this thesis to my supervisor Prof. Nobuaki Minematsu and Assistant Prof. Daisuke Saito and my loving family, who have kept giving me great spiritual and material support. I would like to thank all Minematsu-Saito Lab members who have given help to me in both lives and researches, especially Shohei Toyama and Fumiya Shiozawa who have helped me a lot on solving technical problems. Besides I would like to thank all kind people I met who have ever helped me in Japan. I am very grateful to all of them.

This work was supported by JSPS KAKENHI Grant Numbers JP16H03084, JP16H03447, and JP26240022. Thanks to all students and teachers who had participated in this experiment.

Abstract

Shadowing is a practicing strategy which requires a speaker to repeat what he/she heard as soon as possible. Recent years, shadowing has shown its effectiveness in second language learning, and has been adopted as a language education tool. This work mainly focuses on the automatic assessment of shadowing speech using DNN (Deep Neural Network) -based approaches, including DNN-based GOP (Goodness of Pronunciation) score and DNN-based DTW (Dynamic Time Wrapping) distance. Both approaches are tested to have higher correlation with manual scores than traditional GMM (Gaussian Mixture Model) -based GOP scores in a relatively large shadowing speech corpora, which contains 125 speakers. In addition, the DTW approach utilizes the model utterances only, which indicates the DTW approach has language and transcription independency. This is very meaningful to expressive shadowing speech automatic assessment and minor language shadowing speech automatic assessment.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AA Automatic Assessment

CC Correlation Coefficient

CMN Cepstral mean normalization

DCT Discrete Cosine Transform

DNN Deep Neural Network

EM Expectation–maximization

fMLLR feature-space maximum likelihood linear regression

GMM Gaussian Mixture Model

LDA Linear discriminant analysis

MFCC Mel-Frequency Cepstral Coefficients

ML Maximum Likelihood

MLLT Maximum likelihood linear transform

SD Standard Deviation

STFT short-time Fourier transform

SVR Support Vector Regression

TOEIC Test of English as International Communication

Chapter 1

Introduction

1.1 What is shadowing?

Shadowing is a rather new word, and to my knowledge, its definition has not been fixed yet. Lambert [13] defines shadowing as a paced, auditory tracking task which involves the immediate vocalization of auditorily presented stimuli. On the other hand, Tamai defined it as an active and highly cognitive activity in which learners track the speech that they hear and vocalize it as clearly as possible while simultaneously listening [9]. Briefly speaking, shadowing is a practice strategy which requires a speaker to repeat immediately after hearing the speech.

1.2 Shadowing in language learning

Shadowing is usually considered to include processes of speaking, listening and comprehension of speech simultaneously [20], it has been employed as a practicing strategy among simultaneous interpreters to learn how to listen and speak simultaneously. Later it was also adopted by language teachers. Recent decades have seen the effectiveness of shadowing in language learning [8, 9, 11]. [8, 9] showed shadowing can improve students' listening comprehension. [8] also suggested that shadowing can enhance learners' phoneme perception ability. [11] showed that shadowing can improve learners' intonation, fluency, word pronunciation and overall pronunciation. Comparison study suggested that shadowing could be more or at least no less effective than extensive reading, reading aloud and listening in terms of improving speakers' corresponding language skills, i.e. reading comprehension, speaking, and listening comprehension [8, 22, 12].

The reason why shadowing could benefit language learning probably has its foundation in its processing mechanism. Other than simply repeating, shadowing has been shown to involve complex production-perception interaction, automatic semantic and syntactic processing [21, 6], and some people even performed sophisticated error correction during shadowing [19, 18]. This, plus the fact that shadowing is a combined process of speaking, listening and comprehension, suggests that analytical results of shadowing speech can represent the speakers' overall language proficiency better than those of reading speech [17].

1.3 The main problem of applying shadowing practices

Given the benefits of shadowing in language learning, it is getting widely adopted by teachers in many countries, especially for English learners in Japan. In practical shadowing practices, learners need feedbacks for their shadowing speech. This is usually given by their language teacher. However, listening to and assessing all students' utterances is nearly impossible for a daily-level application. Besides, students also want to practice at home, where the teacher cannot give suggestions immediately. Given these circumstances, A fast and precise automatic assessment (AA) technology of shadowing speech is desirable.

1.4 Structure of this thesis

This work focuses on improving the precision automatic scoring of shadowing speech. Chapter 2 introduces the basic knowledge of speech signal processing. Chapter 3 introduces the definition of AA of shadowing speech, and some related previous works. Chapter 4 gives the detail of the corpus used in the experiment, including how it was collected and how the manual scores were given. Chapter 5 gives two proposed approaches in this work, DNN (Deep Neural Network) -based GOP score and DTW distance between model and learner's utterances. Chapter 6 gives experiment settings and results. Chapter 7 summarizes the whole thesis and gives some future works.

Chapter 2

Speech signal processing

This chapter gives a brief introduction to the basic knowledge in speech signal processing, which lays the foundation of shadowing speech automatic assessment.

2.1 Feature extraction

The raw signal of human speech is difficult to analyze directly, so the first step of speech processing is usually feature extraction. The most commonly used feature in speech recognition tasks is MFCC (Mel-Frequency Cepstral Coefficients). Followings are typical steps to compute MFCC features for a speech utterance:

Speech signals to short frames

The characteristics of speech signal are constantly changing, and we often only analyze signals in a small window at a time. This window is usually referred as **frame**. A typical frame has a length of 20 to 40ms, since too short frame length causes unreliability in spectral estimation, while too long frame length gives less information of dynamic characteristics. Adjacent frames are often overlapped by frame shift. The shift distance is typically set to 5 or 10ms.

Figure 2.1 gives an illustration of frame length and frame shift. However, at the cutting edge of frames, signals are not continuous. This may lead to a bad spectral estimation. One way to solve this problem is to add a **window function** onto the original signal, in which **hamming window** is a common choice:

$$w(x) = 0.54 - 0.46\cos 2\pi x, \qquad \text{where} \quad 0 \leq x \leq 1,$$

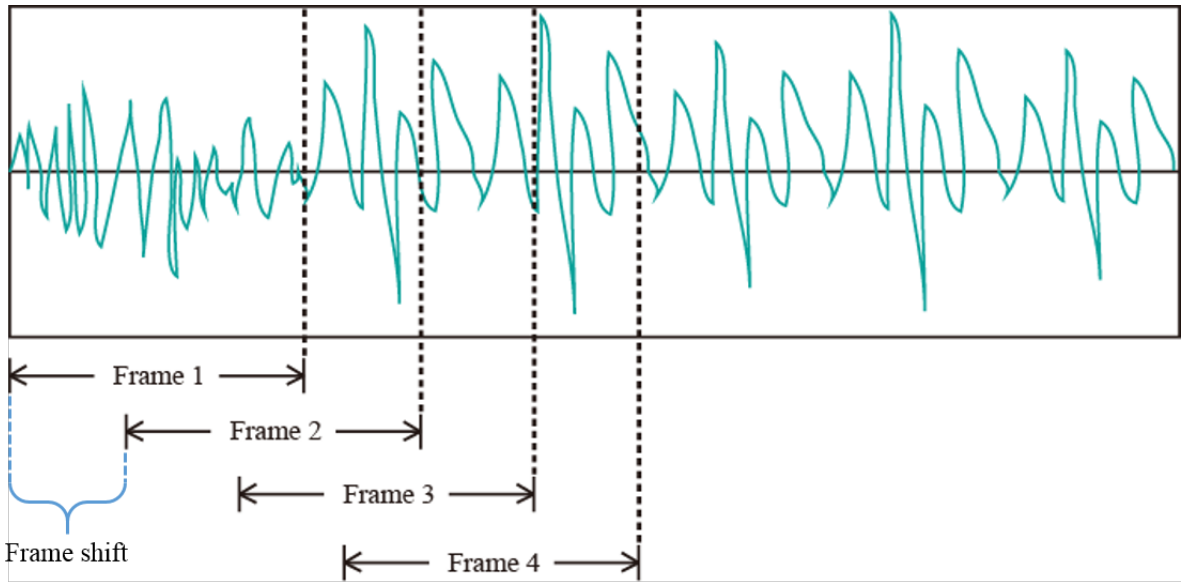


Fig. 2.1 An illustration of frame length and frame shift.

Hamming window makes the change of signals in a frame more smoothly, hopefully results in a better spectral estimation.

Compute power cepstrum

The next step is to calculate the **power spectrum** of each frame [4]. This is motivated by the human cochlea (an organ in the ear) which vibrates at different spots depending on the frequency of the incoming sounds. Depending on the location in the cochlea that vibrates (which wobbles small hairs), different nerves fire informing the brain that certain frequencies are present. This is similarly achieved by applying STFT (short-time Fourier transform) to signal frames, and then transforming it from complex frequency domain into power frequency domain.

Apply the mel filterbank

According to [4], the power spectrum still contains a lot of information not required for Automatic Speech Recognition (ASR). In particular the cochlea cannot discern the difference between two closely spaced frequencies. This effect becomes more pronounced as the frequencies increase. For this reason we take clumps of periodogram bins and sum them up to get an idea of how much energy exists in various frequency regions. This is performed by **Mel filterbank**: the first filter is very narrow and gives an indication of how much energy exists near 0 Hertz. As the frequencies get higher our filters get wider as we become less concerned

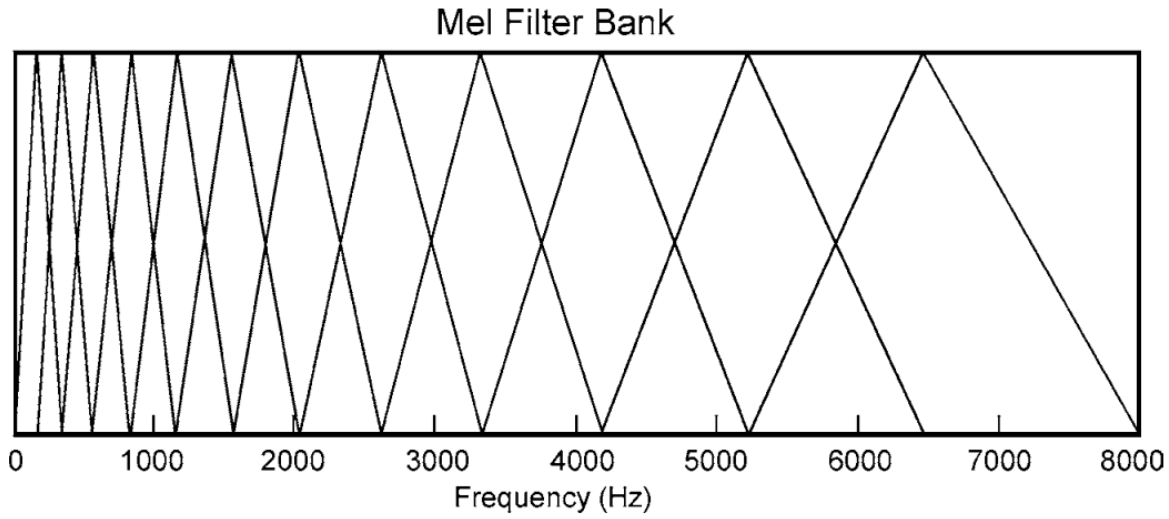


Fig. 2.2 An illustration of Mel filterbank.

about variations. We are only interested in roughly how much energy occurs at each spot. The following gives a formula from frequency to Mel scale:

$$M(f) = 1125\ln(1+f/700),$$

and Figure 2.2 gives an illustration of mel filterbank.

Once we have the filterbank energies, we take the logarithm of them. This is also motivated by human hearing: we don't hear loudness on a linear scale [4].

Take the DCT

The final step is to compute the **DCT** (Discrete Cosine Transform) of the log filterbank energies [4]. This is because our filterbanks are all overlapping, the filterbank energies are quite correlated with each other.

After taking DCT, keep the first 12 coefficients (expect c_0) instead of all of them. This is because the higher DCT coefficients represent fast changes in the filterbank energies and it turns out that these fast changes actually degrade ASR performance, so we get a small improvement by dropping them [4]. Finally, those taken out coefficients are called MFCC.

2.2 Acoustic model

After extracting features, the next thing we need is a bidirectional mapping between human language notations to features. This is generally modeled by probabilistic models. Assume the

intended phoneme sequence is X, the observed features are O, we model their relationship as $P(O,X)$. This can be further broken down into:

$$P(O,X) = P(O|X)P(X),$$

where $P(O|X)$ is usually called acoustic model and $P(X)$ is called language model.

In this section, the traditional GMM-HMM acoustic model and recently popular DNN acoustic model will be explained. The language model will be introduced in the next section.

2.2.1 GMM-HMM model

The HMM is the most popular and successful stochastic approach to speech recognition in general use [7]. The existence of elegant and efficient algorithms for both training and recognition may be the main reason. The HMM, the acoustic model, is required to determine, in conjunction with the language model, the most likely word sequence given some speech data. Specifically within this process, the acoustic model is required to give the probability of each possible word sequence.

A typical three-emitting-state HMM is shown in Figure 2.3. Emitting state cannot be observed directly, but can be estimated by the emitted features. The relationship is modeled by GMM. Suppose f_m is the probability distribution associated with emitting state s_m when given feature vector x, then GMM modeled f_m can be expressed as:

$$f_m(x) = \sum_{n=1}^{N} a_n \frac{1}{(2\pi)^{K/2}|\Sigma_n|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_n)^T\Sigma_n^{-1}(x-\mu_n)\right),$$

where N is the mixture component number, K is the dimension of feature vector, a_n is the weight, μ_n and Σ_n are the mean vector and variance matrix for component n, respectively.

Then the GMM-HMM model can be characterized by [7]:

1. N, the number of states in the model. S_i represents state i in HMM.
2. A, the state probability transition matrix, where a_{ij} represents the transition probability from state S_i to S_j.
3. B, the output probability distribution associated with each emitting state, where

$$b_m(x) = f_m(x),$$

if S_m is the current emitting state and x is observed feature vector.

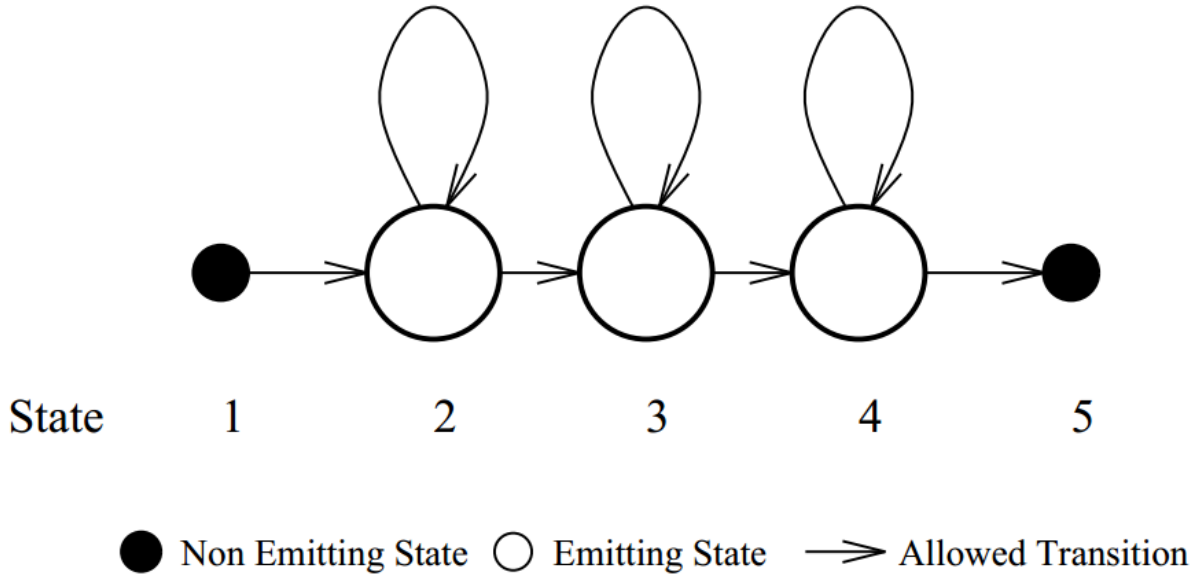


Fig. 2.3 An example of three-emitting-state HMM [7].

4. π, the initial state distribution.

All standard HMMs can be described using above parameters. However, in most speech recognition tasks, HMMs are constrained into **left-to-right** HMMs. That is, the probability transition matrix A is not full. This is consistent with the structure of phoneme pronunciation, and can help prune the HMM structure.

HMMs are usually built for every phoneme, and can be combined together to generate word/sentence phoneme. Figure 2.4 shows an example of combined HMM for two phonemes, /y/ and /i/.

Under all of the assumptions in GMM-HMM model, now we can write the formula $P(O|X)$ for HMM:

$$p_{hmm}(O|X) = \pi \prod_{m=i}^{M} f_{\tau_m}(O_m) a_{\tau_{m-1}\tau_m},$$

where M is the number of feature frames, O_m is the feature at frame m, and τ_m is the emitting state number at frame m.

The parameters can be estimated using EM (Expectation–maximization) algorithm, probably with ML (Maximum Likelihood) criterion [7].

2.2.2 DNN model

According to [10], over the last few years, advances in both machine learning algorithms and computer hardware have led to more efficient methods for training deep neural networks

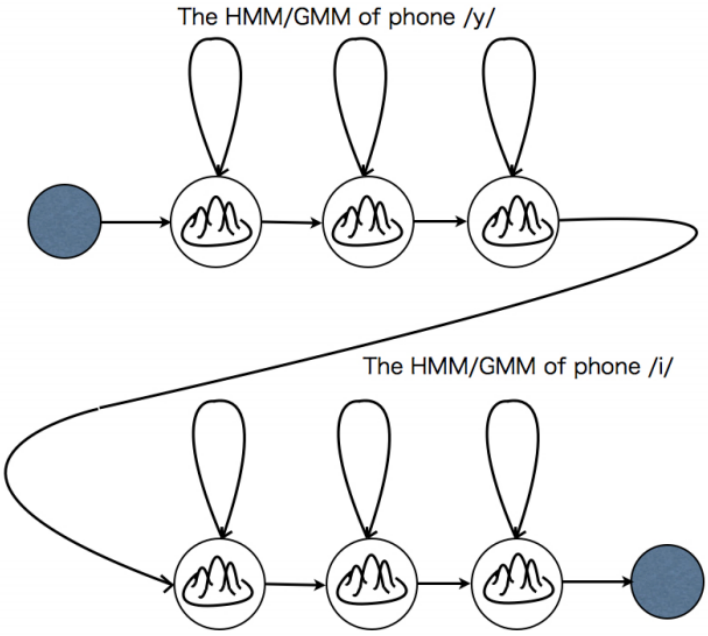


Fig. 2.4 A combined HMM for phone /y/ and /i/ [24].

(DNNs) that contain many layers of non-linear hidden units and a very large output layer. The large output layer is required to accommodate the large number of HMM states that arise when each phone is modelled by a number of different "triphone" HMMs that take into account the phones on either side. Even when many of the states of these triphone HMMs are tied together, there can be thousands of tied states. Using the new learning methods, several different research groups have shown that DNNs can outperform GMMs at acoustic modeling for speech recognition on a variety of datasets including large datasets with large vocabularies.

Technically speaking, a DNN is a neural network with more than one hidden layer. Figure 2.5 gives an illustration of DNN. The left-most layer is called input layer, which accepts a vector (where its dimension is the same as the number of units in input layer) as input. Layers in the middle are called hidden layers, which take the output of its last layer as input, and pass the output to the next layer. The right-most layer is called output layer, which handles the real output of the whole network. Forward-feeding is probably the most common task for DNN. When the parameters in DNN are fixed, DNN accepts a vector as input, passes it through all the hidden layers and gives the result in output layer. The formula for hidden layers $i+1$ is:

$$x_{i+1,j} = h(\sum_{q} w_{iq}^{(i)} x_{iq} + b_j^{(i)}),$$

where x_{ij} is the value of j-th unit in layer i, $w^{(i)}$ is the weights for layer i, $b^{(i)}$ is the bias for layer i, and h is the activation function. Finally, in the output layer, these units are usually

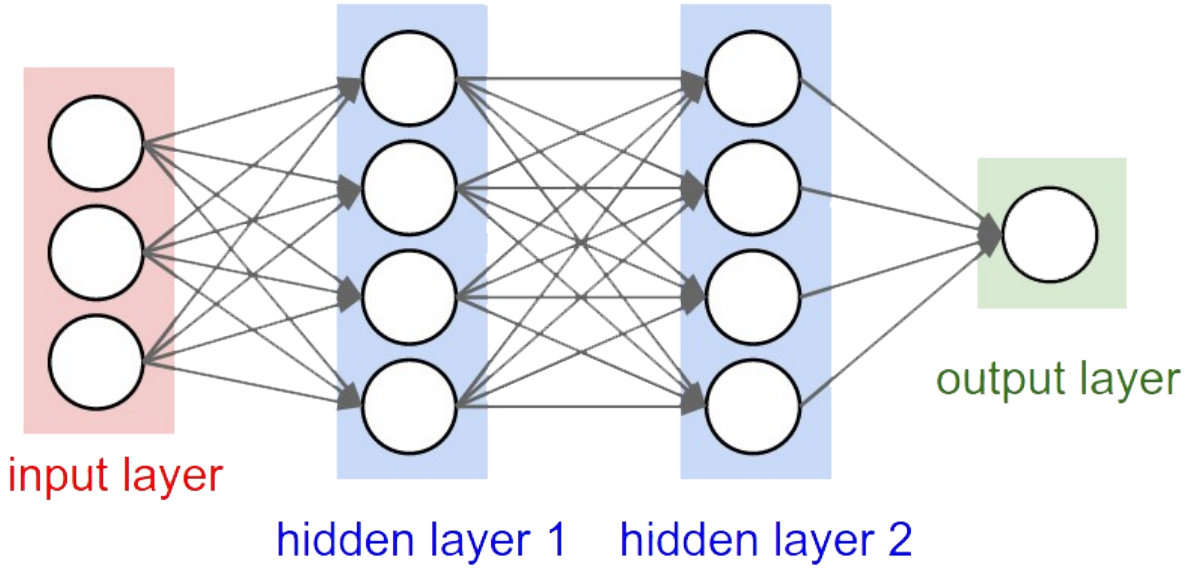


Fig. 2.5 An illustration of a two-hidden-layer DNN.

normalized by, for example, the softmax function:

$$s(x_{ij}) = \frac{e^{x_{ij}}}{\sum_k e^{x_{ik}}}.$$

One of the most important part in DNN is the activation function. This function is the only non-linear part in a standard DNN. Depending on tasks, the following activation function are used:

1. identity: $h(x) = x$
2. sigmoid: $h(x) = \frac{1}{1+e^{-x}}$
3. tanh: $h(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
4. ReLu: $h(x) = \max(0,x)$

All of these activation functions are differential (except for ReLu at 0), which makes them suitable for DNN training. More details about DNN training and integrating DNN and HMM can be found in [10].

2.3 Language model

As previous mentioned, not only the acoustic model $P(O|X)$, the language model $P(X)$ also affects the probability distribution of the target $P(O,X)$.

Language model is a model which measures the naturalness of word sequences quantitatively [28]. Please consider the following examples:

1. There is a cat running on the road.
2. There is a cat flashing on the road.

Perhaps most would agree that the first sentence is more natural than the second one. Humans can judge this using the daily life experience, but how about machines? The relationship among observed words may be one of the criterion. In this case, the word *flashing* is considered less relevant to the context *cat* and *road* than the word *running*, making it harder to appear in the sentence. A machine can judge this kind of relationship statistically by feeding it a large number of texts.

Consider a simple case of the above relationship. Every word in the word sequence is only relevant to the words before it, then the appearing probability of the whole sequence is the product of each word in it given the words before it. This can be quantitatively described as:

$$P(X) = \prod_{i=1}^{n} P(x_i|x_1,\ldots,x_{i-1}),$$

where $X = \{x_1,\ldots,x_n\}$ is the word sequence.

This can be further simplified by assuming that one word is only revelant to limited words coming before it. This simplified model is called **n-gram** when only the last N words are considered. As a result, $P(x_i|x_1,\ldots,x_{i-1})$ can be expressed as:

$$\begin{aligned} P(x_i|x_1,\ldots,x_{i-1}) &= P(x_i|x_{i-N-1},\ldots,x_{i-1}) \\ &= \frac{\text{Count}(x_{i-N-1},\ldots,x_i)}{\text{Count}(x_{i-N-1},\ldots,x_{i-1})} \end{aligned} \tag{2.1}$$

Here, $Count(x_{i-n},\ldots,x_i)$ is the number of appearances for word sequence $x_{i-n},\ldots,x_i$ in the corpus.

This model may seem to be too simple for a good estimate of $P(X)$, but the fact is, it works out very well, especially when a large amount of learning corpora is provided. In addition, the computation cost for n-gram model is very low. These features have made n-gram the most widely used language model in automatic speech recognition tasks.

2.4 Decoding and forced alignment

There are two basic tasks for speech processing, **forced alignment** and **decoding**. They are both commonly used in model training, speech recognition and speech assessment tasks.

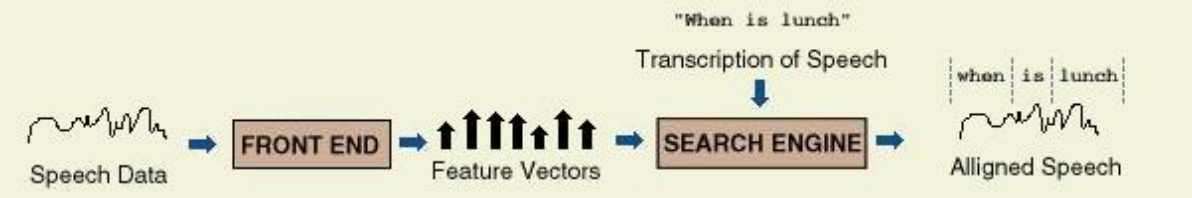


Fig. 2.6 The flow chart for a standard forced alignment [1].

Decoding is a task that when only the utterance is given, we find the most possible word sequence for it. The most common way to do this is to build a decoding network combining both acoustic and language models, then apply Viterbi Algorithm to search for the best path.

Forced alignment on the other hand, requires both the utterance and its transcription. The purpose is to find the boundaries for each phoneme, as shown in Figure 2.6. The searching algorithm is also Viterbi Algorithm. The result of forced alignment is usually used to train a larger acoustic model.

Chapter 3

Automatic assessment of shadowing speech

In this chapter, we will introduce the basic ideas and previous works of AA of shadowing speech. Comparing to normal speech, shadowing speech is usually content-fixed, which means we already know the transcriptions when assessing. Another difference is that shadowing speech is required to be time aligned with the model speech. These features lead to some different approaches of assessment. The purpose of shadowing automatic assessment is to give learners manual-like feedbacks.

Basically, there two approaches to assess shadowing speech automatically: **scoring** and **error detection**. The concepts and related works will be introduced in next two sections. Then the purpose of this thesis will be explained.

3.1 Scoring

3.1.1 Introduction

Scoring means assessing learners' utterances by giving scores. Although the purpose is trying to give manual-like scores, which are usually given by language teachers, the scoring strategy varies for different scorers. Actually, before starting scoring utterances, scorers often have meetings to discuss, to determine which aspects will be scored and how. The same kind of discussion is also needed for automatic scoring.

Suppose the model utterance says "I lighted a candle.", some commonly used scoring aspects of shadowing speech are listed below:

1. **Phoneme**.
   Phoneme aspect is assessed based on the ability to produce each phoneme clearly. For

example, in English shadowing, the difference between phone /l/ and /r/ cannot be distinguished for low-level Japanese learners. Also, some unexpected phonemes may be inserted, for example Japanese phone /ru/ may be inserted at the end of word *candle* for Japanese learners. In our collected data, this happens more frequently than in reading aloud speech recording, probably because shadowing requires higher cognitive load so that learners don't have time to prepare for a clear pronunciation.

In phonetics, the spectral envelope corresponds to the shape of vocal tract, and the shape of vocal tract determines vowels and many consonants. MFCCs are often considered as a good representation of spectral envelope, so to score the goodness of phoneme pronunciation, MFCCs are widely used as in speech assessment tasks.

2. **Prosody**.
   Prosody means the pitch change, or F0 change physically, of speech. It consists of many concepts, such as **tone**, **accent**, **stress** and **intonation**. Not all of these factors are required for a specific language. For example, English has no tone, but is very sensitive to stress and intonation. The meaning of the word *increase* changes depending on its stress position: /in'kri:s/ is a verb meaning "to become larger", while /'inkri:s/ is a noun meaning "an amount by which something increased". On the other hand, Chinese is a tone language, but less sensitive to stress. Stress in Chinese is usually used for emphasizing specific words (or specific syllables).

   Prosody is very important for communication since it has functions of indicating the main subject and borders between semantic segments. It is thus, important for shadowing speech scoring. Nearly all of these prosody concepts are related to pitch. This gives us an idea that F0 information could be useful for scoring. Actually, in speech recognition of tone languages such as Chinese, F0 can be directly fed as an input feature.

3. **Correctness**.
   Correctness means the ability to produce correct word sequences given by the model utterance. Correctness may contain several kinds of errors, such as grammatical errors, omission errors, insertion errors and so on. For example, for a model utterance "I lighted a candle.", some possible errors in the aspect of correctness can be:

   - "I lighted a . . . ". This is an omission error.
   - "I lighted some candles". This one contains two substitution errors.
   - "I light a candle". This one is more complicated, since it could be regarded as both an substitution or a grammatical error.

- "I (hmm) (mirr) . . . ". Some pronunciations are not clear enough to distinguish in the utterance. We refer this as mimic error, since the speaker doesn't understand the meaning of what he heard and he is just trying to say something.

It's difficult to make a complete category for these correctness errors since multiple errors can happen at the same time and the borders of them are not clear.

Correctness can be automatically detected by using a grammar network with extra paths for possible errors. This will be discussed in later sections.

4. **Delay**.
   Delay is the difference between starting time positions of the model utterance and the learner's utterance. It is one of the major differences between shadowing speech and other speeches, such as reading aloud and repetition. High-level shadowers can repeat the model utterance in a very short delay, sometimes as short as 150ms [20] to 254ms [18].

   The scoring of delay is rather simple. Teachers could just compare the starting time positions of two utterances and judge whether delay is in acceptable levels. The same can be done for machines, using forced alignment to detect the starting and ending time positions for each word.

Not all of these factors are needed to be taken account into. Low-level shadowers often struggle with standalone pronunciations so more attention can be paid on the phoneme and correctness aspects. On the other hand, high-level shadowers tend to have more grammatical errors and insertions though the meaning of the whole sentence doesn't change.

In the next section, a universal assessment criterion of transcribed speech will be introduced.

3.1.2 GOP score

Introduction

GOP (Goodness of Pronunciation) score is an objective score indicating the clarity of utterance/word/phone. It is universal and easy to compute, and thus widely used in speech assessment tasks.

In its computation, it is assumed that the orthographic transcription is known and that a set of HMMs is available to determine the likelihood $P(O^{(p)}|p)$ of the acoustic segment $O^{(p)}$ correspond to each phone p [30]. Under this assumption, GOP score is defined as the

normalized log posterior probability of phone p when given the segment $O^{(p)}$. That is,

$$\text{GOP}(p) = \frac{1}{D_p} \log(P(p|O^{(p)})), \tag{3.1}$$

where D_p is the number of frames of audio segment $O^{(p)}$.

Break down equation 3.1 by Bayes' theorem and we get:

$$\text{GOP}(p) = \frac{1}{D_p} \log \left(\frac{P(O^{(p)}|p)P(p)}{\sum_{q \in Q} P(O^{(q)}|q)P(q)} \right), \tag{3.2}$$

where Q is the set of all phones.

Assuming all phones have the same prior probability (i.e., $P(q) = P(p)$ for all $p, q \in Q$), and the sum in denominator can be approximated by its maximum, we have:

$$\text{GOP}(p) \approx \frac{1}{D_p} \log \left(\frac{P(O^{(p)}|p)}{\max_{q \in Q} P(O^{(q)}|q)} \right). \tag{3.3}$$

The numerator is exactly the likelihood of phone p. The denominator is the biggest likelihood among all phones in Q, which is exactly the decoding likelihood of audio segment $O^{(p)}$.

By this approximation, a GMM-HMM based acoustic model is capable of computing GOP scores fast.

Application in shadowing speech assessment

Luo et al. [16] adopted GOP scores as measurements of English shadowing proficiencies of Japanese learners. The flow chart of its computation system is shown in Figure 3.1.

First MFCC features are extracted from the raw waveforms. Then these features are used to decode the utterance using a GMM-HMM based acoustic model and a phone-loop network grammar. The decoded phoneme sequence and their likelihoods are computed. The likelihood for each phoneme is the denominator of Equation 3.3. The transcription of the utterance is used together with the MFCC features to perform forced alignment so the time information and likelihood of each phoneme in the transcription could be determined. The likelihood here is the numerator of Equation 3.3. Finally these likelihoods are combined together and normalized, and then we obtain the GOP scores for each phoneme in the transcription.

There are 27 participants in the experiment. They are language teachers, intermediate leaners and beginners from Japan. They are all asked to participate TOEIC (Test of English as International Communication) tests so we can know their true English proficiencies. During the

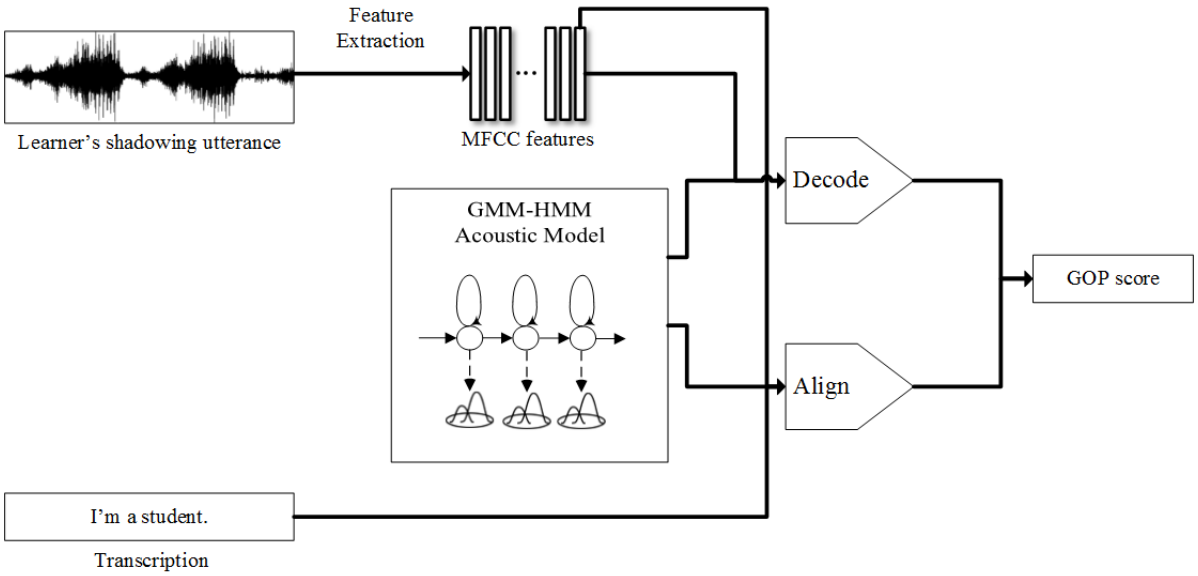


Fig. 3.1 Flow chart for GMM-HMM acoustic model based GOP computation in [16].

experiment, the participants are required to shadow 21 sentences which topic is very familiar for Japanese. Phoneme-level GOP scores are computed first, and then the scores are averaged in utterance-level. Finally the utterance-level GOP scores are averaged by speakers so we have the GOP scores for all 27 speakers.

The result was quite good at that time. The best correlation of speaker-level GOP scores and their TOEIC scores is 0.82, which is close to the inter-rate correlation between language teachers' manual scores. However, the size of this data set is too small to give a convincing conclusion about the relationship between GOP scores and TOEIC scores. There are also some other problems about this study. We will discuss them in the later sections.

3.2 Error detection

Introduction

Error detection is another way of giving feedbacks to the learners. Scoring only cares about giving a score to the input utterance, while error detection tries to find out specific errors in it. Usually the detected error is fed back together with some suggestions about how to improve it. This kind of assessment is ideal for learners' self improvements, but also very difficult to realize.

Table 3.1 Error definitions of shadowing speech in [27].

Name	Description	Example
Substitution	Substitution means one word is substituted by another.	symptoms ⇒ sentences
Omission	Omission means one expected word is missed.	had (been) poisoned
Grammatical Error	Errors related to tense and grammar.	works ⇒ worked
Insertion	Insertion means one unexpected word is found.	(the) symptoms
Repetition	Repetition means a word is fully or partially repeated.	very (very) expensive
Multi-to-one	Multi-to-one means several words are arranged to a set of syllables.	two hundred dollars ⇒ two hundo
Mimic	Mimic means a word is shadowed without understanding its meaning.	
Spoken noise	Filled pause.	<uh>, <en>
Non-spoken noise	Noise other than spoken noise.	<microphone>
Whispering	Whispering means the speaker shadowed in very low voice.	

Related works

Shi et al. [27] have made a study of detecting simple omission errors and improving the correlation between GOP scores and TOEIC scores.

The basic idea of this study is, the number of omission errors have deep relationship with the shadowing proficiencies. To investigate this, this study first collected shadowing utterances from 20 Japanese students (10 men and 10 women). There are 21 utterances for each speaker. All occurrences of errors in these utterances are labeled according to a pre-defined error table (Table 3.1). The statistics of labeled results are shown in Figure 3.2 and Figure 3.3.

We can see the most frequent error type is absolutely *omission*. It is easy to interprete since shadowing is a highly cognitively loaded task so that when speakers cannot catch up with the model utterance they simply tend to keep silent. From Figure 3.3, it can be concluded that the higher the learner's shadowing proficiency is, the fewer omission errors he would have made. This suggests a strong relationship between the shadowing proficiency and the number of omission errors. Thus this study aims to detect omission errors and utilize them to improve the scoring result.

An omission-tolerant aligning grammar network is prepared to detect omissions. Every word in the network could be replaced by a short pause. Viterbi algorithm will search for the most possible path for the whole word sequence so the short pauses in alignment result are omitted words. The accuracy of omission detection is about 70%.

After omission detections, some extra features are computed for each utterance. For example, the omission rate of words, the duration of silence part. The GOP scores are one of these features, computed the same way as in [16]. Finally these features are used to estimate learners' TOEIC scores via SVR (Support Vector Regression).

In the experiment, the GOP scores are used as the baseline. About 40 speakers participated. When using the same content in [16], the new approach (Corr.=0.83) just outperforms the baseline (Corr.=0.82) a little bit. However, in the experiment of another textbook, the correlation coefficient of baseline drops to 0.61. The new approach improves it to 0.74 at best, which is quite effective.

This study shows us a possibility of doing error detection for shadowing speech. In the next section, we will describe the purpose of this thesis, including the problems in previous studies and how to improve them.

3.3 Aim of this thesis

This thesis will focus on introducing our works about automatic scoring of shadowing speech. As previously mentioned, these previous works all suffer from the following problems:

1. **Insufficient amount of data**.
   None of the previous experiments have more than 40 participants, which is considered insufficient for investigating the relationship between shadowing proficiency and automatically generated scores.

   In this our experiment, we collected English shadowing utterances from 125 college students in Japan for a better examination.

2. **Instability of GMM-based GOP score**.
   [27] has shown that the GOP score is not always well correlated with the shadowing proficiency.

   In this study, the DNN-based acoustic model is adopted so that estimation of phoneme posterior probabilities are more accurate and the GOP score can be computed without approximating.

3. **Reliability of TOEIC test**.
   Despite the fact that speaking ability is essential to shadowing, TOEIC test doesn't contain any speaking test. So one with high TOEIC score is not necessarily good at shadowing, and vice versa. The reliability of adopting TOEIC score as the true shadowing proficiency is doubtful.

   In this study, we use manual scores by language teachers instead of TOEIC scores. Manual scores also give us evaluations in sentence and phrase level, while TOEIC scores can only be provided in speaker level.

4. **Dependence on transcriptions of model utterances and acoustic models of the target language**.
   Although this is not a big problem for shadowing since shadowing contents are always transcribed, some usages are limited for this kind of approach. For example, this approach probably doesn't work well for shadowing of minor languages such as Icelandic, because there are no largely collected corpus to train a good acoustic model. This approach is also not suitable for expressive shadowing, because the transcription does not contain information about emotion control such as intonation change and vowel prolonging.

   We proposed a new approach to score shadowing speech by measuring the distance between the model and learners' utterances instead of using GOP scores. This approach utilizes the model utterance, not using transcription directly. It is thus suitable for language-independent shadowing assessment and expressive shadowing assessment.

These solutions would be explained in detail in the next chapter.

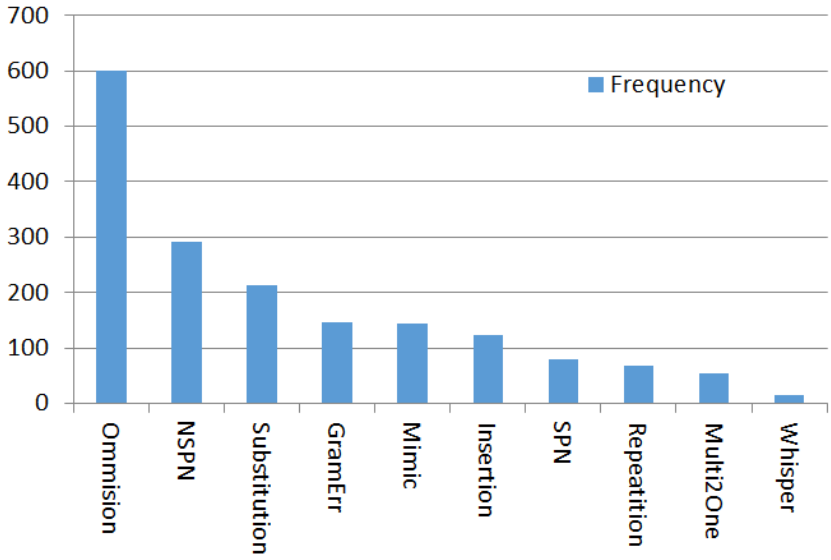


Fig. 3.2 Labeled error statistics (overall) in [27].

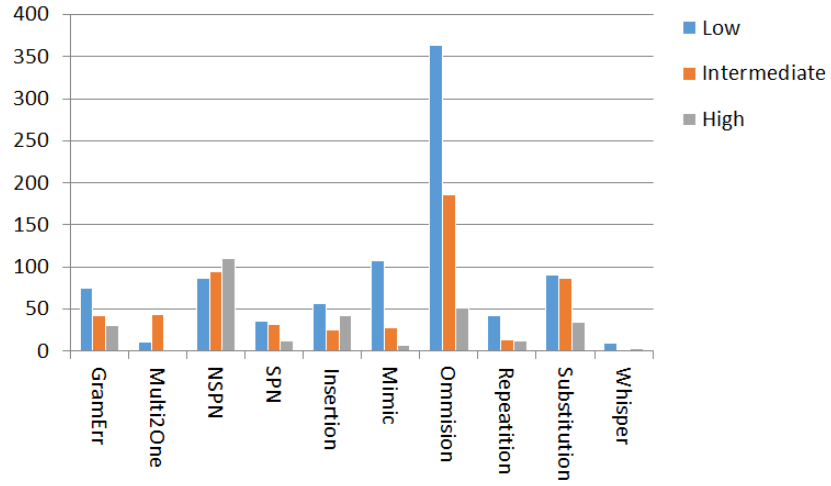


Fig. 3.3 Labeled error statistics (proficiency grouped) in [27].

Chapter 4

Corpus description

4.1 Corpus collecting

In order to examine the relationship between automatic scores and manual scores in a larger scale, we collected English shadowing speeches from 125 college students from Japan. They are all English learners from 3 universities, called University K, University S and University A.

These students are required to do parallel shadowing, which means no transcriptions are shown. The shadowing contents are 55 sentences from 4 passages, whose titles are “My name is Akira”, “MacDonald”, “Valentine” and “Fugu”. They are all considered familiar topics to Japanese learners. Some of the sentences are picked up and shown in Table 4.1. As you may see, some of them are easy and others are difficult. Each passage is shadowed 4 times (without transcriptions). There is a small quiz between the second and third recording to help learners understand the meaning of each passage. 4 or 5 selection questions are contained in a quiz. Table 4.2 gives some example questions in these quizzes.

To collect the utterances fast and correctly, an online shadowing recording website is made. This website is capable of shadowing practice and shadowing recording. Locally recorded utterances will be sent to servers in our laboratory. Introductions and manuals are distributed to the students before the real recording. Figure 4.1 is a screenshot for this recording website.

Because unexpected noises will affect the automatic assessment dramatically, we have also paid attention to the recording devices to make sure that utterances are recorded in good environments. Based on our observations, most learners are not familiar with recording devices and do not know how to use them correctly. To solve this problem, we prepared ear-hook type microphones to fix the distance between the speaker’s mouse and the microphone. This prevents pop noises effectively in our tests. Figure 4.2 is a photo of such kind of microphone set (made by ourselves).

Table 4.1 Some sentences picked up from the shadowing contents.

Passage	Sentence index	Content
My name is Akira	1	Hi, my name is Akira.
My name is Akira	4	I'm studying photography too. Shall we exchange some photos we've taken, and discuss them on the Internet.
MacDonald	1	The MacDonald's house has been broken into.
MacDonald	10	It was you who kicked the door, wasn't it?
Valentine	1	February 14th is a day for people who have fallen in love.
Valentine	12	People wrote their own words on the cards. Usually a kind or funny message.
Fugu	1	In 1996, three men in California were taken into a hospital with strange symptoms.
Fugu	4	The hospital doctors thought the men had been poisoned but couldn't work out what's wrong with them.

4.2 Manual scoring

4.2.1 Introduction

Due to the unreliability of TOEIC scores, we decide to manually score each collected utterance. However, to compare with previous works, each learner was still asked to take a mini-TOEIC test, which contains only the listening part of the regular one. Their scores are rescaled from 0 to 100 for convenience.

We would like to manually score all 55 utterances for every speaker at first, but soon we found it would be too much work to do. Then we decide to decline the amount of sentences. Finally 10 sentences are chosen manually by an English teacher who has taught Japanese students English for over 20 years. These sentences are chosen in a criterion of difficulty balance. That is, both easy and high-level sentences are chosen. Furthermore, only the forth time recorded utterances will be scored. Now we have $10 * 125 = 1250$ utterances to be scored in total. By our observation, a few students tend to only shadow in the first half of a sentence, and keep silent in the other half. This is because when they could not understand the first part, they are most likely to miss the other part too. For this reason, we further divided every sentence into 2 or 3 phrases depending on its syntactic structure. As a result, we have 27 phrases among these 10 sentences.

Table 4.2 Examples of quiz questions used in shadowing recording.

Question and Choice	**Contents**
Question-1	If you see how a person sleeps, what can you predict?
A	The person's kindness
B	The person's character
C	The person's health condition
D	The person's sleeping hours
Question-2	What is the personality of a person who sleeps like a solder?
A	friendly
B	outgoing
C	easy-going
D	quiet

The next problem is how to score these phrases. Our scorers are three language teachers born and grown up in America. Two of them (named SA and SB) are American-Japanese halves, while the other (named SC) has experience in teaching Japanese. All of them are familiar with English learners in Japan, so they are able to judge those utterances objectively from an educational perspective.

After a discussion, the scorers agree to assess in the following three aspects:

- **Phoneme (P)**. Phoneme indicates how close the pronunciation is to standard American English. The score ranges from 1 (worst) to 5 (best). For example, strong Japanese accent would lead to a low score.
- **Prosody (S)**. Prosody indicates how well the supra-segmental factors are controlled. This includes the intonation change and the stress position. The score ranges from 1 (worst) to 5 (best).
- **Correctness (C)**. Correctness indicates how well the shadowing content is reconstructed. For example, omissions and insertions of phonemes/words would lead to a low score. The score ranges from 1 (worst) to 5 (best).

These scoring aspects very similar to the introduction in Chapter 3, except that the delay aspect is ignored. By summing up these three scores, the total score of a phrase ranges from 3 to 15.

Fig. 4.1 A screenshot for the shadowing recording website.

4.2.2 Analysis

We analyzed the manual scores and found some interesting results.

Statistics information

Table 4.3 summarizes the score results for all three scorers (SA, SB and SC). Mean is average and SD is standard deviation. SA seems to score more strictly on phoneme and more loosely on prosody and correctness than the other two scorers. SB and SC have very close result on all three aspects. Phoneme mean score is relatively low, which is consistent with the fact that most of the participants are low and medium English learners. Surprisingly, all three scorers give high scores for correctness aspect. This is probably because the utterances are produced after practicing three times, so the leaners have enough time to remember the contents.

Correlations within scorers

We also had a look into the correlations between each pair of three scorers. Table 4.4, Table 4.5 and Table 4.6 are the correlation coefficients (CC) in phrase level, sentence level and speaker level, respectively. P is phoneme, S is prosody and C is correctness score. P+S+C is the sum of them. Phrase level means the cor. is computed using phrase as the basic unit (i.e., the amount of data is $125 * 27 = 3375$). Sentence level means we first merge the phrase scores into sentence scores by their average, and then compute the correlation coefficients. (i.e., the amount of data

Fig. 4.2 A photo of microphone set used in shadowing recording.

Table 4.3 Statistics from three scorers (SA, SB and SC) for each aspect.

	Phoneme (P)	**Prosody (S)**	**Correctness(C)**
SA-Mean	1.9	4.2	4.1
SA-SD	0.61	0.56	0.54
SB-Mean	2.9	2.8	3.7
SB-SD	0.64	0.82	0.68
SC-Mean	2.7	2.9	3.7
SC-SD	0.71	0.78	0.69

is $125 * 10 = 1250$). Speaker level means we further merge the sentence scores into speaker scores by their average, and then compute the cor. (i.e., the amount of data is 125).

As we expected, highest CCs are in the order of speaker, sentence and phrase level. CCs of P+S+C are also higher than the three aspects standalone for all three levels. This is because the more samples are, the reliable the CC is. Also, scorer SB and SC have higher CC than theirs between SA. This is consistent with the fact that statistics (mean and SD) of SB and SC are closer than SA. Another trend is that CC for correctness seems always higher than phoneme and prosody. This can be backed up by the fact that the correctness scores from three scorers are all relatively high.

Correlations with TOEIC scores

We computed CCs between speaker-level manual scores and their TOEIC scores to investigate the reliability of TOEIC tests. Table 4.7 is the CCs between TOEIC scores and manual scores

Table 4.4 Correlation coefficients between each pair of scorers in phrase level.

	P	**S**	**C**	**P+S+C**
SA_SB	0.47	0.42	0.67	0.71
SA_SC	0.43	0.47	0.65	0.69
SB_SC	0.56	0.54	0.70	0.74

Table 4.5 Correlation coefficients between each pair of scorers in sentence level.

	P	**S**	**C**	**P+S+C**
SA_SB	0.57	0.46	0.72	0.73
SA_SC	0.52	0.54	0.72	0.73
SB_SC	0.66	0.62	0.77	0.80

Table 4.6 Correlation coefficients between each pair of scorers in speaker level.

	P	**S**	**C**	**P+S+C**
SA_SB	0.74	0.63	0.85	0.87
SA_SC	0.72	0.73	0.87	0.86
SB_SC	0.84	0.72	0.86	0.87

Table 4.7 Correlation coefficients between manual and TOEIC scores.

	SA	**SB**	**SC**	**SA+SB+SC**
TOEIC	0.44	0.46	0.48	0.48

of each scorer. The manual scores are P+S+C, which is the sum of phoneme, prosody and correctness scores. SA+SB+SC means speaker-level manual scores are the average of these three.

As we expected, TOEIC scores have low correlation with manual scores of any scorer (CC=0.48). This means at least under the current settings of experiment conditions, TOEIC scores are not suitable for representing learners' shadowing proficiencies. This is very important since we have adopted TOEIC scores in many previous works. If we could find one kind of automatic score that has higher CC than TOEIC score with the manual scores (and as we did), this automatic score would be very meaningful for estimating the true proficiencies of shadowing learners.

In next chapters, we will introduce two kinds of improved automatic scores of shadowing speech. Both of them have shown much higher correlation than TOEIC scores in our experiment.

Chapter 5

Proposed approaches

5.1 DNN-based GOP score

5.1.1 Introduction

DNN-based GOP score is an automatic score improved from GMM-based GOP score. Unlike the GMM, DNN allows GOP computation without approximation. That is, $\text{GOP}(p) = \frac{1}{D_p}\log(P(p|O^{(p)}))$ can be directly modeled by DNN model.

To explain how such kind of DNN works in speech recognition and assessment, we first have to how GMM works. At first, only context-independent phones, i.e., the monophones are modeled. For example, the feature distribution of /a/ is represented by a GMM without considering which comes before and after /a/. This is okay when the task is simple, such as singleton phoneme recognition. However, in more complicated tasks such as large vocabulary continuous speech recognition, this model is too simple to handle all variations of each phone. In real life, the physical properties of phone /a/ can be significantly affected by the phones come before and after it. A straight forward approach is to model context-dependent phones, for example, biphones and triphones. Biphone is a phone with another adjacent phone to it, i.e., /b-a/ or /a+c/. Triphone is a phone with its two adjacent phones, i.e., /b-a+c/. /b/ is the phone comes before /a/ and /c/ is the phone comes after /a/.

Triphone model does bring improvement for modeling, but there is a new problem: we do not have enough data for train every triphone well. Suppose we have 40 phones in English, then there will be $40 * 40 * 40 = 64000$ combinations for all triphones, which is too large to train each them. Some of them have not appeared in the train corpus at all. So how do we handle them? The most common solution is, tie similar phones together. For example, phone /b-a+c/ and /d-a+c/ are considered very similar since the central and right phones are the same, and the left phones are both stop consonants. A decision tree is used to group triphones based on

such kind of rules. Finally, the tied phones, usually referred as **senones**, are the smallest unit possessing different GMMs. Their number could vary from hundreds to thousands, depending on different tasks.

DNN model takes observed feature vector as its input, and outputs the posterior probability of all senones. That is, $P(s|O)$ for all $s \in S$, where S is the set of all senones and O is the observed feature vector. The structure of this kind of DNN is illustrated in Figure 5.1.

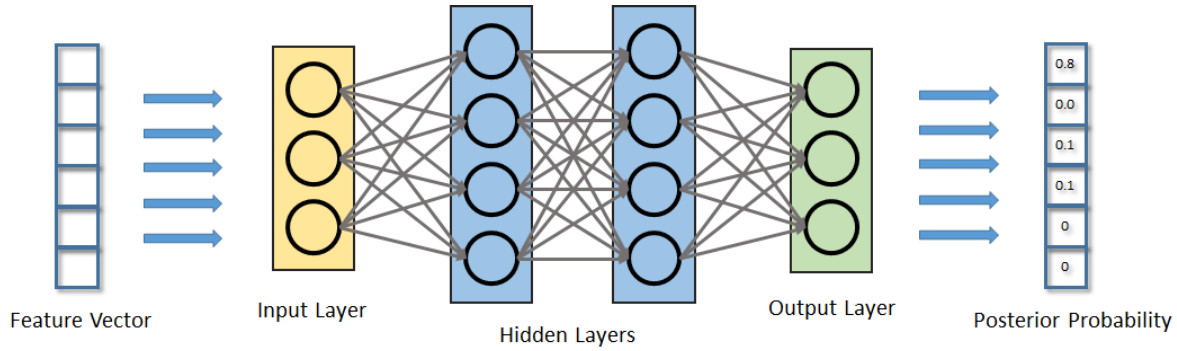


Fig. 5.1 An illustration of DNN structure in GOP computation.

5.1.2 Computation steps

In this section, we will explain how to use DNN model to compute GOP scores in detail. Figure 5.2 is a flowchart of all computation steps.

Feature extraction

The first step is to extract features from the raw waveform. Although MFCCs and its delta features are typical features for modeling, some feature-space transformations are used for a better result. In particular, CMN (Cepstral Mean Normalization), LDA (Linear Discriminant Analysis), MLLT (Maximum Likelihood Linear Transform) and fMLLR (feature-space Maximum Likelihood Linear Regression) are adopted. CMN makes sure that cepstral coefficients are summed up to 0. LDA compresses the dimension of spliced MFCC features. MLLT compensates for the loss caused by using diagonal Gaussian variance matrices for GMM model. fMLLR performs a speaker self adaptation for a better feature estimation. See [26] for more details of these techniques.

GMM-HMM-based forced alignment

The next step is align the utterance to find out phone for each frame. Notice that this is done in context-independent phone level, which means we look for the most possible phone for each

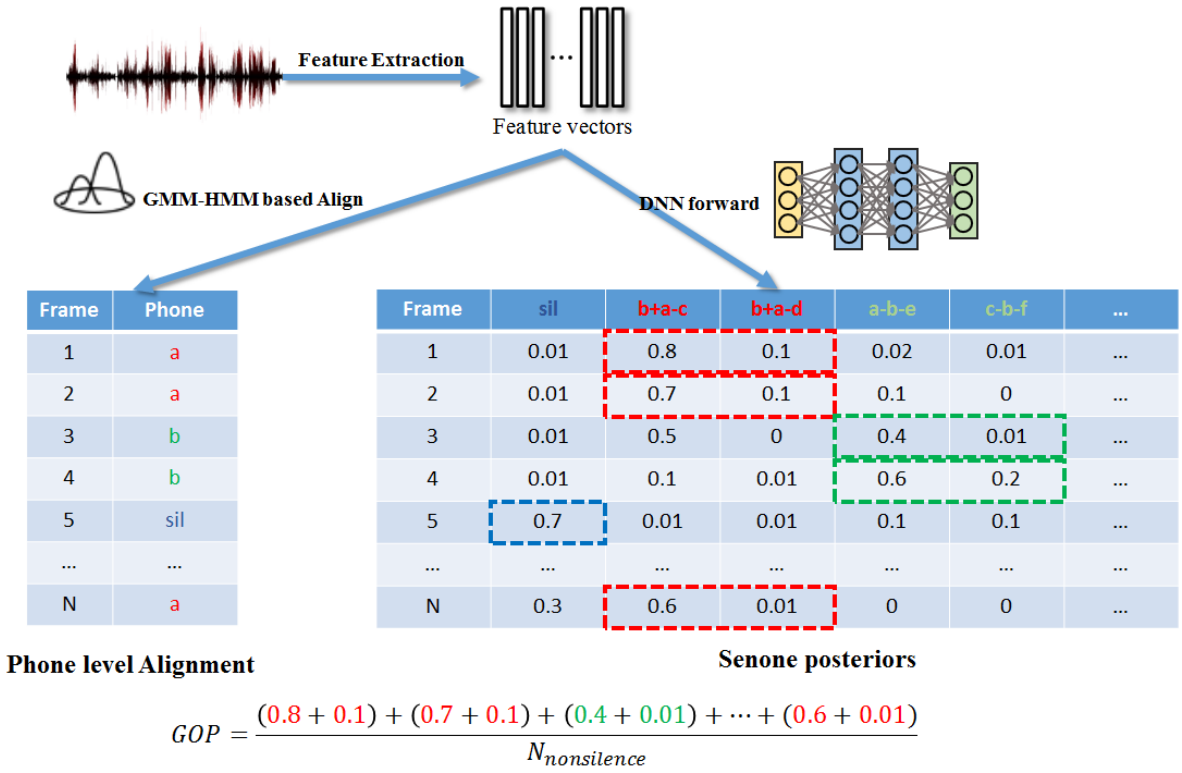


Fig. 5.2 A flowchart of DNN-based GOP computation steps.

frame instead of senone. This is important because later the posterior calculation will be done in senone level. We must keep the mapping relationship between senone and context-independent phone in mind.

You may have noticed that we do not adopt DNN-HMM model to perform forced alignment. This is because alignment results are considered accurate enough even by GMM models. But still, we leave this as a future work.

DNN-based senone posterior

Then we compute the posterior vectors for each frame. The output posterior vector x satisfies $\sum_{i=1}^{N} x_i = 1$, where N is the number of senones, and x_i is the posterior probability for senone i. This can be computed simply by feeding feature vectors forward through the network since this is just what DNN models.

GOP calculation

With all information obtained above, now we are ready to calculate the GOP score for the utterance. The alignment result can be regarded as the "correct answer" of each frame, while the

posterior vector can be regarded as the "actual product" of each frame. As shown in Figure 5.2, by the definition of GOP score (Equation 3.1), we simply sum up the posterior probabilities of senones corresponding to the phone of current frame in alignment. Because we do not want to estimate the goodness of the speaker's silence part, when current phone is silence, we ignore it. Finally, the summed up probability is divided by the number of non-silence frames, and we get the GOP score for current utterance.

Someone may ask what if the speaker missed some words. Would not the alignment result become a mess? The answer is, even if some expected words are missing in the shadowing utterance, the forced alignment still tries to align the missing words to some unrelated parts of the utterance. This will probably result in a low GOP score, which is still consistent with what we expect.

At the point, the DNN-based GOP score is complete. The result of this approach will be introduced in next chapters. Now we will introduce another approach of scoring shadowing utterances in the next section.

5.2 DTW distance of DNN-based posteriors

5.2.1 Introduction

If we look into GOP score deeply, we will find that its essence is to estimate how far the actual pronunciation is from the standard phone model. In other words, we measure the distance of the utterance and the standard acoustic model. However, this approach does not make use of the model utterance (the utterance played for shadowers). The distance between the model and learner's utterances can also be one kind of scores. The problem is, how do we define the distance between two utterances? Trivial definitions such as the difference of waveform amplitude or MFCCs suffer from the speaker-dependent property affection and non-aligned time position. The first problem can be solved using DNN-based senone posteriors, which speaker identity is considered to be removed. The second one can be solved using a general aligning technique called DTW.

DTW is a technique trying to find the "closest" mapping path of two sequences. Here, "closest" means the minimum accumulation of local distances between each element of two sequences. The element can be anything, for example, scalar, vector and matrix, as long as the distance of two elements can be defined. Several works compared two utterances through DTW after they were converted to sequences of posterior vectors [25, 29, 14, 15]. For example, [25] adopted DTW to evaluate the accentedness of non-native speech using senone posteriors. [14]

adopted DTW to detect mispronunciation. And of course, the same technique could be applied to shadowing.

5.2.2 Mathematical formulation

First let us have a look at the general definition of DTW.

Suppose there are two sequences $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$, where n and m are the lengths of X and Y, respectively. Suppose a local distance $\mathrm{Dis}(a,b) \in \mathbb{R}$ is well defined for all $a \in X$ and $b \in Y$. Then the optimization target (minimum DTW distance of X and Y) is:

$$\mathrm{MinD} = \min_P \sum_{i=1}^{K} \mathrm{Dis}(X_{p_i^1}, Y_{p_i^2}) \tag{5.1}$$

$$\text{Where} \quad P = \{(p_1^1, p_1^2), \ldots, (p_K^1, p_K^2)\}, \tag{5.2}$$

$$(p_1^1, p_1^2) = (1,1), (p_K^1, p_K^2) = (n,m) \tag{5.3}$$

P is usually called the DTW path. p_i^1 and p_i^2 are the i-th point on the path, representing indices in X and Y, respectively. Our target is to find a path that minimizes the accumulation distance under some constraints.

Let us look at Figure 5.3 for a better understanding. Imagine there is a board with size $n * m$. Grid (i,j) of the board represents the local distance $\mathrm{Dis}(X_i, Y_j)$. For convenience, denote $\mathrm{D}(i,j) = \mathrm{Dis}(X_i, Y_j)$. The local distance can be anything, as long as it is meaningful in terms of measuring the distance between elements in two sequences. For example, if X and Y are both sequences of scalars, $\mathrm{Dis}(X_i, Y_j)$ can be defined as $|X_i - Y_j|$. If X and Y are both sequences of vectors with the same dimension, $\mathrm{Dis}(X_i, Y_j)$ cane be defined as the Euclid distance between X_i and Y_j.

Now our job is to find a path from $(1,1)$ to (n,m) which minimizes the sum of all local distances on it. But this problem is very trivial since we can put only two grids, $(1,1)$ and (n,m) in it. So we need to add some constraints for the path. For example, just like in Figure 5.3, one grid can only be located at the right, top or right-top of its last. This kind of constraints is called local constraints. Some commonly used are listed in Figure 5.4. Type (a) is used in the example in Figure 5.3. Now the obtained path actually represents a best mapping in terms of least difference.

Suppose the local constraint is type (a). Define $f(i,j)$ to be the minimum accumulation distance from $(1,1)$ to (i,j), then the transition equation and the border conditions can be

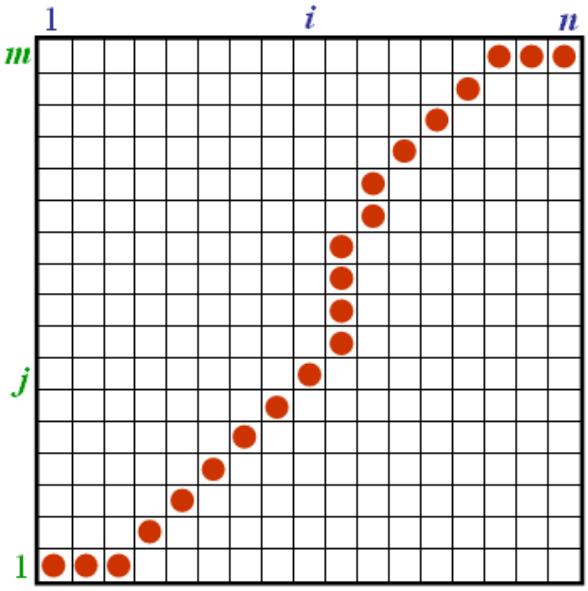


Fig. 5.3 An illustration of dynamic time wrapping.

expressed as:

$$f(i,j) = \begin{cases} \min\left(f(i-1,j-1)+2d, f(i,j-1)+d, f(i-1,j)+d\right), & n \geq i > 1, m \geq j > 1 \\ f(i-1,j)+d, & i > 1, j = 1 \\ f(i,j-1)+d, & i = 1, j > 1 \\ d, & i = 1, j = 1 \end{cases} \tag{5.4}$$

where $D(i,j)$ is denoted as d.

This problem can be easily solved by dynamic programming. $f(n,m)$ will be the minimum distance from $(0,0)$ to $(1,1)$. Based on the choice of min, we could track which local path is actually passed and recover the whole DTW path. Finally, the value of $f(n,m)$ is the DTW path for sequence X and Y under the local distance difinition Dis.

5.2.3 Apply DTW to shadowing speech

When applying DTW to score shadowing speech, we have three things to consider:

1. **Element of sequence**.
2. **Local distance**.
3. **Local constraint**.

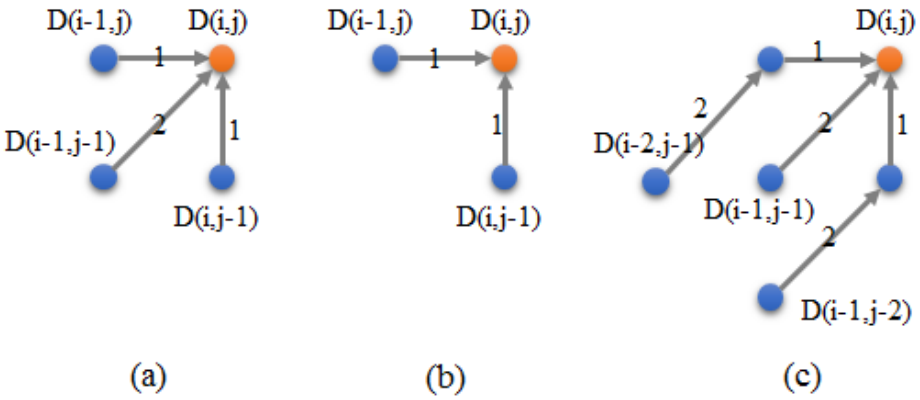


Fig. 5.4 Some commonly used local constraint in DTW.

To eliminate the speaker identity from the utterances, we decide to use senone posteriors as the feature of each frame, i.e., the element of sequence. One another advantage of using senone posteriors is that we can control the granularity. For example, a typical English speech recognition system has about 3000 senones. We can change the amount by modifying the strategy of the decision tree. If we want a very precise comparison, we can change the amount to 5000 or even more. On the other hand, if we want a rough comparison, we can change the amount to 1000 or less.

As for local distance, because the posteriors are vectors, we can use Euclid distance (ED) to measure the difference of two posterior vectors. But since posteriors are also probabilities at the same time (i.e., $\sum_i a_i = 1$), we have some other choices. Bhattacharyya distance (BD) and Kullback–Leibler divergence (KL-divergence or KL-div) are widely used indices of measuring the similarity of two probability distributions. Suppose a and b are two vectors satisfying $a, b \in \mathbb{R}^n$, $\sum_{i=1}^n a_i = 1$ and $\sum_{i=1}^n b_i = 1$, then we can express all three metrics as following:

$$\mathrm{D}_{\mathrm{EUC}}(a,b) = \sqrt{\sum_i (a_i - b_i)^2} \tag{5.5}$$

$$\mathrm{D}_{\mathrm{BD}}(a,b) = -\log\left(\sum_i \sqrt{a_i b_i}\right) \tag{5.6}$$

$$\mathrm{D}_{\mathrm{KLx}}(a,b) = \sum_i a_i \log\left(\frac{a_i}{b_i}\right). \tag{5.7}$$

Note that both $\mathrm{D}_{\mathrm{EUC}}$ and D_{BD} are symmetric for a and b, while D_{KL} is not. However, D_{KL} would not change much if we swap a and b.

The final problem is which local constraint to use. Because the learner's pronunciation can be either longer or shorter than the model utterance, we choose the conventional local constraint to do our experiment, type (a).

At this time point, we are already able to compute DTW distance between model and learners' utterances. This approach does not require to know the transcription of the model utterance. We further extend this approach for a larger application. See the next section.

5.2.4 Language-independent scoring using DTW

Before speaking of how to implement language-independency using DTW, let us begin with the acoustic model. Usually, there are thousands of senones in a acoustic model and such a large number of sound classes are prepared for posterior calculation. These senones can represent much more kinds of sound classes than those defined in phonology. English and Japanese are usually considered very different languages. English has 15 vowels while Japanese has only five. So in English education in Japan, English vowels are often explained by referring to the combination of five vowels. For example, English /ei/ is not found in the Japanese vowel system but it can be approximately produced by intermediate Japanese vowel /e/ and /i/. This means that English /ei/ could probably be found in a Japanese utterance saying /eiei/ because of co-articulation. More generally, the results of co-articulation, senones are capable of representing phonemes partially of another language.

This is gives us a hint, that for the same sound segment, the posterior probability distributions are similar in most languages over the world. Figure 5.5 gives an illustration of this similarity. A correctly produced English vowel /ei/ will most likely have a beautiful distribution in English model. That is, senones with /ei/ as their central phone have high probabilities while others are not. Interestingly, Japanese acoustic model will give a similar distribution for the same English /ei/, except that high probabilities go to senones with adjacent /e/ and /i/ in it.

This property helps with scoring shadowing speech with language independency. We can score English shadowing speeches using Japanese, or any other acoustic model. "What's the point? Doesn't the native English acoustic model give you the most precise result?" Someone may ask. The answer is, yes, the native English model does estimate the English utterances the best, but how about some language you have not even heard about? Some minor languages do not have enough corpus to build a good acoustic model, and even if it had, there may be difficulty to collect them and tune the model. Perform scoring on that language may consume much time and money. However, with this new approach, their shadowing utterances can be scored using English, Japanese or any other well-developed acoustic models, as long as the model utterance is provided (which is probably true since this is shadowing).

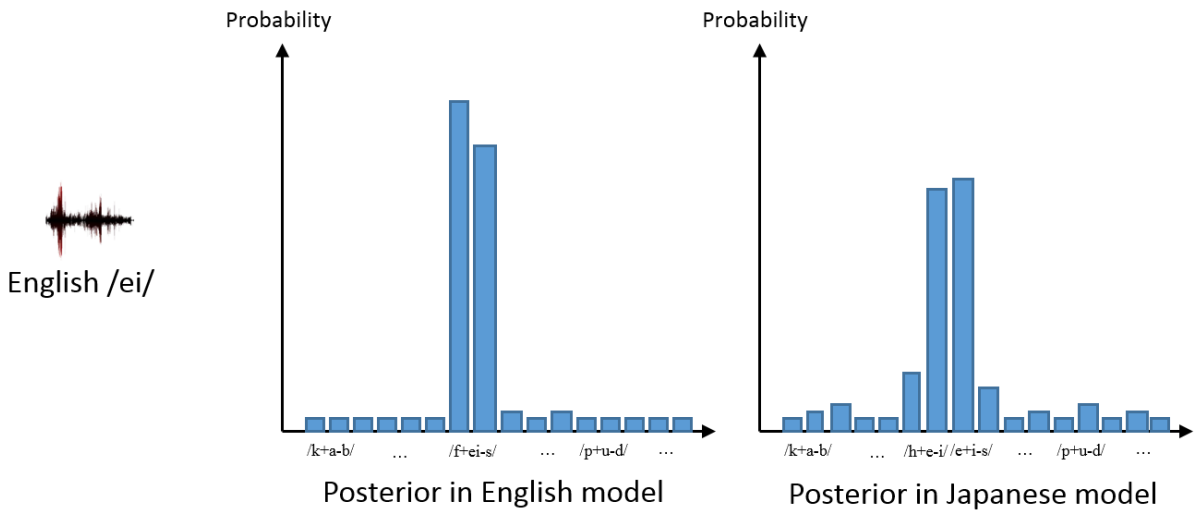


Fig. 5.5 An illustration of posterior distributions in both English and Japanese acoustic models for English /ei/.

In later experiment, we will also compute Japanese posteriors of these English shadowing speeches, and apply DTW to them. The result is expected to be a little worse than the one based on English model.

Chapter 6

Experiment

6.1 Experiment settings

We prepared five acoustic models for the experiment:

- **GMM_HTK**.
  GMM_HTK is a set of English phoneme GMM-HMMs trained with the WSJ (Wall Street Journal) recipe [3] of HTK [2]. The training data includes both WSJ and TIMIT corpus [5]. This model is used in the previous works.

- **GMM_KAL**.
  GMM_KAL is a triphone GMM-HMM model trained with the WSJ recipe of Kaldi Speech Recognition Toolkit [23]. This model is preliminary for the DNN model, since all the alignments for training and feature extraction are done by it. Many normalization and feature-space transformation techniques are adopted in this model.

- **DNN_ENG**.
  DNN_ENG is a English DNN model which outputs senone (tied triphone state) posterior probability, trained with the WSJ recipe of Kaldi Toolkit. The parameters are remained default. There are about 3k senones in the DNN output layer.

- **DNN_JP9K**.
  DNN_JP9K is a Japanese DNN model trained with the CSJ (Corpus of Spontaneous Japanese) recipe of Kaldi Toolkit. The parameters are remained default, which result in about 9k senones for the DNN output layer.

- **DNN_JP3K**.
  DNN_JP3K is a Japanese DNN model trained with the CSJ (Corpus of Spontaneous

Japanese) recipe of Kaldi Toolkit. It is almost identical to DNN_JP9K, except that the leaf number of decision tree is limited, results in about 3k senones in the DNN output layer.

Acoustic model configuration details are listed in Table 6.1. The configurations of DNN models are listed in Table 6.2.

Table 6.1 Details of acoustic model configuration used in experiment.

Model Name	Language	Corpus	#senones	Input feature
GMM_HTK	English	WSJ+TIMIT	8000	MFCC+CMVN
GMM_KAL	English	WSJ	3458	MFCC+CMN+MLLT+fMLLR
DNN_ENG	English	WSJ	3458	MFCC+CMN+MLLT+fMLLR
DNN_JP9K	Japanese	CSJ	9429	MFCC+CMN+MLLT+fMLLR
DNN_JP3K	Japanese	CSJ	2856	MFCC+CMN+MLLT+fMLLR

Table 6.2 DNN configurations used in experiment.

Model Name	DNN Structure	Extra	Train Method
DNN_ENG	Input:440-dim Hidden:2048-dim+Sigmoid Output:3386-dim+softmax	sMBR	Stochastic-GD
DNN_JP9K	Input:1400-dim Hidden:1905-dim+Sigmoid Output:9429-dim+softmax	sMBR	Stochastic-GD
DNN_JP3K	Input:1400-dim Hidden:1905-dim+Sigmoid Output:2856-dim+softmax	sMBR	Stochastic-GD

Basically, GMM_HTK is trained for comparing results with our previous works. GMM_KAL is the model generating alignments and input features for DNN models. DNN_ENG are used to compute DNN-based GOP scores, and all of the DNN models are used to compute DTW distances between learners' and model utterances.

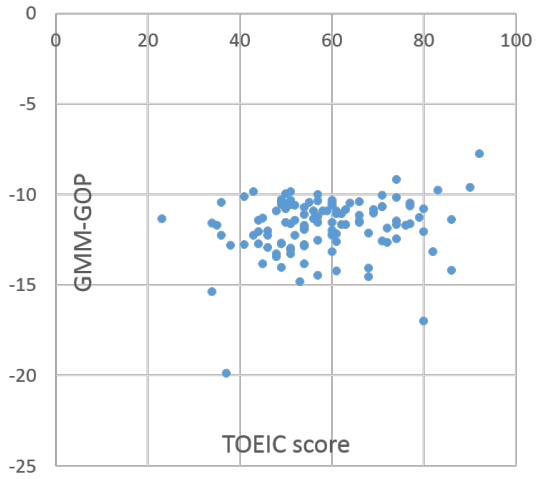
As for the collected corpus, since one learner had problems with his recording devices, finally only 124 learners' data is used.

6.2 Results

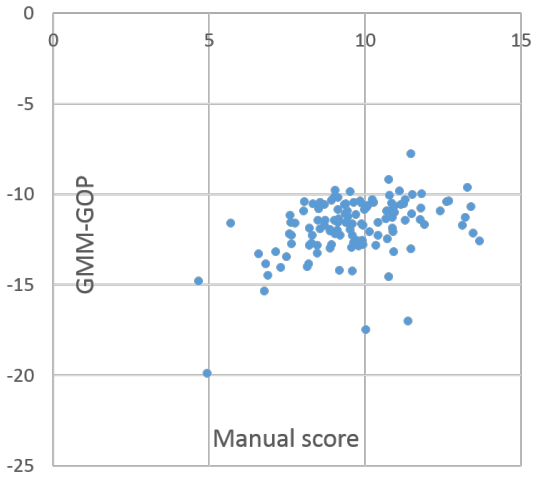
6.2.1 GMM-based GOP scores

GMM-based GOP scores are computed using model GMM_HTK. The score is produced in sentence level, so we have $124 * 10 = 1240$ scores (10 for each learner). These scores are further merged into speaker level, and finally we obtained 124 scores (applied by log).

Figure 6.1a and Figure 6.1b show the plots where the Y-axis is the GMM-based GOP score and the X-axis is learners' TOEIC scores and speaker-level manual scores. Note that the speaker-level manual scores are the sum of three assessing aspects (i.e., P+S+C) and averaged by three scorers.



(a) The plot of GMM-based GOP scores and TOEIC scores. CC=0.19.



(b) The plot of GMM-based GOP scores and manual scores (P+S+C). CC=0.41.

Fig. 6.1 The results for GMM-based GOP scores.

The CCs between TOEIC scores and manual scores are both not high, although the one with manual scores is a little better then TOEIC scores. The reason may be that the feature used for GMM_HTK is very plain without much preprocessing, so it is not robust against noisy speeches. Our corpus is most collected in CALL classrooms and home by students themselves, so some noise is inevitable. This also gives us a hint that this mini-TOEIC test can hardly reflect the true proficiency of shadowing learners.

6.2.2 DNN-based GOP scores

DNN-based GOP scores are computed using model DNN_ENG. The computation details are described in Section 5.1.2.

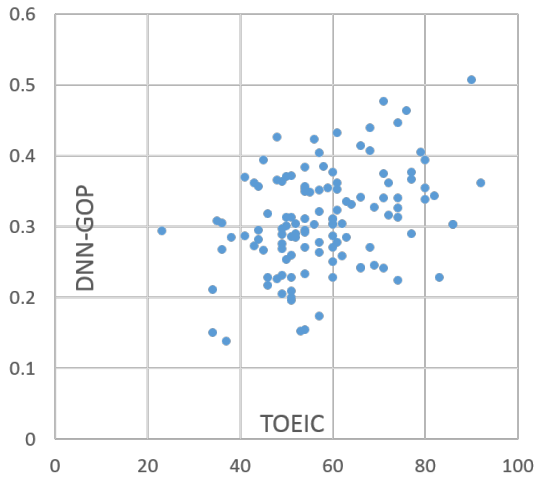
First we take a look at the correlation between speaker-level DNN-based GOP scores and TOEIC scores. The result is in Figure 6.2a. As you can see, the CC is as low as 0.37, indicating a weak correlation with TOEIC scores. On the other hand, the correlation between DNN-GOP and manual scores (which is the sum of three aspects) is plotted in Figure 6.2b. The CC is as high as 0.83, which is close to the inter-rating CC among three scorers (the best is 0.87). Several conclusions can be made so far:

1. TOEIC scores are not suitable for this task. DNN-GOP is a much better indicator in terms of correlation with manual scores (P+S+C) at least by these three scorers.

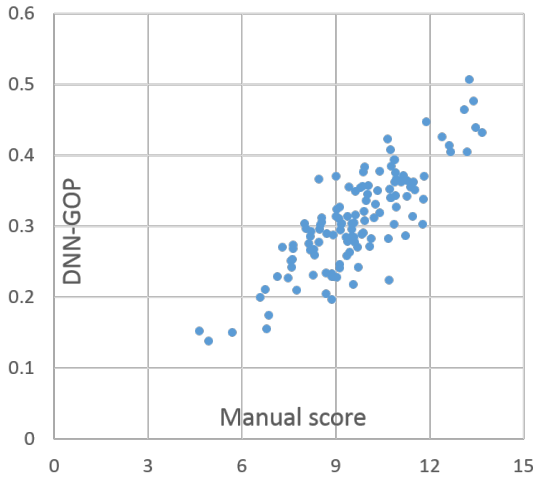
2. DNN-GOP with plenty of preprocessing transformations gives a very good result. The CC is almost as high as those between manual scores from three scorers.

3. GMM-GOP did not work out well in current settings of experiment. The low CC=0.41 is much lower than our expectation.

The reason why GMM-GOP did not work out well may be the noise as previously mentioned. We would like to investigate the performance of GMM and DNN models under the same condition, but we have faced some technical problems when implementing the GMM one. Anyway, we remained this as a future work.



(a) The plot of DNN-based GOP scores and TOEIC scores. CC=0.37.

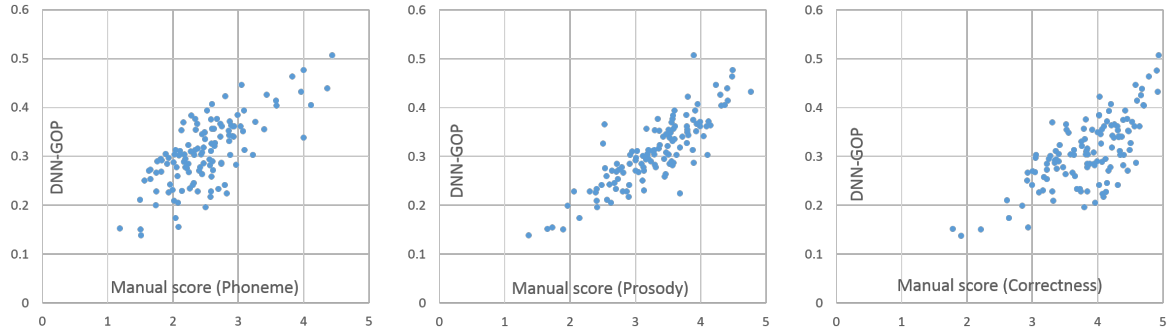


(b) The plot of DNN-based GOP scores and manual scores (P+S+C). CC=0.83.

Fig. 6.2 Plots for DNN-based GOP scores and TOEIC scores and manual scores.

To further investigate the relationships between DNN-GOP and three aspects of the manual scores, we computed CCs for each of the aspect. The results are shown in Figure 6.3.

Figure 6.3a, Figure 6.3b and Figure 6.3c are plotted graphs for phoneme, prosody and correctness aspect, respectively. Among of them, the prosody score has the highest correlation with DNN-GOP (CC=0.84).



(a) The plot of DNN-based GOP scores and manual scores (Phoneme). CC=0.72.

(b) The plot of DNN-based GOP scores and manual scores (Prosody). CC=0.84.

(c) The plot of DNN-based GOP scores and manual scores (Correctness). CC=0.70.

Fig. 6.3 Plots for DNN-based GOP scores and three aspects of manual scores.

On the other hand, although the automatic estimation is usually from automatic scores to manual scores, we performed a linear regression using phoneme, prosody and correctness scores to estimate DNN-GOP in order to find the importance of each aspect. The analysis result is:

$$\text{DNN-GOP} = 0.024P + 0.072S + 0.004C - 0.0048 \tag{6.1}$$

where P represents phoneme, S represents prosody and C represents correctness score.

We can easily tell that about 70% of the weights is on prosody aspect, while nearly none weight goes to correctness aspect. This is consistent with the previous result that prosody score has the highest correlation with DNN-GOP. This is actually not what we expected first because the input feature used in acoustic models are all MFCCs. None of the prosodic features like F0 are fed. The key reason for this may be the importance of stress in English. Stress is categorized in prosody aspect, but represented by the power of spectrum, which is where MFCCs come from. It seems that the three teachers all focused on the stress for prosodic aspect in their scoring.

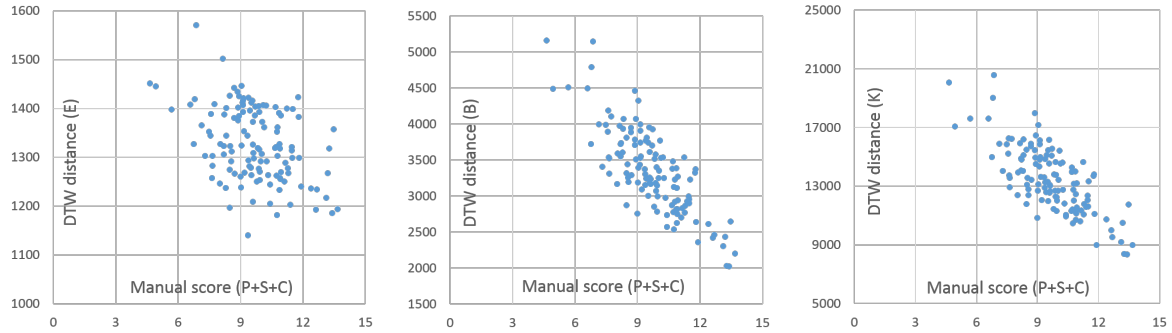
One another thing is that the correctness aspect contributes little to the DNN-GOP. The reason may be that every learner did quite well in their forth recording in terms of understanding the contents, resulting in high correctness scores for most learners. This makes this aspect less discriminative in DNN-GOP estimation.

6.2.3 DTW distance using native acoustic models

The next experiment is to investigate the relationship between DTW distances and manual scores. Here, the DTW distances are computed using English acoustic model DNN_ENG, which means the each dimension of posterior vector represents the probability for a senone in English. Due to the unreliability of TOEIC scores (proved in previous experiments), we no longer do the same computation for the TOEIC scores. Details of the computation process can be found in Section 5.2.

We computed distances using all three kinds of local distances: Euclid distance (E, Equation 5.5), Bhattacharya distance (B, Equation 5.6), Kullback–Leibler divergence (K, Equation 5.7).

The results are shown in Figure 6.4. Note that different from DNN-GOP scores, distances are the shorter, the better. So the CCs between DTW distances and manual scores are all negative. We should focus on their absolute values. BD and KL-div based DTW distances seem to work out very well, especially for BD (CC=−0.80), which has close performance to DNN-GOP (CC=0.84). Again, we have to emphasize that, DTW distance computation does not require any information about the language nor the transcription of shadowed utterances, making it a more difficult task for DNN-GOP. Despite that, close performance is achieved by using Bhattacharya distance. On the other hand, Euclid distance-based DTW distances performed much worse comparing to BD and KL-div. This is probably because Euclid distance is too general to consider the features of probability vectors.



(a) DTW distances (Euclid distance) and manual scores (P+S+C). CC=−0.42.

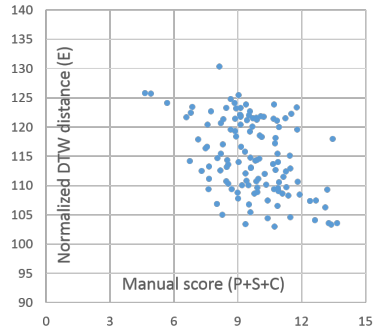
(b) DTW distances (Bhattacharya distance) and manual scores (P+S+C). CC=−0.80.

(c) DTW distances (KL-divergence) and manual scores (P+S+C). CC=−0.76.

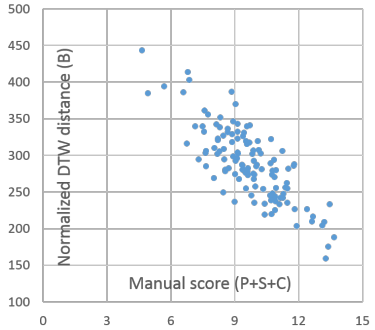
Fig. 6.4 Plots for DTW distances and manual scores.

However, the above experiment forgot to take the duration of model utterances into account. For example, which distance is larger if we have a 10s model utterance and an one-hour model utterance? The answer is, even if the one-hour one is shadowed better, the effect of accumulated

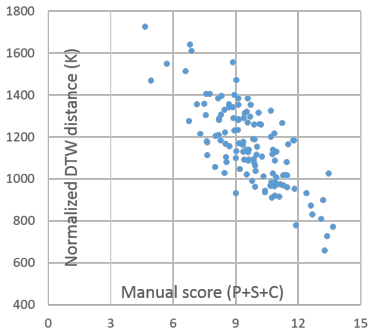
noise and errors would definitely larger than the 10s one due to the duration, which is not what we expected. So we further normalized the DTW distances by the duration of their model utterances. The results are shown in Figure 6.5. There are no significant changes in the CCs (The one for Euclid distance is slightly better). But we will still normalize the DTW distances in later experiments because it is more stable.



(a) Normalized DTW distances (Euclid distance) and manual scores (P+S+C). CC=−0.43.



(b) Normalized DTW distances (Bhattacharya distance) and manual scores (P+S+C). CC=−0.80.



(c) Normalized DTW distances (KL-divergence) and manual scores (P+S+C). CC=−0.76.

Fig. 6.5 Plots for normalized DTW distances and manual scores.

To investigate how these DTW distances are obtained, we print out the DTW paths for three local distances in Figure 6.6. It can be clearly seen that, this is a good speaker and both BD and KL-div tracks the mapping paths well. By contrast, EUC does not track the path well. We can see many corners on its path, which means the two audio segments seem the same to the EUC approach.

6.2.4 DTW distance using non-native acoustic models

Next we repeat the DTW distance computation under the same conditions of previous one. However, we changed the English acoustic model DNN_ENG into Japanese models DNN_JP9K and DNN_JP3K. We believe that both Japanese models will generate similar results to DNN_ENG. Someone may ask why Japanese models, but not French models or Chinese models. This is not because our learners are from Japan. Actually, it is better to use acoustic models from a third language so we can investigate the language independency more clearly. But unfortunately, we have not collected the training or shadowing corpus from other languages, so we trained Japanese models using CSJ instead. Still, we remain this as a future work.

We plotted two graphs of manual scores (P+S+C) and speaker-level DTW distances computed using DNN_JP9K and DNN_JP3K. Since BD performs the best in previous experiments,

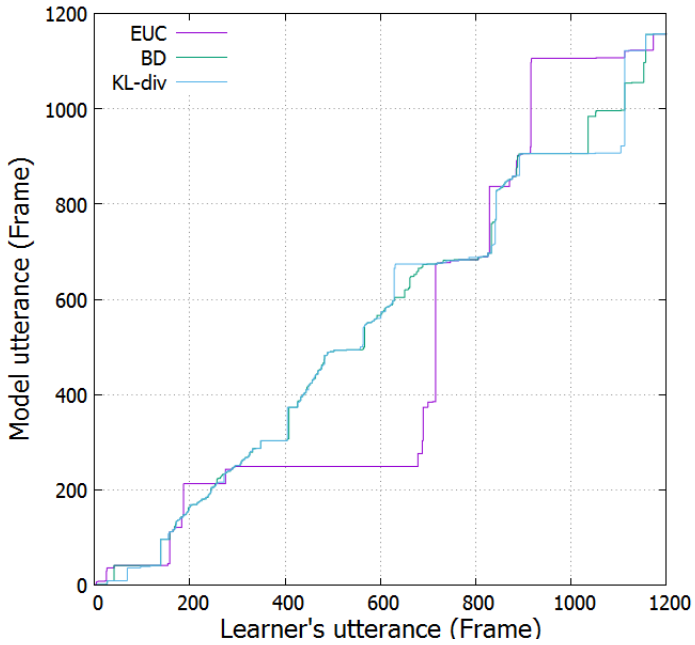


Fig. 6.6 The DTW paths between one learner's and model utterance.

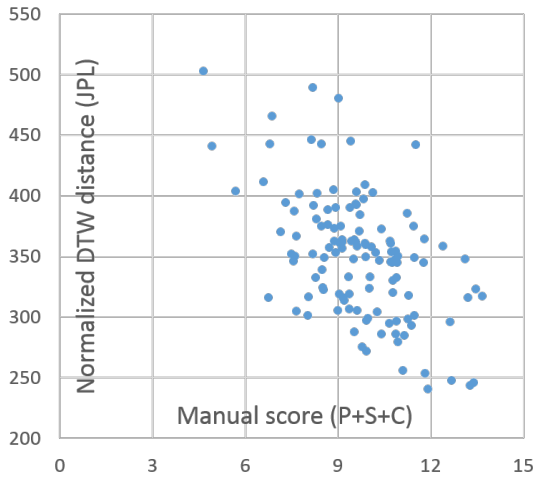
we only adopted BD as the local distance. The distances are normalized by the durations of their model utterances. Note that the only difference between DNN_JP9K and DNN_JP3K is their amounts of senones.

The results are shown in Figure 6.7. Figure 6.7a is the result based on DNN_JP9K and Figure 6.7b is based on DNN_JP3K. The one with fewer senones (about 3k) had much better performance than the larger one (CC=-0.74 against CC=-0.54), which is surprising. There may be two reasons for this result.

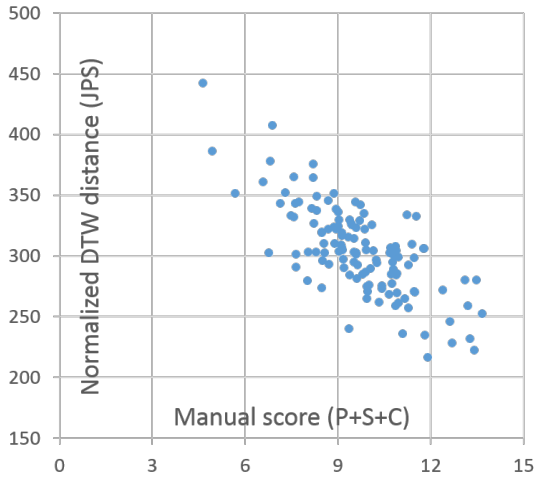
1. The performance of these two models. In our experiment, we ran some Japanese speech recognition tasks to see the WERs (Word Error Rate) of these two Japanese models. DNN_JP3K is slightly better DNN_JP9K, which means the former one can produce posterior probabilities more precisely.

2. The granularity of assessment. Unlike native speakers, second-language learners usually cannot handle the articulation very well. So using a very strict acoustic model may bot be suitable in shadowing assessment. On the other hand, a loose model can take some pronunciation errors in some degree, which is similar to manual scoring strategy.

The most important thing is, DNN_JP9K actually achieved a very close result to the one with native acoustic model (CC=-0.80). This confirms our thoughts about the language independency of this approach. The posterior distributions are still similar even in two very different languages, English and Japanese.



(a) Normalized DTW distances (DNN_JP9K) and manual scores (P+S+C). CC=−0.54.



(b) Normalized DTW distances (DNN_JP3K) and manual scores (P+S+C). CC=−0.74.

Fig. 6.7 The correlation between non-native based DTW distances and manual scores.

Table 6.3 Summarization of all results.

CC	TOEIC	P+S+C	P	S	C
GMM-GOP	0.19	0.41	-	-	-
DNN-GOP	0.37	0.83	0.72	0.84	0.70
DTW-E	-	-0.43(-0.42)	-	-	-
DTW-B	-	-0.80(-0.80)	-	-	-
DTW-K	-	-0.76(-0.76)	-	-	-
DTW-JPL	-	-0.54	-	-	-
DTW-JPS	-	-0.74	-	-	-

We also print out the DTW path based on DNN_JP9K and DNN_JP3K for the same learner's utterance in Figure 6.8. Although the path seems a little irregular at the end, it has nearly the same path shape as the one in Figure 6.6. This confirms the effectiveness of the language independency approach.

6.2.5 Summarization

We summarized all results in Table 6.3. Symbol explanation:

- **CC**. Correlation coefficient.
- **TOEIC**. Speaker-level TOEIC score.

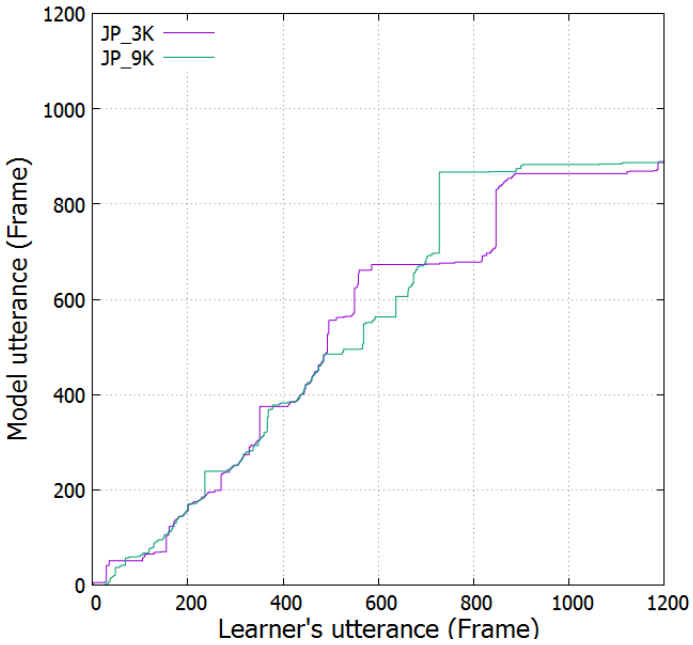


Fig. 6.8 The non-native model based DTW paths between one learner's and model utterance.

- **P**. Speaker-level phoneme aspect score.
- **S**. Speaker-level prosody aspect score.
- **C**. Speaker-level correctness aspect score.
- **P+S+C**. Speaker-level sum score of phoneme, prosody and correctness.
- **GMM-GOP**. Speaker-level GMM-based GOP score.
- **DNN-GOP**. Speaker-level DNN-based GOP score.
- **DTW-E**. Speaker-level DTW distance based on DNN_ENG using Euclid distance as local distance.
- **DTW-B**. Speaker-level DTW distance based on DNN_ENG using Bhattacharya distance as local distance.
- **DTW-B**. Speaker-level DTW distance based on DNN_ENG using KL divergence as local distance.
- **DTW-JPL**. Speaker-level DTW distance based on DNN_JP9K model using Bhattacharya distance as local distance.

- **DTW-JPS**. Speaker-level DTW distance based on DNN_JP3K model using Bhattacharya distance as local distance.

Chapter 7

Conclusions and future works

7.1 Conclusions

In this thesis, we mainly aimed to: 1. improve the accuracy of shadowing speech scoring; 2. find a language and transcription independent approach to perform shadowing speech scoring.

First, we collected English shadowing utterances from 125 college students in Japan. To investigate the reliability of using TOEIC scores as the ground truth of learners' shadowing proficiencies, we manually scored each utterance in three aspects: phoneme, prosody and correctness. The result showed that the correlation coefficient between speaker-level manual scores and their TOEIC scores is as low as 0.48, which means TOEIC scores cannot represent learners' proficiencies well.

Our first approach is using GOP score based on DNN model. A significant improvement of correlation coefficient with manual scores is seen comparing to traditional GMM-HMM based GOP score. The second approach is using DTW distance between learner's and model utterances to assess. The merit is that the transcription of model utterance is no longer needed, and the language of acoustic model could be different from the shadowing language. This approach further allows scoring on expressive shadowing, which is not possible for the GOP approach. The result shows that although the accuracy of DTW approaches using both native and non-native language model drops down a little bit, it is still very close to the GOP approach. Language-independent shadowing scoring is realized to some degree.

7.2 Future works

- Use the same input features for GMM and DNN models. In this work, GMM and DNN models have different input features. They both take MFCCs at first, but the DNN one

applied many feature-space transformations to make the features more robust. We would like to make these conditions same, however it is difficult since they are trained by different toolkits. In the future, we would like to train the models by a single toolkit to make a fair comparison.

- Optimize the number of senones in DNN models. In this work, only one English DNN and two Japanese DNNs with different amount of senones are investigated. The result showed that the performances of these two Japanese DNN models are very different, giving us an idea that there should be a DNN with the best granularity in terms of maximizing the correlation with manual scores. We would like to find it by making the number of senones variable in the future.

- Add an extra regression stage. In this work, the computed automatic scores are directly used as the estimation of learner's shadowing proficiency. It is reasonable to perform another regression by introducing more features, such as word omission rate and silence duration. [27] has adopted this technique to improve the accuracy of GMM-GOP scores. We are planning to do the same things too in the future.

References

[1] Forced alignment. https://www.isip.piconepress.com/projects/speech/software-/tutorials/production/fundamentals/v1.0/section_04/s04_04_p01.html.

[2] Hidden markov model toolkit. http://htk.eng.cam.ac.uk/, .

[3] Htk wall street journal training recipe. https://www.keithv.com/software/htk/, .

[4] Mel frequency cepstral coefficient (mfcc) tutorial. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/.

[5] Timit acoustic-phonetic continuous speech corpus. https://catalog.ldc.upenn.edu/ldc93s1.

[6] Peter W Carey. Verbal retention after shadowing and after listening. *Attention, Perception, & Psychophysics*, 9(1):79–83, 1971.

[7] Mark John Francis Gales. *Model-based techniques for noise robust speech recognition*. PhD thesis, University of Cambridge Cambridge, 1995.

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

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

[10] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[11] 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.

[12] A Kuramoto, O Shiki, H Nishida, and H Ito. Seeking for effective instructions for reading: The impact of shadowing, text-presented shadowing, and reading-aloud tasks. *LET Kansai Chapter Collected Papers*, 11:13–28, 2007.

[13] Sylvie Lambert. Shadowing. *Meta: Journal des traducteurs/Meta: Translators' Journal*, 37(2):263–273, 1992.

[14] Ann Lee and James Glass. A comparison-based approach to mispronunciation detection. In *Spoken Language Technology Workshop (SLT)*, pages 382–387. IEEE, 2012.

[15] Ann Lee and James R Glass. Pronunciation assessment via a comparison-based system. In *SLaTE*, pages 122–126, 2013.

[16] Dean Luo, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose. Automatic assessment of language proficiency through shadowing. In *Chinese Spoken Language Processing, 2008. ISCSLP'08. 6th International Symposium on*, pages 1–4. IEEE, 2008.

[17] Dean Luo, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose. Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences. In *SLaTE*, pages 37–40, 2009.

[18] William Marslen-Wilson. Linguistic structure and speech shadowing at very short latencies. *Nature*, 1973.

[19] William D Marslen-Wilson. Sentence perception as an interactive parallel process. *Science*, 189(4198):226–228, 1975.

[20] William D Marslen-Wilson. Speech shadowing and speech comprehension. *Speech communication*, 4(1-3):55–73, 1985.

[21] Holger Mitterer and Mirjam Ernestus. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, 109(1): 168–173, 2008.

[22] Takayuki Nakanishi and Atsuko Ueda. Extensive reading and the effect of shadowing. *Reading in a Foreign Language*, 23(1):1, 2011.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[24] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[25] Ramya Rasipuram, Milos Cernak, Alexandre Nanchen, et al. Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities. In *INTERSPEECH*, number EPFL-CONF-209089, 2015.

[26] Shakti P Rath, Daniel Povey, Karel Veselỳ, and Jan Cernockỳ. Improved feature processing for deep neural networks. In *INTERSPEECH*, pages 109–113, 2013.

[27] Shuju Shi, Yosuke Kashiwagi, Shohei Toyama, Junwei Yue, Yutaka Yamauchi, Daisuke Saito, and Nobuaki Minematsu. Automatic assessment and error detection of shadowing speech: Case of english spoken by japanese learners. In *INTERSPEECH*, pages 3142–3146, 2016.

[28] Shohei Toyama. Use of global and acoustic features conveying non-linguistic messages to adapt language models for spontaneous speech recognition. Master's thesis, The University of Tokyo, 2017.

[29] Raphael Ullmann, Ramya Rasipuram, Hervé Bourlard, et al. Objective intelligibility assessment of text-to-speech systems through utterance verification. In *INTERSPEECH*, number EPFL-CONF-209096, 2015.

[30] Silke M Witt and Steve J Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2):95–108, 2000.

Appendix A

Publications

International conferences

- Shuju Shi, Yosuke Kashiwagi, Shohei Toyama, Junwei Yue, Yutaka Yamauchi, Daisuke Saito and Nobuaki Minematsu, “Assessment and Error Detection of Shadowing Speech: Case of English Spoken by Japanese Learner”. In *INTERSPEECH*, pp. 3142–3146, 2016.
- 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 *INTERSPEECH*, pp. 1422-1426, 2017.
- Yutaka Yamauchi, Junwei Yue, Kayoko Ito and Nobuaki Minematsu, “Investigation of teacher-selected sentences and machine-suggested sentences in terms of correlation between human ratings and GOP-based machine scores”. In *SLaTE*, 2017.

National conferences and meetings

- Shuju Shi, Junwei Yue, Yosuke Kashiwagi, Shohei Toyama, Yutaka Yamauchi, Daisuke Saito and Nobuaki Minematsu, “Automatic Assessment and Error Detection of Shadowing Speech”. In 情報処理学会音声言語情報処理研究会資料, 2016-SLP-112(7), pp. 1-6 (2016-7).
- Junwei Yue, Fumiya Shiozawa, Shohei Toyama, Yutaka Yamauchi, Kayoko Ito, Daisuke Saito and Nobuaki Minematsu, “DNN-based GOP and Its Application to Automatic Assessment of Shadowing Speeches”. In 情報処理学会音声言語情報処理研究会資料, 2017-SLP-115(13), pp. 1-6 (2017-2).

- Nobuaki Minematsu, Junwei Yue, Yutaka Yamauchi, Kayoko Ito, and Daisuke Saito, “Experimental attempts to realize effective recording of utterances produced by multi-speakers’ simultaneous shadowing”. In 日本音響学会講演論文集, pp. 257-258, 2016.

- Junwei Yue, Shuju Shi, Yosuke Kashiwagi, Shohei Toyama, Yutaka Yamauchi, Daisuke Saito and Nobuaki Minematsu, “A study on improving the performance of automatic assessment of shadowing speech”. In 日本音響学会講演論文集, pp. 145-148, 2016.

- Junwei Yue, Fumiya Shiozawa, Shohei Toyama, Yutaka Yamauchi, Kayoko Ito, Daisuke Saito and Nobuaki Minematsu, “Manual Scoring of Shadowing Speeches and Automatic Score Estimation using DNN-based GOP”. In 日本音響学会講演論文集, pp. 349-352, 2017.