NOISE-ROBUST AND STRESS-FREE VISUALIZATION OF PRONUNCIATION DIVERSITY OF WORLD ENGLISHES USING A LEARNER'S SELF-CENTERED VIEWPOINT

Yuichi SATO, Yosuke KASHIWAGI, Nobuaki MINEMATSU, Daisuke SAITO, Keikichi HIROSE

The University of Tokyo, Tokyo, Japan

{yuichisato,kashiwagi,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

The term of "World Englishes" describes the current and real state of English and one of their main characteristics is a large diversity of pronunciation, called accents. We have developed two techniques of individual-based clustering of the diversity [1, 2] and educationally-effective visualization of the diversity [3]. Accent clustering requires a technique to quantify the accent gap between any speaker pair and visualization requires a technique of stress-free plotting of the speakers. In the above studies, however, we developed and assessed these two techniques independently and in this paper, we assess our technique of automatic accept gap prediction when it is used for our stress-free visualization. Further, since CALL applications today are not always used in a quiet environment, we introduce a feature enhancement (denoising) technique to improve noise-robustness of accent gap prediction. Results show that our accent gap prediction shows correlation of 0.77 to IPA-based manually-defined accent gaps and that, by applying feature enhancement to noisy input utterances, our technique can predict the accent gap that could be obtained in a clean condition, when the SNR is larger than 10 [dB].

Index Terms: World Englishes, pronunciation clustering, visualization, feature enhancement, noise-robustness

1. INTRODUCTION

English is often used as a tool of international communication and this fact inevitably causes a large variation to English, depending on the language background of speakers and listeners. If we focus on the phonological and phonetic aspect, pronunciation diversity is called accents. Recently, more and more teachers accept the concept of World Englishes (WE) [4, 5] and they regard US and UK accents just as two major examples of accented English. If one accepts the concept of WE as it is, he can claim that there does not exist the standard pronunciation of English. In this situation, there will be a great interest in how one type of pronunciation compares to other varieties, not in how that type of pronunciation is incorrect compared to US or UK pronunciation.

These days, we can easily find good online resources of WE such as the TED talk archive [6] and a series of online lectures at many universities [7, 8]. If we can build an accent-based browser of WE, with which spoken documents are searched for by querying the speakers' accent characteristics, it will become a good tool for learners to know the current and real state of English and for international business persons to make themselves accustomed to the pronunciation diversity of WE. We can find several textbooks that introduce the pronunciation diversity of WE for international business persons.

978-1-4673-8279-3/15/$31.00 ©2015 IEEE

For this aim, so far, we have developed two techniques of individual-based clustering of the pronunciation diversity [1, 2] and educationally-effective visualization of the diversity [3]. Generally speaking, clustering of N items requires the distance matrix among the N items. The first technique was developed to predict the accent gap between any speaker pair to obtain the pronunciation distance matrix among speakers [1, 2]. Visualization of a distance matrix is often done by using MDS (Multi-Dimensional Scaling) or drawing its dendrogram. In either case, a result of visualization includes stress or distortion. This is inevitable because the N items often lie in a high dimensional space and visualization is a process of projecting the N items' geometrical distribution in a high dimensional space onto a two dimensional plane. For example, if learners are scattered via MDS, some parts of the resulting chart have stress but learners cannot know which parts of the chart include stress. Pedagogically speaking, this is a serious problem. Then, we proposed a technique to realize stress-free visualization of the distance matrix by introducing a learner's self-centered viewpoint [3].

In these two previous works, however, the two techniques were developed and assessed independently. In this paper, we firstly assess our automatic prediction of the accent gap between two speakers in the context of our stress-free visualization of learners. Secondly, we introduce a feature enhancement (denoising) technique to improve noise-robustness of accent gap prediction. These days, CALL applications are used not only in private and quiet rooms but also in public rooms such as classrooms. In these environments, some noises are inevitably added to speech input to CALL systems. Practically speaking, noise-robustness is required [9] as it is required for speech recognition systems.

The rest of this paper is structured as follows. Section 2 describes our previous works [10, 1, 2, 3] on IPA-based accent gap quantification, individual-based clustering of WE, and educationally-effective visualization of WE. In Section 3, we set up an experimental environment to assess our accent gap prediction for our stress-free visualization and some results are shown. In Section 4, after brief explanation of the feature enhancement technique that we apply here, its effectiveness will be validated. Section 5 concludes this paper with some future directions.

2. RELATED WORKS

2.1. Problem formulation of accent gap prediction

In our previous works, the problem of accent gap prediction between two speakers was formulated as regression problem to predict the reference accent gap of the two speakers automatically only by using their utterances. Here, the reference gap was obtained by comparing IPA transcripts of the two speakers' utterances of the same and common paragraph. For this task, the Speech Accent Archive (SAA) [11]

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

---

pliːˌ kʰəl stɛla æsk əɹ tu bɹĩŋ ðiːs θĩŋz wɪθ həɹ fɹɪ̃m ɡə stɔːɹ sɪks spũːnz əˌ fɪɛʃ snoʊ piːz faːɪv θɪk slæbˌz əv blu ʧʃiːz ɛ̃nd̬ meɪbi ə snaːk fəɹ hɚ bɹʌðəɹ bɔːb wi ɔlso nid ə smɔl plæstɪk sneɪk ɛ̃n ə bɪg tʰɔɪ fɹɔːg fɚ ðə kʰɪːdˌz ʃi kɛ̃n skup ðiˌz θĩŋz ĩntũ θɹi ɹɛd bæːgz ɛ̃n wi wɪ goʊ miːt həɹ wɛ̃nzdeɪ æt ðə tɹeĩn steɪʃə̃n

Fig. 1. The elicitation paragraph of 69 words and an example of IPA narrow transcription (speakerID = german17)

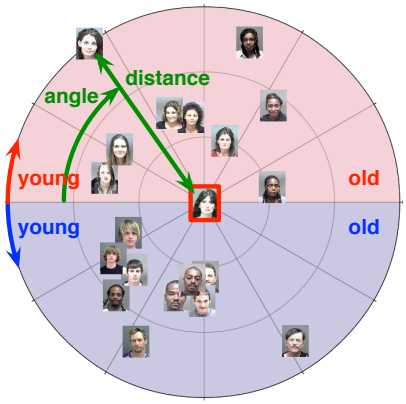
was used because it is a collection of readings of the same paragraph, provided by approximately 2,000 native and non-native speakers of English all over the world. Figure 1 shows the common paragraph used for recording and an example of IPA narrow transcription.

2.2. IPA-based accent gap quantification

By comparing two IPA transcripts of two speakers, it is possible to quantify the accent gap between them [12, 13], where good correlation was observed between the quantified gaps and subjectively-defined gaps although prosodic characteristics are not described well on IPA transcripts. In [10], we realized another quantification method by using Dynamic Time Warping (DTW) between two transcripts. For this, the distance matrix among all the kinds of IPA phones used in the SAA had to be prepared. We built an HMM for each of the most frequent 153 phones, not phonemes, in the SAA, which can cover 95% of the phone instances found in the SAA. For each of the other 5% phones, we substituted the HMM with no diacritic that shares the same base phone. These HMMs were trained using an expert phonetician's twenty productions of each of the 153 phones and were used to prepare the distance matrix among the phones in the SAA. In [14], our DTW-based method of accent gap quantification was compared to some conventional methods using other strategies [12, 13]. Our method showed better correlation to the accent gap subjectively rated by human listeners.

2.3. Automatic prediction of accent gaps between speakers

The IPA-based accent gap calculated via DTW was automatically predicted without IPA transcripts [1, 2]. As mentioned in Section 2.1, this problem was treated as regression problem. What kind of features should be used for accent gap prediction? It should be noted that acoustic differences between the SAA utterances of two speakers are not good features for prediction [15]. This is because acoustic differences are strongly influenced by non-linguistic factors such as differences of age and gender, which are totally irrelevant to accent gap prediction. To avoid the non-linguistic influences, in [1, 2], we used pronunciation structure analysis, which was proposed in [15]. Generally speaking, non-linguistic differences, such as differences in speaker and microphone, can be modeled mathematically as static feature transformation such as frequency warping. Pronunciation structure analysis characterizes an utterance only by contrastive features, which are mathematically proven to be independent of any invertible static feature transformation.



The pronunciation of a speaker in a red rectangle is compared to those of some speakers in the SAA. She is placed at the origin and the accent gap from her to a speaker in the archive is represented as distance between them. The angle of each archive speaker indicates his/her age. The archive speakers of the same gender are plotted in the upper semicircle and vice versa.

Fig. 2. Visualization of pronunciation diversity from a speaker's self-centered viewpoint [3]

To predict the IPA-based accent gap, *differential* contrastive features between two speakers were used for Support Vector Regression [1, 2]. The performance was evaluated in two modes of speaker-pair-open and speaker-open. In the former mode, the correlation of IPA-based gaps to automatically predicted gaps was much higher than that of IPA-based gaps to phoneme-based gaps, which were calculated via DTW of phone**mi**c transcripts, not phone**ti**c transcripts. Phone**mi**c transcripts are broader and more abstract description than phone**ti**c transcripts. Experimental procedures in [1, 2] will be explained in Section 3.

2.4. Educationally-effective visualization of the diversity

By using IPA-based gaps or automatically predicted gaps between any speaker pair out of N speakers, we can obtain the accent gap matrix or the pronunciation distance matrix among the N speakers. Two well-known methods to visualize a distance matrix are drawing an MDS-based scatter chart and a dendrogram from the matrix. Both methods try to project the geometrical distribution of the N speakers in the original high dimensional space onto a two-dimensional plane. If those methods are used for learners in a language class and the result is fed back to them, they will receive one and the same visualization result. It is expected, however, that different learners may pay special attention to different parts of the result. A learner's main interest will be in the relations from *himself* to others, which should be emphasized for visualization, compared to the other relations. Learner-dependent visualization will be practically preferable.

A problem exists both in the above two methods. Projection of a geometrical shape in a high dimensional space onto a two-dimensional plane usually causes distortion or stress. This stress can be avoided for a learner in the N speakers by visualizing only a part of the distance matrix, which should be related to that specific learner. In other words, stress is inevitable when one attempts to visualize the entire matrix of the N speakers. Suppose that that specific learner is speaker n, $\{d_{nj}\}$ in the matrix are relations from that learner to others and $\{d_{nj}\}$ can be visualized even on a one-dimensional plane with no stress.

In [3], for speaker n, we used $\{d_{nj}\}$ and other non-linguistic attributes of the N speakers for effective visualization. Figure 2 shows

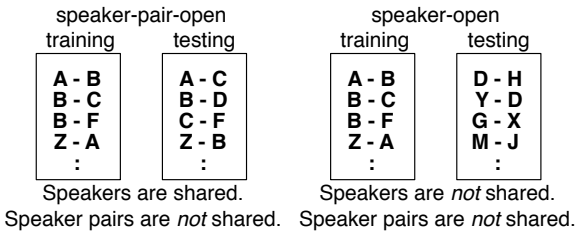


Fig. 3. Two modes of speaker-pair-open and speaker-open

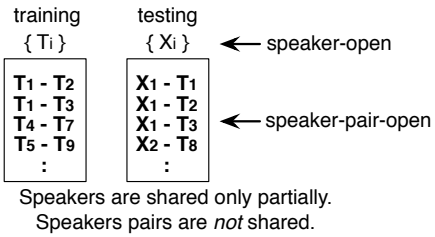


Fig. 4. A new mode of accent gap prediction

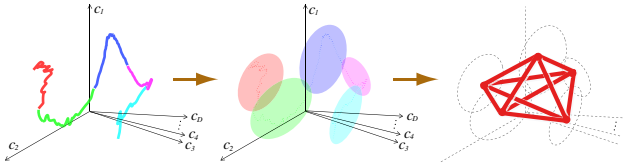


Fig. 5. Pronunciation structure extraction

Table 1. Three kinds of correlation for accent gap prediction

speaker-pair-open	speaker-open	phoneme-based
0.87	0.50	0.76

an example. Here, the age and gender are also referred to. We have a strong reason why we adopted the non-linguistic factors of age and gender in Figure 2. It is interesting that a learner's listening ability is sometimes overfitted to a specific speaker, i.e., his teacher. A learner can understand easily what his teacher says but understand poorly what other teachers say. Learners' robustness of listening against differences of age and gender is known to be lower than that of native speakers [16, 17]. Considering this fact, we introduced age and gender attributes to visualization.

3. ACCENT GAP PREDICTION FOR STRESS-FREE VISUALIZATION

As described in Section 1, our three previous works were conducted independently and, especially, technical assessment of our method of accent gap prediction and that of stress-free visualization was done separately. In this section, the former technique is assessed when it is combined with the latter one.

3.1. Three modes of automatic accent gap prediction

In [2], automatic accent gap prediction between two speakers was examined in two modes, which are a speaker-pair-open mode and a speaker-open mode. Difference between the two is illustrated in Figure 3. The task of accent gap prediction takes two speakers as input and predicts the accent gap between them. So, in the former mode, training speaker *pairs* and testing speaker *pairs* are not overlapped at all. However, training speakers are allowed to be found in testing speaker *pairs* and testing speakers can be found in training speaker *pairs*. Openness is guaranteed only in terms of speaker *pairs*. On the other hand, in a speaker-open mode, all the available speakers are divided into training speakers and testing speakers, and training speaker pairs are formed only from the training speakers. As for testing speaker pairs, only the testing speakers are used. In this mode, openness is guaranteed in terms of speakers.

When we can use N speakers for experiments, the number of speaker pairs is $N(N-1)/2$. If we divide these pairs into two halves for training and testing, the number of training speaker pairs is $N(N-1)/4$ in a speaker-pair-open mode. In the other mode, since the number of training speakers is $N/2$, that of training speaker pairs is $N(N-2)/8$, which is smaller than the half amount of training data in a speaker-pair-open mode.

Another large difference exists between the two modes, which is related to the regression mechanism of Support Vector Regression (SVR). In a speaker-pair-open mode, when speaker pair A-B is found in the testing data, speaker pairs of A-$\{x\}$ ($x \neq B$) and B-$\{y\}$ ($y \neq A$) can be found in the training data. In SVR, input features are mapped into a very high-dimensional feature space, where inner product between an input sample and each of all the training samples is calculated by using a kernel function. Values of inner product can be regarded as similarity scores and regression is done by using these scores as weights. When one wants to predict the accent gap of A-B in a speaker-pair-open mode, the prediction performance is expected to be affected by whether $\{x\}$ include a speaker who is close to B or $\{y\}$ include a speaker who is close to A in the training data.

On the other hand in a speaker-open mode, when A-B is found in the testing data, the training data includes neither of A or B. The prediction performance is easily expected to be influenced by whether or not a speaker pair who are close enough to A-B is found in the training data.

We can claim that the task of accent gap prediction in a speaker-pair-open mode comes to treat speaker-wise pronunciation diversity and that the task in a speaker-open mode has to handle speaker-pair-wise pronunciation diversity. In other words, in the former mode, the magnitude of pronunciation diversity is estimated to be $O(M)$ and it is to be $O(M^2)$ in the latter mode, where M is the magnitude of speaker diversity. Due to these two kinds of increased difficulty, the regression performance in a speaker-open mode is much lower than that in a speaker-pair-open mode. Table 1 shows three kinds of correlation [2], correlation of predicted gaps to IPA-based gaps in the two modes, and that of phoneme-based gaps to IPA-based gaps. The phoneme-based gaps are calculated by conducting DTW over phone**mi**c transcripts of the SAA utterances, which are converted from the original SAA phone**ti**c transcripts.

Stress-free visualization [3] was proposed to locate a new speaker, who is a central speaker in Figure 2, adequately in the archive speakers of WE. In this case, it is reasonable to consider that all the archive speakers have their own IPA transcripts while a new speaker does not. Training of SVR is done by using all the archive speakers and testing is done by predicting the accent gap between that new speaker and each of the archive speakers. Strictly speaking, the two modes investigated in [2] cannot be applied directly to this experimental setup, which is illustrated in Figure 4. In this new mode, a testing speaker is always a new speaker and is not included in the training data, and in this sense, accent gap prediction is done in a speaker-open way. However, accent gap is always predicted between a new speaker and a known archive speaker, who is used

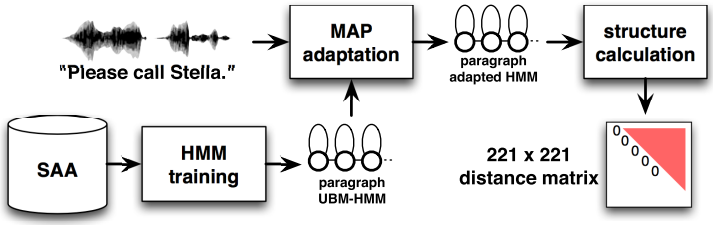


Fig. 6. Procedure to calculate the pronunciation structure

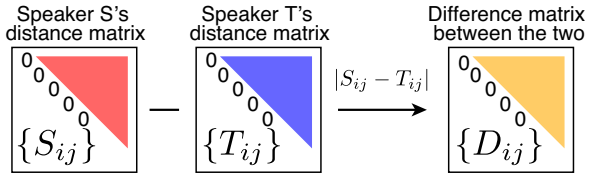


Fig. 7. Difference matrix between two speakers' matrices

as training speaker of SVR. In this sense, accent gap prediction is done only in a speaker-pair-open way. Namely, the mode adopted in [3] can be placed between the two modes examined in [2] and the performance in the new mode will be somewhere between their performances.

In the following sections, the regression performance in the new mode is experimentally investigated.

3.2. Pronunciation structure analysis

Figure 5 conceptually illustrates the process of pronunciation structure extraction from an input utterance [15]. The utterance, which is a feature vector sequence, is converted to a sequence of distributions. From every pair of them, f-divergence-based distances are calculated. The resulting distance matrix is called speech structure or pronunciation structure. Due to transform-invariance of f-divergence [18], the structure was shown to be very independent of static and non-linguistic variations [15]. This structural representation was already applied to speech recognition and synthesis [19, 20], pronunciation scoring [21], pronunciation error detection [22], pronunciation clustering [23], and dialect analysis [24]. In this paper, the Bhattacharyya distance is used as one of the f-divergences.

3.3. Procedure of accent gap prediction

The pronunciation structure was extracted from each of spoken paragraphs of the SAA. Here, the paragraph-based speaker-independent HMM was trained firstly and it was used as Universal Background Model (UBM). Then, it was adapted through MAP (Maximum A Posteriori) adaptation to each speaker. The initial model for the UBM-HMM was prepared by concatenating American English (AE) phoneme HMMs trained with the WSJ corpus [25] by referring to the phoneme sequence derived from the CMU pronunciation dictionary [26]. The initial model was updated through ML-based parameter reestimation by using all the 369 available speakers of the SAA[1]. This UBM-HMM was then adapted to each of the 369 speakers. Acoustic features used for paragraph-based HMMs were MFCC + ΔMFCC. Figure 6 schematizes the procedure adopted in this paper to calculate the pronunciation structure. The number of states of a paragraph-based HMM is $3N$, where N is the number of phonemes of the SAA paragraph (=221).

[1]Many speakers in the SAA deleted some words in the SAA paragraph or inserted new words. They were not used in the experiments.

For each speaker-adapted paragraph-based HMM, the averaged Bhattacharyya distance (BD) between every pair of the phoneme instance HMMs was calculated, where the i-th phoneme instance HMM in the paragraph HMM is a three-state HMM spanning from the $(3i-2)$-th state to the $3i$-th state of that paragraph HMM. BD was calculated by using MFCC features only. Finally in Figure 6, the pronunciation structure of a spoken paragraph of the SAA was obtained as 221×221 distance matrix. As illustrated in Figure 7, from two distance matrices of speakers S and T, we can derive a difference matrix D to characterize the accent gap between them.

$$D_{ij} = |S_{ij} - T_{ij}|, \quad (i < j). \tag{1}$$

For SVR, all the elements of $\{D_{ij}\}$ were used as input features and the total number of the features is 24,310 ($=221\times220/2$). The target of prediction is the pronunciation gap calculated by using the IPA transcripts of S and T. ϵ-SVR in LIBSVM [27] was used with the radial basis function kernel of $K(x_1, x_2) = \exp(-\gamma|x_1 - x_2|^2)$.

3.4. Results and discussion

Using all the available speakers of the SAA corpus, 5-fold cross-validation experiments were done in the new mode explained in Section 3.1. Here, for each of the testing speakers, the accent gaps between him/her and the training known speakers were predicted. Using these gaps and their IPA-based reference gaps, the correlation for that testing speaker was calculated. By conducting cross validation, the averaged correlation of the predicted gaps to IPA-based gaps over all the testing speakers was obtained as 0.77. This is surely lower than the performance in a speaker-pair-open mode but still very comparable to that of phoneme-based accent gap calculation. Although the practical usefulness of the proposed technique for learning WE is not discussed here, the performance obtained experimentally may be able to be interpreted in the following way.

Phonemes are often explained as the minimum linguistic units that ordinary listeners can perceive, and they are defined dependently on the native language of those listeners. Similarly, phones with diacritics are said to be the minimum linguistic units that expert phoneticians can perceive and they are independent of languages that are spoken. Phoneme-based accent gap calculation was done via DTW between American English (AE) phonemic transcripts that were converted from the SAA IPA phonetic transcripts. Logically speaking, we can claim that AE phonemic transcripts can be regarded as results of ordinary American listeners' perception while IPA phonetic transcripts are surely results of expert phoneticians' perception. Since the correlation of phoneme-based gaps to IPA-based gaps and that of automatically predicted gaps to IPA-based gaps is very comparable, our proposed method of predicting the accent gap for stress-free visualization may be comparable to ordinary AE listeners' performance of prediction. Further, we can say that the performance shall be improved by using additional features already examined in [2].

4. NOISE-ROBUST PREDICTION OF ACCENT GAPS

In the current section, we aim at improving noise-robustness of accent gap prediction by introducing a technique of feature enhancement or noise suppression. Here, as Deep Neural Network (DNN)-based feature enhancement, we examine Deep Denoising Auto-Encoder (DDAE), originally proposed for noise-robust speech recognition [28].

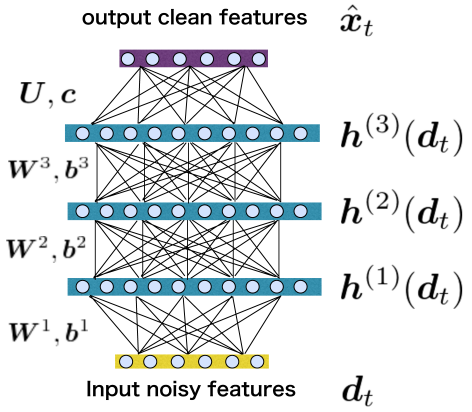


Fig. 8. Deep denoising auto-encoder tested in this paper

4.1. Deep denoising auto-encoder

DDAE uses DNN as feature transformer and attempts to reconstruct clean features directly from their noisy version. Figure 8 shows the DDAE tested in this paper. It has three hidden layers and the number of nodes of each hidden layer was fixed to 1024. This DDAE can estimate clean features as

$$\hat{\boldsymbol{x}}_t = \boldsymbol{U}\boldsymbol{h}^{(3)}(\boldsymbol{d}_t) + \boldsymbol{c}\,, \quad (2)$$

$$\boldsymbol{h}^{(3)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(3)}\boldsymbol{h}^{(2)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(3)})\,, \quad (3)$$

$$\boldsymbol{h}^{(2)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(2)}\boldsymbol{h}^{(1)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(2)})\,, \quad (4)$$

$$\boldsymbol{h}^{(1)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(1)}\boldsymbol{d}_t + \boldsymbol{b}^{(1)})\,, \quad (5)$$

where $\boldsymbol{U}$ and $\boldsymbol{W}^{(n)}$ are weight matrices and $\boldsymbol{c}$ and $\boldsymbol{b}^{(n)}$ are bias vectors. $\boldsymbol{d}_t$ is an input speech feature vector, composed of seven consecutive frames of MFCC+Δ+$\Delta\Delta$ (7×39=273 dimensions). Output feature vector $\hat{\boldsymbol{x}}$ is a 39-dimensional vector. The DDAE was pre-trained with Restricted Boltzmann Machine (RBM) for each layer and finally fine-tuned by applying back-propagation based on the minimum mean square error criterion.

4.2. Accent gap prediction in noisy environments

For this experiment, noise addition was done to all the utterances of the 369 speakers. Considering practical situations of using CALL applications, we selected two types of noise from the JEIDA-NOISE database [29], computer noise and machine noise. The SNR levels of the resulting noisy utterances were set to 0, 5, 10, 15, and 20 [dB] and these utterances will be depicted as NU-1. The feature-enhanced version of NU-1 through DDAE will be called as EU-1. For accent gap prediction, the UBM was trained only with clean utterances, which was MAP-adapted to every paragraph utterance of NU-1 and EU-1.

For training of DDAE, utterances of 1,016 speakers in the SAA, which are not overlapped with the above 369 speakers, were used with the above two kinds of noise. The SNR used for training DDAE was also from 0 to 20 [dB], meaning no acoustic mismatch with respect to the type and level of noise. It should be noted, however, that a single network of DDAE was used commonly for all the types and levels of noise. For testing DDAE, due to lack of time, a single testing set out of the five sets of 5-fold cross validation was used.

To test DDAE in the case of a new type of noise, we additionally selected another kind of practical noise, babble noise, and added it to the 369 speakers' utterances at the SNR levels of 0 to 20 [dB], which are referred to as NU-2 henceforth. Their feature-enhanced version through DDAE is EU-2, where DDAE was trained only with computer noise and machine noise.

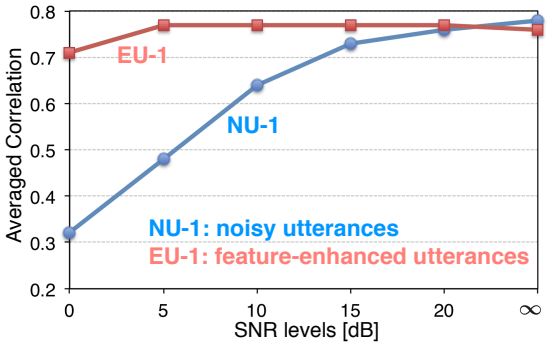


Fig. 9. Effects of DDAE in closed-noise environments

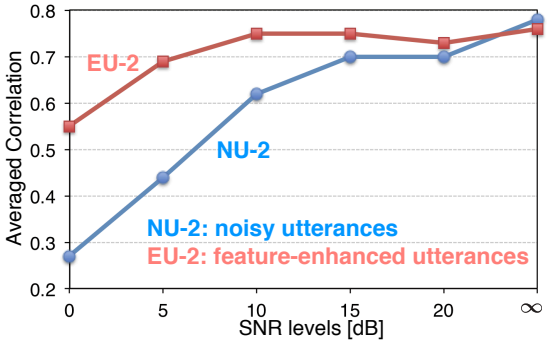


Fig. 10. Effects of DDAE in open-noise environments

4.3. Results and discussion

Figure 9 and Figure 10 show the correlations in closed-noise environments (computer noise and machine noise) and those in open-noise environments (babble noise), respectively. Different from speech recognition applications, CALL applications are expected *not* to be used in environments with unexpected and heavy noise such as cars, trains, airplanes, streets, restaurants, etc. It is also highly expected that users will remove or turn off noise sources such as radios before using CALL applications. Therefore, performance assessment of DDAE in closed-noise environments can be said to be still practical. Clearly shown in Figure 9, DDAE can improve the correlation very effectively. If the SNR of input speech is larger than 5 [dB], the accent gap that could be obtained in a clean condition can be predicted with DDAE. The correlation at the SNR being 5 [dB] is 0.77 while that in a clean condition is 0.78.

In open-noise environments, where babble noise is used, we can say that DDAE is still very effective. It seems that the SNR of 10 [dB] is required to predict the accent gap of a clean condition. The correlation at the SNR being 10 [dB] is 0.75 while that in a clean condition is 0.78.

5. CONCLUSIONS

In many classes of English, utterances of a single accent are often accepted as model utterances. In Japan, General American (GA) is often used and in Europe, Received Pronunciation (RP) is widely used. In this situation, learners will regard mistakenly that type of English as *the* English and will expect that other English users will use that accent when they speak to those learners. Once the learners