A MEASURE OF PHONETIC SIMILARITY TO QUANTIFY PRONUNCIATION VARIATION BY USING ASR TECHNOLOGY

Tianze Shi[1,2], Shun Kasahara[2], Teeraphon Pongkittiphan[2]

Nobuaki Minematsu[2], Daisuke Saito[2], Keikichi Hirose[2]

[1] Tsinghua University, [2] The University of Tokyo

{shitianze, kasahara, teeraphon, mine, dsk_saito, hirose}@gavo.t.u-tokyo.ac.jp

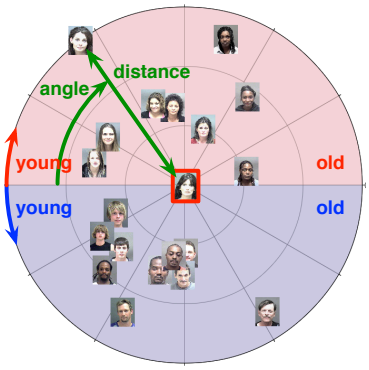
ABSTRACT

It attracts researchers' interest how to define a quantitative measure of phonetic similarity between IPA transcripts of the same sentence read by two speakers. This problem can be divided into how to align two transcripts and how to quantify alignment gap. In this paper, we introduce a method of similarity calculation using phone-based or phoneme-based acoustic models trained with the algorithm used to develop Automatic Speech Recognition (ASR) systems. Use of acoustic models will introduce an issue of speaker dependency because speech spectrums always convey the information of the training speakers' age and gender, which is totally irrelevant to phonetic similarity calculation. We examine how independent our method is of training speakers and how close the calculated similarity is to the similarity subjectively rated through a listening test. We also compare our method to recent works and show our method can give higher correlation by 4 points to human-rated similarity.

Keywords: Phonetic similarity, DTW, HMM, speaker-independency, native-likeness

1. INTRODUCTION

If one can use a good method of measuring phonetic similarity between two IPA transcripts of the same sentence read by a native speaker and a non-native speaker, the similarity will be used as degree of nativeness [14]. If that method is applied to two differently accented non-native speakers, the similarity will be used as quantitative measure of accent gap. If it is applied to N native and non-native speakers, one can get a similarity matrix or a distance matrix of pronunciation diversity found in the N speakers. Using this matrix, one can visualize the diversity of pronunciation. In our previous study [4], we used a method of calculating phonetic similarity between transcripts to obtain a pronunciation distance matrix among speakers of World Englishes, where a main focus was put not on validity of phonetic similarity calculated by our method but on effectiveness of visualizing the pronunciation diversity of World Englishes. Figure 1 shows an example of visualization.



The pronunciation of a speaker in a red rectangle is compared to those of many speakers in an archive of World Englishes. 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.

Figure 1: Visualization of pronunciation diversity from a speaker's self-centered viewpoint [4]

In this paper, we examine how valid our method of similarity calculation is by comparing its results to subjectively rated similarity. Since we use acoustic models trained by using utterances of a single speaker, speaker dependency will be an inevitable problem. If speaker dependency is strong, similarity scores calculated by acoustic models of a speaker and those by another speaker will have different values. In the current paper, empirical results demonstrate high speaker-independency of our model. Further, our method gives higher correlation to human-rated similarity than recently proposed methods.

2. ALIGNMENT OF TRANSCRIPTS

Phonetic similarity between transcripts is often derived by aligning two transcripts based on edit distance such as Levenshtein distance [3, 14]. Here, a phonetically valid definition of segment-to-segment

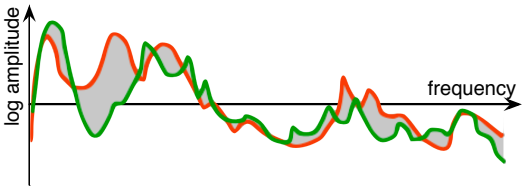
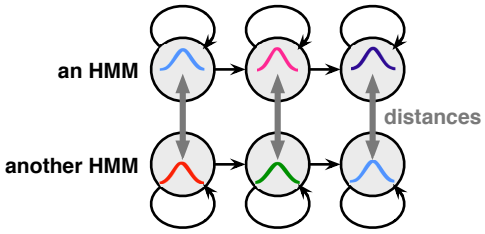


Figure 2: Spectral gap between two spectrums



Black arrows and gray arrows represent allowed transitions and state-to-state distances, respectively.

Figure 3: 3 state-level distances bet. 2 HMMs

distance is key to the algorithm. Then, theoretically inspired measures based on distinctive features were introduced [1], and data-driven measures were developed based on confusion matrices [17] and dialectologically defined variation matrices [15].

Unlike these works, we introduce a method of defining segment-to-segment distances by using raw speech features, that are the lower dimensions of cepstrum coefficients. The cepstrum coefficients are derived through inverse Fourier transform of a speech spectrum and the lower dimensions of the coefficients represent its envelope. If we denote the lower cepstrum dimensions of a spectrum as $\{c_i\}$ and those of another as $\{d_i\}$, it is easy to prove that the Euclid distance between them is proportional to the spectral gap (size of the gray area in Figure 2) between two power-normalized spectrum envelopes.

$$\sum_{i=1}^{N}(c_i - d_i)^2 \propto \text{spectral gap in Figure 2}$$

This measure is often used to assess the quality of synthesized speech, where the spectral gap between synthesized and natural speech is calculated [10].

By using the cepstrum coefficients, we build acoustic models of HMM (Hidden Markov Model) of phones and phonemes. For each model, we prepare multiple sample utterances of each segment. In each case, a speech unit, phone or phoneme, is modeled as three states, each of which contains a single Gaussian of cepstrum coefficients. The HMM-to-HMM distance is calculated by averaging the three state-to-state distances, shown in Figure 3. Although we can use any distance measure that can evaluate separability between two distributions, here we use Bhattacharyya distance [2].

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.

pliːz̚ kʰɔl stɛla æsk ə tu bɹĩŋ ðiːs θĩŋz wɪθ həɹ fɹə̃m n̥ə stɔːɹ sɪks spũːnz ə̯ fɪɛʃ snoʊ piːz faːɪv θɪk slæbz̚ əv blu ʧʃiːz ɛ̃nn̥ meɪbi ə snaːk fəɹ hɚ bɹʌðəɹ bɔːb wi ɔlso nid ə smɔl plæstɪk sneɪk ɛ̃n ə bɪg tʰɔɪ fɹɔːg fɚ ðə kʰɪːdz̚ ʃi kɛ̃n skup ðiz̚ θĩŋz ĩntũ θɹi ɹɛd bæːgz ɛ̃n wi wɪ goʊ miːt həɹ wɛ̃nzdeɪ æt ðə ɪɛĩn steɪʃɛ̃n

Figure 4: The elicitation paragraph of 69 words and an example of IPA transcription

Once we have a method of calculating segment-to-segment distances, we can align two transcripts and in our paper, we use the algorithm of Dynamic Time Warping (DTW) [8], which is similar to edit-distance-based alignment. A small difference between the two is that in edit-distance-based alignment, penalty of insertion or deletion is usually prepared based on prior knowledge but in DTW, those penalties are not needed and by using specific local path configurations, any segment in a sequence is mapped automatically to a segment in the other.

In speech technology, DTW is often used to align two sequences of cepstrum vectors, which mean two utterances. In this paper, the same algorithm is applied to two sequences of IPA symbols. Alignment of two phoneme sequences by DTW with phoneme HMMs is often done in studies of spoken term detection [5] but as far as we know, alignment of two IPA transcripts by DTW with phone HMMs was introduced for the first time in our work [4, 11] and its validity is tested in this paper in detail.

3. CORPUS AND MODELS

3.1. Speech material: Speech Accent Archive

The Speech Accent Archive (SAA) [13] is a resourceful dataset of readings in English collected from speakers with diverse language backgrounds. The archive provides both speech samples and their IPA transcripts with diacritic marks. We show the elicited paragraph and an example of transcription in Figure 4. The number of base phones in the transcripts is approximately 100 and when we consider the diacritics, the number goes up to more than 550.

Out of more than 2,000 speakers available in the SAA, by following [14], we extracted 115 native U.S.-born English speakers and 280 speakers whose native language is not American English (AE). For comparing alignment with phone HMMs to that with

phoneme HMMs, we also prepared the phonemic version of the IPA transcripts in the SAA. Here, phones were carefully mapped to AE phonemes that are defined in the CMU dictionary [9].

3.2. Acoustic models of phones and phonemes

We built an HMM for each of the most frequent 153 phones (with diacritics) in the SAA, which cover 95% of the phonetic instances found in the SAA. These models were built from each of two expert male phoneticians, P01 and P02. This means that the two sets of HMMs are speaker-dependent models. They were asked to pronounce each phone twenty times paying close attention to diacritic differences among phones sharing the same base phone.

To compare phone-based and phoneme-based similarity calculation, we also built HMMs for a total of 39 AE phonemes. Further, by using the algorithm widely used for training HMMs, we built more fine-grained phoneme HMMs [8]. Here, the phoneme is defined depending on its preceding and succeeding phonemes. If the number of phonemes is N, that of context-dependent phonemes is N^3 ($=59,319$ when $N=39$). This type of HMMs are called *triphone* models in ASR community although strictly speaking, they should be *triphoneme* models. To avoid confusion, we call them *triphoneme* models hereafter. If it is reasonable to consider diacritic variations of a base phone as context-based variations of that phone, *triphoneme* models may function in a similar way as phone models do.

We built both speaker-dependent (mono)phoneme models and triphoneme models from two General American (GA) speakers, F12 (female) and M08 (male). They are from the ERJ (English Read by Japanese) corpus [7], where speech samples of some native speakers are included as reference. They read 460 sentences and the two sets of HMMs were trained by using these continuous speech samples while phone HMMs were trained from isolatedly uttered phone instances. For training triphoneme models, the number of triphonemes can be reduced by merging the N^3 triphonemes using a top-down clustering tree. In this paper, the number of triphonemes was set to approximately 10,000 while that of monophonemes was 39 for each speaker.

Another set of monophoneme models were built by using a large number of speakers from the WSJ corpus [12], which are speaker-independent models. Speaker-independent *monophone* models are, however, very difficult to build because recording from a large number of expert phoneticians is impractical.

The HMMs prepared for this paper are summarized in Table 1, where 12 MFCCs and 12 ΔMFCCs were used as acoustic features in all the cases.

Table 1: The HMMs prepared for this paper

model ID	#models	spk-dependency
P01-monophone	153	dependent
P02-monophone	153	dependent
M08-monophoneme	39	dependent
M08-triphoneme	10K	dependent
F12-monophoneme	39	dependent
F12-triphoneme	10K	dependent
WSJ-monophoneme	39	independent

Table 2: Pearson corr. of segment-to-segment distances between training speakers ($p < .001$)

speakers	type	corr.
P01-P02	monophone	0.86
F12-M08	monophoneme	0.97
M08-WSJ	monophoneme	0.97

4. EXPERIMENTS AND DISCUSSION

4.1. Segment-to-segment distances

We first investigated speaker-independency of our method by calculating correlation between segment-to-segment distances of a set of HMMs and those of another set, where the two sets were obtained from different speakers. The results are shown in Table 2. By comparing P01-P02 and F12-M08, we can say that the phoneme-based distances appear to be more speaker-independent than the phone-based ones, although the latter distances will be found to be as speaker-independent as the former ones in the following section. Correlation between speaker-dependent models and speaker-independent models (M08-WSJ) is also extremely high. High independency is considered to be because we only focus on spectral contrasts (See Figure 2) and, if two speakers are of the same accent, we can claim that the magnitude of contrast are very similar between the two.

4.2. Transcript-to-transcript distances

Next, we examined speaker-independency with respect to transcript-to-transcript distances, which can be regarded as quantitative measure of accent gap. We also examined whether some differences can be found between phone-based models and phoneme-based models. For every pair of all the 395 IPA transcripts, we performed DTW and calculated word-unit distances between every corresponding word pair. Here, since longer words have larger distances, for normalization, we took the logarithm of the distances. Then, we calculated their average. It should be noted that while comparison can be done between phone-based models and between phoneme-based models as in Section 4.1, it is possible here even between a phone-based model and a phoneme-based

Table 3: Pearson correlation of transcript-to-transcript distances between different kinds of models ($p < .001$)

models	type	corr.
P01(mono)-P02(mono)	phone-phone	0.99
M08(mono)-M08(tri)	phoneme-phoneme	0.99
M08(tri)-P01(mono)	phoneme-phone	0.90

model. When phoneme-based models were used, as mentioned in Section 3.1, not phonetic transcripts but phonemic ones were aligned.

Table 3 shows three kinds of comparisons, phone-phone, phoneme-phoneme, and phoneme-phone. Unlike Section 4.1, P01-P02 (phone-phone) shows extremely high speaker-independency. It simply indicates that phone-based similarity calculation and phoneme-based similarity calculation between transcripts are equally speaker-independent.

M08(mono)-M08(tri) shows that accent gap measured with monophoneme models and that with triphoneme models are extremely correlated, indicating no difference between the two models.

It is very interesting to the authors that M08(tri)-P01(mono) shows less correlation compared to the above two cases. This correlation reduction is definitely not due to speaker difference (M08-P01) but due to speech unit difference (phoneme-phone). As we discussed in Section 3.2, triphoneme models (#models=10K) were expected to function similarly to monophone models (#models=153) but M08(tri)-P01(mono) shows that this expectation is not completely correct. Which one of triphoneme or monophone is more valid to calculate phonetic similarity between transcripts? We attempted to answer this question by comparing the two kinds of similarity scores obtained by DTW to subjective similarity scores rated through a large listening test.

4.3. Correlation to subjective native-likeness scores

We discuss how valid our method is for phonetic similarity calculation by making two comparisons, 1) phoneme-based and phone-based in our method, and 2) our method and other recently proposed methods in [14, 16]. Discussion is done by investigating the similarity scores calculated automatically and those obtained from human listening.

For this aim, we followed the experimental settings in [14] and used our method to predict the native-likeness score, where an input non-native transcript was aligned to all the 115 native transcripts and the average of the 115 scores was calculated as *machine* native-likeness score. This score is compared to its *human* native-likeness score, which is provided by [16], where subjective ratings were collected from over 1,000 native AE listeners. 286 samples were rated and each was rated by more than 50 participants. The human native-likeness score is the average of these ratings and the higher the score is, the more native-sounding the sample is.

Table 4: Pearson correlation between machine scores and human scores ($p < .001$)

method	corr.	corr. (logarithm)
phone-based	-0.81	-0.83
AE phoneme-based	-0.77	-0.80
PMI	-0.77	-0.81
NDL	-0.75	-0.82

Table 4 shows the results. PMI (pointwise mutual information) [14] and NDL (naïve discriminative learning) [16] are recently proposed methods, where algorithms were developed with non-acoustic definition of segment-to-segment distances [14] and association strengths between cues and outcomes [16]. Clearly shown, phone-based gives higher correlation than phoneme-based and phoneme-based gives similar correlation to phone-based PMI and NDL. We can say that for phonetic similarity calculation, use of phonetic knowledge is more effective for defining segments than simple context-based sophistication of phonemes. As suggested in [16], introduction of log transformation improves the performance and our phone-based method shows the best result again.

After the experiments, we examined a supervised framework of prediction, i.e., regression. Since we had 115 native transcripts (native speakers) and each transcript was divided into 69 words, we prepared 115 speaker-dependent similarity scores and 69 word-dependent scores as explanatory variables. Use of sophisticated frameworks such as Support Vector Regression and Ridge Regression can avoid the overfitting problem. It was surprising that these frameworks did not show any improvement although our method is totally unsupervised. To improve correlation, we suggest resorting to non-phonetic aspects of accents, such as speech prosody.

5. CONCLUSIONS

We examined the validity of our method of phonetic similarity calculation using DTW and acoustic models of phones. Results showed high speaker-independency and superiority over phoneme-based models and recently proposed methods. Our method tested in this paper requires phoneticians' manual transcription but in [6], we already proposed another method of *automatic* prediction of IPA-based phonetic similarity calculated by our method introduced in this paper. No transcript is needed for automatic prediction. Interested readers should refer to [6].

6. REFERENCES

[1] Bailey, T. M., Hahn, U. 2005. Phoneme similarity and confusability. *Journal of Memory and Language* 52(3), 339–362.

[2] Bhattacharyya, A. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* 35, 99–109.

[3] Heeringa, W. J. *Measuring dialect pronunciation differences using Levenshtein distance*. PhD thesis Rijksuniversiteit Groningen.

[4] Kawase, Y., Minematsu, N., Saito, D., Hirose, K. 2014. Visualization of pronunciation diversity of world englishes from a speaker's self-centered viewpoint. *Oriental COCOSDA* Phuket. 149–153.

[5] Konno, K., Itoh, Y., Kojima, K., Ishigame, M., Tanaka, K., Shi-Wook, L. 2014. High priority in highly ranked documents in spoken term detection. *Proc. Asia-Pacific Signal and Information Processing Association* Angkor Wat.

[6] Minematsu, N., Kasahara, S., Makino, T., Saito, D., Hirose, K. 2014. Speaker-basis accent clustering using invariant structure analysis and the speech accent archive. *ISCA tutorial and research workshop of Odyssey* Joensuu. 158–165.

[7] Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., Makino, S. 2004. Development of english speech database read by japanese to support call research. *the 18th International Congress on Acoustics* Kyoto. 557–560.

[8] Rabiner, L., Juang, B.-H. 1993. *Fundamentals of Speech Recognition*. Prentice Hall.

[9] Rudnicky, A. 2007. The cmu pronunciation dictionary, release 0.7a. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[10] Saito, D., Watanabe, S., Nakamura, A., Minematsu, N. 2012. Statistical voice conversion based on noisy channel model. *IEEE Trans. Audio, Speech and Language Processing* 20(6), 1784–1794.

[11] Shen, H.-P., Minematsu, N., Makino, T., Weinberger, S. H., Pongkittiphan, T., Wu, C.-H. 2013. Speaker-based accented english clustering using a world english archive. *Proc. ISCA workshop of SLaTE* Grenoble. 184–188.

[12] Vertanen, K. 2006. Baseline wsj acoustic models for htk and sphinx: Training recipes and recognition experiments. Technical report Cambridge, UK: Cavendish Laboratory.

[13] Weinberger, S. H., Kunath, S. A. 2011. The speech accent archive: towards a typology of english accents. *Language and Computers* 73(1), 265–281.

[14] Wieling, M., Bloem, J., Mignella, K., Timmermeister, M., Nerbonne, J. 2014. Measuring foreign accent strength in english. *Language Dynamics and Change* 4(2), 253–269.

[15] Wieling, M., Margaretha, E., Nerbonne, J. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics* 40(2), 307–314.

[16] Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., Baayen, R. H. 2014. A cognitively grounded measure of pronunciation distance. *Public Library of Science (PloS) one* 9(1), e75734.

[17] Žgank, A., Horvat, B., Kačič, Z. 2005. Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity. *Speech Communication* 47(3), 379–393.