Normalization and matching routine for comparison of native speaker and non-native speaker tongue trajectories

Shusuke Moriya, Yuichi Yaguchi, and Ian Wilson

University of AIZU, Aizuwakamatsu, Fukushima, Japan
{ m5191128, yaguchi, wilson}@u-aizu.ac.jp

Abstract. The main purpose of this research is to specify articulation difference between native and non-native speakers by digitizing tongue motions and analyzing the difference between utterances. Differences in tongue motion directly influence speaker's pronunciation, therefore it may be possible to improve non-native speaker's efficiency of pronunciation practice with the relevant feedback and visualization. It is necessary for comparison of native and non-native speakers' tongue motions to that end, however, normalization is absolutely necessary to remove the influence of anything except tongue motion before comparison, because every person has a unique shape and size. In this paper, we use coronal cross section of the tongue taken by ultrasound scanner to carry out the following: first record the ultrasound of speaker's tongue motion using the corpus "The Boy Who Cried Wolf." Then, sample tongue motion by using a histogram of oriented gradients and Karhunen-Loeve expansion. Next, apply eight prepared normalizations to tongue motions. Finally, compare each tongue motion per frame via dynamic time warping and correlation coefficient. The experimental result allowed us to compare with speaker's tongue motions in sentences which were recorded in different environments or by different speakers and to point out non-native speaker's speaking errors.

Keywords: Image Processing, Midsagittal Ultrasound Tongue Image, Histogram of Gradients, Correlation Coefficient

1 Introduction

English is spoken all over the world and so it is important as a standard for the global communication. Therefore, non-native speakers study English for second language. However, it is very hard to improve their speaking skill. Most of them adopt the method of modeling our tongue and mouth motion on a native speaker's one by trial and error from detecting speaking error with their own hearing. In other words, detecting their own speaking error depends on their or someone's hearing. If tongue motion is visualized, they should improve a speaking skill more efficiently.

The sectional tongue images which is taken by ultrasound, CT, or MRI are used for visualization of tongue motion. Especially, a lot of papers refering to

ultrasound have been advanced [1]. Transformation of tongue and mouth makes various sounds. The tongue, especially, assumes an important role for speaking and its shape is transformed greatly. For these reasons, there are many papers such as the study of distinction of phonemes [2] [3] and the study of making sounds by transforming artificial tongue and vocal tract [4]. Therefore, comparing tongue motion between native and non-native speakers makes this an appropriate approach for non-native speakers to improve their speaking skill. However, every person has a unique shape and size. In addition, an image capturing error affects a tongue image directly. It is impossible to compare between tongue images without normalization.

Detecting someone's speaking error requires sampling pure tongue motion while extracting the influence of the environment and tongue shape. Under the present condition, numeric data of tongue motion was completely separated by environment and subjects in midsagittal ultrasound tongue image space (MUTIS). In this paper, therefore, our goal is developing the algorithm which removes them influence from numeralized data. Under our research process, at first, we obtain the ultrasound tongue image from participants who read out the story "The boy who cried wolf". Then, we plot tongue motion as 1872 dimension vectors in MUTIS with histograms of oriented gradients [5]. We use the K-L expansion [6] to compress these vectors to 128 dimension vectors. Compressed vectors are normalized by eight prepared methods. Next, we compare normalized vectors with dynamic time warping (DTW) [7]. At this time, we apply correlation coefficient to the result of comparison and emphasize features of error. Finally, we pick up the best normalization method from eight of them.

In this paper, at first, we describe the method of data correction and plotting a tongue motion from ultrasound image to MUTIS. In addition, we list the normalization methods, then report the result.

2 Method

The voice is formed from vowels that are produced by comparatively open configuration of the vocal tract and consonants that are articulated with completely or partial closure. And, most changes in voice are caused by transformation of the tongue above the vocal tract. Therefore, it is possible to classify the pronunciations via analyzing tongue trajectory in the vocal tract.

2.1 Data Collection

Through the data collection, participants ultrasound tongue images were video-recorded via the probe stabilization method [8]. The image ratio of width to height is 4:3. The reason "The Boy Who Cried Wolf" [9] was used is that it is widely known that the sentence keeps the balance of phonemes. Normally, ultrasound tongue images are taken by transducer which is held to the speaker's jaw. However, these ultrasound tongue images are distorted because of hand and jaw shake. Therefore, the original photographic device that is composed of

Fig. 1: Recording Environment. 1. Ultrasound camera, 2. Corpus is shown on a display, 3.Combination of helmets and transducer, 4. iMac for recording ultrasound movies.

a helmet and a transducer like the Fig.1, was used to solve the problem. The device reduced unintentional transducer motion and distortion of images, and tongue images became more accurate.

2.2 Extraction of tongue motions on MUTIS

The numeralization method that is based on Histogram of oriented gradients and spectrum vector field [10] was used for capturing the tongue curve on ultrasound tongue image. Before numeralization, the original image had unnecessary labels along both the left side and the top side. (Fig.2(a)). It was necessary to remove these labels to make pure ultrasound tongue images. Next, tongue images were de-noised via median filter and log filter because, the high stationary noise that occurs by the unevenness of ultrasound obstructs the detection of tongue edge from images. The filtered image is written as

$$P(t) \in (i, j|n), i, j, n \in \mathbb{N},$$

$$0 \leq i \leq I-1, 0 \leq j \leq J-1, 0 \leq n \leq N-1$$

Two gradients $I'_x(t), I'_y(t)$ were made from one filtered image via a sobel filter (Fig.2(b)). These gradients $P'_x(t), P'_y(t)$ describe that the edge intensity $P_e(t)$ and edge angle $P_\alpha(t)$ are written as the following function.

$$P_e(t) = \|P'_x(t) - P'_y(t)\|_{L2}$$
$$P_\alpha(t) = \tan^{-1} \frac{P'_y(t)}{P'_x(t)}$$

Then, images were segmented 20x15 block to create a histogram of the edge angle for each block quantized 8-directions (Fig.2(c)). Combine 3x3 blocks with

normalization for expressing partial tongue shape with the sum of edge intensities. Finally, integrate 13x18 blocks without margin blocks into 8x13x18 = 1872 vector fields for each frame, because it is impossible to combine 3x3 blocks at margin cell. It takes an impractical amount of time to solve the similarity with these 1872-dimension vectors. Therefore, the 1872-dimension vectors were compressed via K-L expansion to create 128-dimension subspace.

Fig.3 shows three participant's tongue trajectories via the previous method in MUTIS. These three trajectory mean tongue trajectories from the start to the end of the corpus. Fig.4 also shows the trajectories separated by words. From the figures, participant's tongue trajectories are separated completely in MUTIS. In addition, the cumulative contribution ratio of dimension compression is about 70 percent and there is a high possibility that information lack distorted tongue trajectory. In this figure, although points of MUTIS are classified by person, shape of each trajectory by word is similar in each class. Thus, we assume that it is possible to compare with only tongue trajectories with normalization.

2.3 Normalization of tongue trajectory

Our previous study [11] shows how the tongue trajectories are plotted in MUTIS via the method in section 2.2. In conclusion of the study, MUTIS is affected by not only pronunciation feature but also personal characteristics in the same space. Therefore, it is essential to normalize distributions of tongue trajectory from the utterance start point before getting similarity of tongue motion. In the process of normalization, the mentioned 128-dimension vectors are written as tongue motion

$$T_{Aw}(t) = (x_{Awt1}, x_{Awt2}, \ldots, x_{Awt128})^T$$

where A is speaker, w is word, and t is time. In this paper, we prepared three basic normalization methods and combined them into eight streams to normalize tongue motions. The first basic normalization method (NM) is solved with the following expression.

$$N_{Aw}(t) = \frac{T_{Aw}(t)}{\sum_{k=2}^{N} T_{Aw}(1) - T_{Aw}(k)}$$

In this method, tongue motions are normalized by calculating an average vector. The second one (DL) is getting time subtraction vectors that is written as

$$\Delta T_{Aw}(t) = T_{Aw}(t) - T_{Aw}(t-1)$$

The third one (KL) is using transformation to the principal axis via K-L expansion. Eight normalization streams are constructed from the combination of these basic methods in order. The following Table,1 shows the constructed streams. We applied them to tongue motions.

DL	NM
$DL \rightarrow KL$	$NM \rightarrow DL$
$KL \rightarrow DL$	$NM \rightarrow KL$
$NM \rightarrow KL \rightarrow DL$	$NM \rightarrow DL \rightarrow KL$

Table 1: Eight Normalization Streams

2.4 The calculation of tongue motions similarity

Dynamic time warping was used for getting similarity of tongue motions because, there was time expansion on pronunciation for each word. This method is also used for a voice matching system. The asymmetrical DTW was adopted in our research and written as

$$D(t,\tau) = \sqrt{L_{Aw}(t)^2 - L_{Bw}(\tau)^2} + \min \begin{cases} D(t-1,\tau) \\ D(t,\tau-1) \\ D(t-1,\tau-1) \end{cases}$$

where D is accumurated distance, A and B are participants, w is word, t and τ are time, and $L_{Aw}(t)$ and $L_{Bw}(\tau)$ are normalized tongue motions.

2.5 Correlation coefficient for error detection

In the case of comparisons with a non-native speaker and native speakers, correlation coefficient can be used for detecting common speaking errors from them. In addition, high correlation coefficient may show that the normalization method indicates more accurate common speaking errors. Therefore, correlation coefficient is barometer of finding the best normalization method.The correlation coefficient C is written as

$$C_t = \frac{\sum_{i=t-7}^{t+7} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=t-7}^{t+7} (x_i - \bar{x})^2} \sqrt{\sum_{i=t-7}^{t+7} (y_i - \bar{y})^2}} | t = 8, 9, \ldots, n-7$$

where n is number of tongue trajectory frames, t is current frame, x, y are local distances of comparison speaker's tongue trajectory through DTW, and $\bar{x}, \bar{y}$ are average of them.

3 Experimentation

3.1 Experimental environment

In experimentation, one English learner of Japanese, and two natives of English recited the story "The boy who cried wolf" and their ultrasound tongue images

were recorded. The ultrasound camera was Toshiba PVQ-381A. The transducer Famio-8 connected to it through DV converter CanopusADV-700. iMac was used for video recording with 30 frames per second. The recorded video size of width to height was 720:480. After removal of unnecessary labels by cropping the image, image size became 640:480. We omitted a detailed explanation of sound recording environment because, our experiment did not demand sound data.

3.2 Corpus

The story keeps the balance of phonemes in "The Boy Who Cried Wolf". It means that all sounds are necessary to speak this story, in other words, non-native speaker's weak point of pronunciation would certainly occur. Therefore, in this paper, we used the first five sentences from "The Boy Who Cried Wolf."

Sentence 1
There was once a poor shepherd boy who used to watch his flocks in the field next to a dark forest near the foot of a mountain.

Sentence 2
One hot afternoon, he thought up a good plan to get some company for himself and also have a little fun.

Sentence 3
Raising his fist in the air, he ran down to the village shouting "Wolf, Wolf."

Sentence 4
As soon as they heard him, the villagers all rushed from their homes, full of concern for his safety, and two of his cousins even stayed with him for a short while.

Sentence 5
This gave the boy so much pleasure that a few days later he tried exactly the same trick again, and once more he was successful.

We obtained the similarity of tongue trajectory from start to end for each sentence.

3.3 Comparison of eight normalization streams

Our purpose is to find the method that treats only tongue trajectory to compare speakers. Consequently, we have to describe how the best normalization works. The ideal normalization must show that similarity between native speakers' tongue trajectory is high and similarity between non-native and each native speakers' one is equivalent. This assumes that native speakers speak with equivalent pronunciation. Therefore, almost the same difference occurs between them and non-native speaker's one. This estimation presupposes that native speakers are from the same region and their tendency of accent is the same. In addition, DTW has similarity as an accumulation distance of tongue motion gap. In other words, a high similarity means that an accumulation distance is low. With the above in mind, in Table.2, NM→KL and DL→KL streams follow the ideal normalization. Therefore, these two normalizations are suitable for the comparison.

	Sentence 1			Sentence 2			Sentence 3		
Stream	A:B	B:C	C:A	A:B	B:C	C:A	A:B	B:C	C:A
None	9.14299	10.3628	10.7264	7.51989	7.24127	8.15609	1.91515	5.25169	5.92162
DL	7.43926	7.46136	8.38080	5.87846	5.43959	6.91317	2.42108	4.67287	5.44039
DL→KL	0.00144	0.00106	0.00143	0.00123	0.00080	0.00119	0.00071	0.00052	0.00059
KL→DL	0.00217	0.00165	0.00225	0.00187	0.00115	0.00176	0.00115	0.00084	0.00096
NM	0.00176	0.00135	0.00176	0.00155	0.00087	0.00155	0.00089	0.00072	0.00065
NM→DL	7.43926	7.46136	8.38080	5.87846	5.43959	6.91317	2.42108	4.67387	5.44039
NM→DL→KL	0.00176	0.00135	0.00171	0.00155	0.00087	0.00151	0.00089	0.00072	0.00065
NM→KL→DL	0.00217	0.00168	0.00221	0.00187	0.00115	0.00176	0.00115	0.00084	0.00096
NM→KL	0.00144	0.00110	0.00144	0.00123	0.00080	0.00119	0.00072	0.00052	0.00059

	Sentence 4			Sentence 5		
Stream	A:B	B:C	C:A	A:B	B:C	C:A
None	9.16012	10.6057	11.9748	8.69147	2.05053	9.05096
DL	6.74165	8.98707	11.2731	7.46197	2.18722	8.11911
DL→KL	0.00207	0.00152	0.00214	0.00155	0.00137	0.00164
KL→DL	0.00315	0.00234	0.00316	0.00242	0.00213	0.00238
NM	0.00289	0.00185	0.00300	0.00200	0.00173	0.00203
NM→DL	6.74165	8.98797	11.2731	7.46197	2.18722	8.11911
NM→DL→KL	0.00289	0.00185	0.00300	0.00200	0.00173	0.00203
NM→KL→DL	0.00315	0.00234	0.00316	0.00249	0.00189	0.00238
NM→KL	0.00207	0.00152	0.00214	0.00156	0.00120	0.00164

Table 2: Determination of Most Effective Normalization Stream. A is English leaner of Japanese. B and C are natives of English. DL→KL and NM→KL streams follow the nearest ideal normalization.

3.4 Discussion

Correlate coefficient was used for selecting best normalization method from previous streams. Fig.5 shows correlate coefficient and result of compasion through the sentence 1. Blue and red lines are comparison of non-native speaker and native speakers, yellow one is correlate coefficient between red and blue line, and green one is products of blue and yellow line. DL→KL stream shows that green line is always below 0.2 point. On the other hand, NM→KL stream that green line approaches 0.4 point at some parts such as first and second highest parts. It means that correlate coefficient is high and local error is also high in the parts. The tongue shape of non-native speaker A is differ hard from native speaker's one in these parts. The difference makes non-native speaker to speak incorrect pronunciations. Therefore, NM→KL stream meets more necessary condition of ideal normalization than DL→KL stream.

There are six ultrasound tongue images that were used for comparison of participants in Fig.6. These trackback images show that tongue shapes at first and second highest part of green line through NM→KL stream. From this figure, it is possible to notice that tongue shapes for pronunciation are completely different in circles. This difference links non-native speaker's speaking errors. Therefore,

the above justifies that NM→KL stream can find non-native speaker's speaking errors. However, two lines were not overlapping completely, so our normalization could not remove the whole influence of anything except tongue motion. For these reasons, NM→KL normalization stream is reasonably possible to extract non-native speaker's speaking errors, but is incomplete.

4 Conclusion

In our research, we just compared tongue motions in each sentence. The resulting granularity is too large to divert it into non-native speakers' pronunciation practice. In addition, normalized data has a little characteristic of a tongue shape or an environment and we were confused whether it was a speaking error or not in some parts. It is because not only the normalization method incompletely, but also, there are many big noises through numeralization.

We have to confirm the reproducibility of comparison in each word and phoneme, and hope to classify participants clearly into speaking level and accent. Therefore, participants who have variable speaking skill and accent, especially non-native speakers who are not English or from Japanese language regions. Finally, we need to develop an image numeralization algorithm which de-noises ultrasound tongue images strongly from a noisy measurement and compresses numeric data with high cumulative contribution ratio.

Acknowledgement

I would like to express thanks and appreciation to Prof, Yuichi Yaguchi and Prof, Ian Wilson for their appropriate advice in my research and Prof. John Brine for advice on the thesis. We also thank students in Image Processing Lab in the University of Aizu.

References

1. B. Denby and M. Stone. Speech synthesis from real time ultrasound images of the tongue. In *in Proc. of IEEE ICASSP 2004*, volume 1, pages 685–688, 2004.
2. M. Rokibul Alam Kotwal, F. Hassan, M.M. Alam, A.R. Khan Jehad, M. Arifuzzaman, and M. Nurul Huda. Recurrent neural network based phoneme recognition incorporating articulatory dynamic parameters. *Advances in Computing and Communications*, pages 349–356, 2011.
3. A. Roy, M. Magimai-Doss, and S. Marcel. Phoneme recognition using boosted binary features. In *in Proc. of IEEE ICASSP 2011*, pages 4868–4871, 2011.
4. K. Fukui, K. Nishikawa, S. Ikeo, E. Shintaku, K. Takada, H. Takanobu, M. Honda, and A. Takanishi. Development of a talking robot with vocal cords and lips having human-like biological structures. In *in Proc. of IROS 2005*, pages 2023–2028. IEEE, 2005.
5. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *in Proc. of IEEE CVPR 2005*, volume 1, pages 886–893, 2005.

6. Jian Yang, David Zhang, and J-Y Yang. A generalised kl expansion method which can deal with small sample size and high-dimensional problems. *Pattern Analysis and Applications*, 6(1):47–54, 2003.
7. D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. *AAAI-94 Workshop on KDD*, pages 229–248, 1994.
8. M. Stone and E.P. Davis. A head and transducer support system for making ultrasound images of tongue/jaw movement. *J. of the ASA*, 98(6):3107–3112, 1995.
9. David Deterding. The north wind versus a wolf: short texts for the description and measurement of english pronunciation. *Journal of the International Phonetic Association*, 36(02):187–196, 2006.
10. M Ihara, T Akasaka, and R Oka. Comparison of features of mel-cepstrum and spectrum vector fields in phoneme recognition based on the bayes discrimination function. Technical report, IEICE Technical Report, 2000.
11. Keita Sano, Yuichi Yaguchi, and Ian Wilson. Comparing l1 and l2 phoneme trajectories in a feature space of sound and midsagittal ultrasound tongue images. *The Journal of the Acoustical Society of America*, 132(3):1934–1934, 2012.

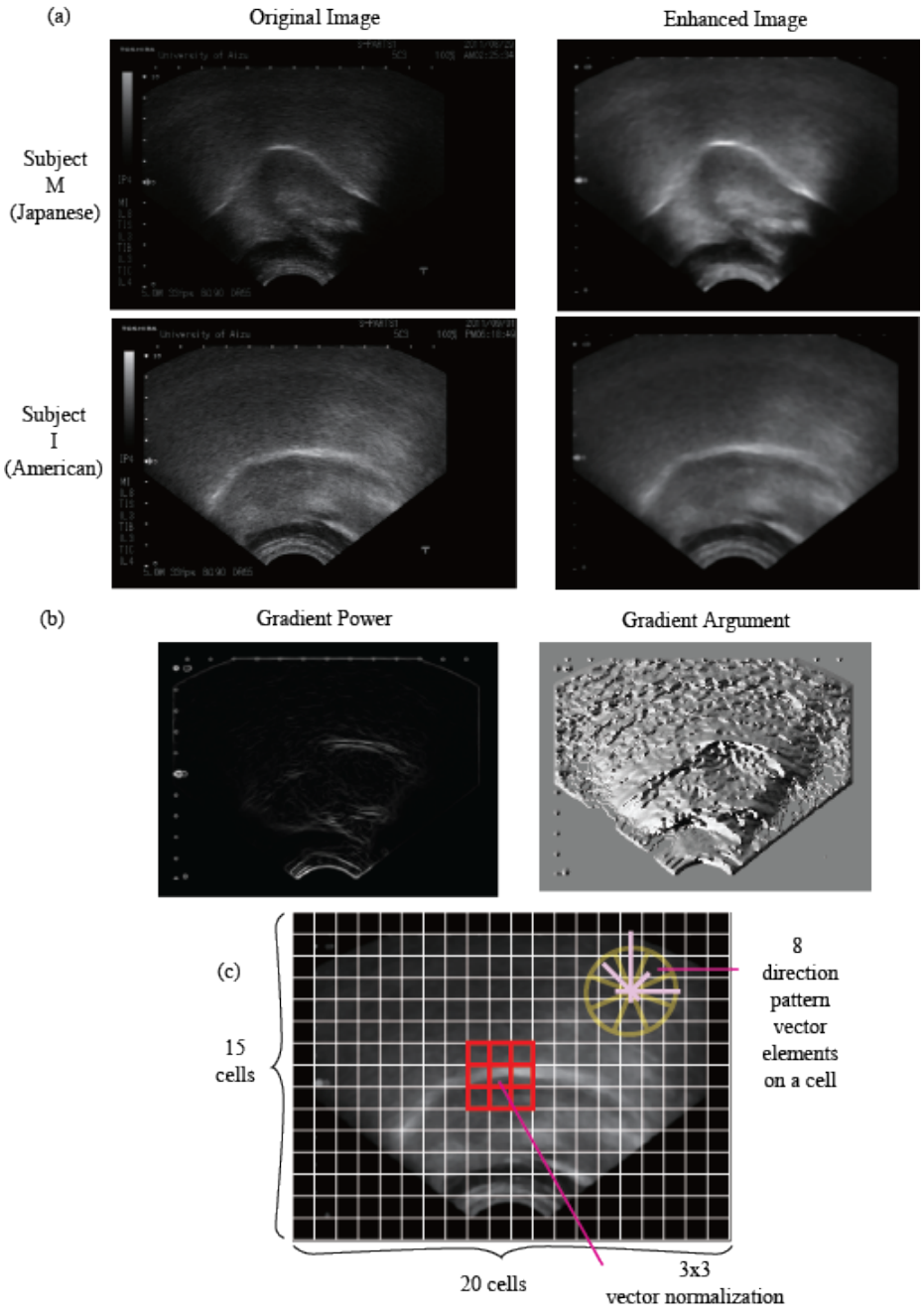


Fig. 2: Constructing Midsagittal Ultrasound Tongue Image Space. Through the process, ultrasound tongue images become 128-dimension vectors as tongue motion.

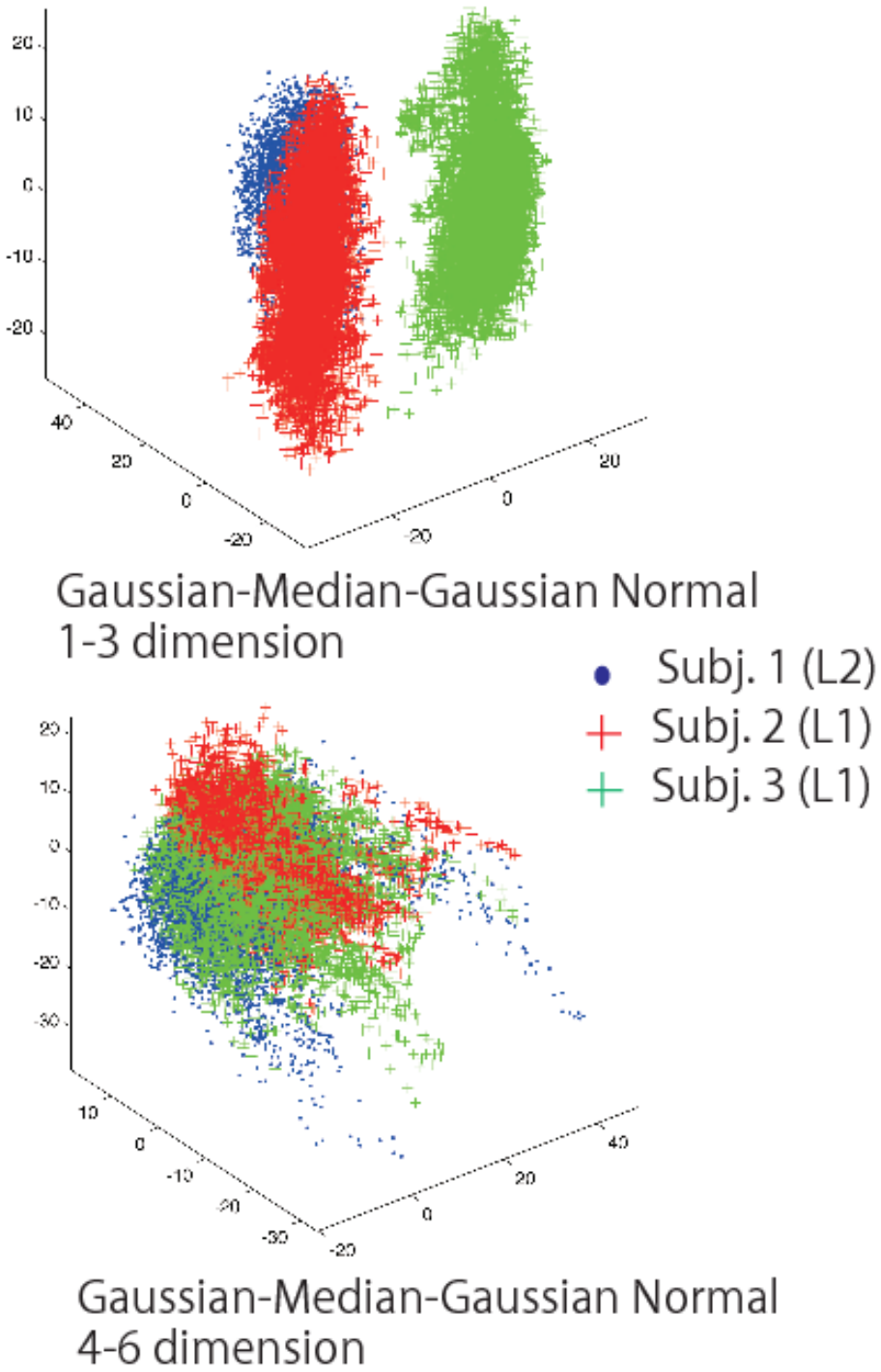


Fig. 3: Plotted Tongue Trajectories of the Full Sentence. Tongue trajectories are completely separated in high dimention of MUTIS.

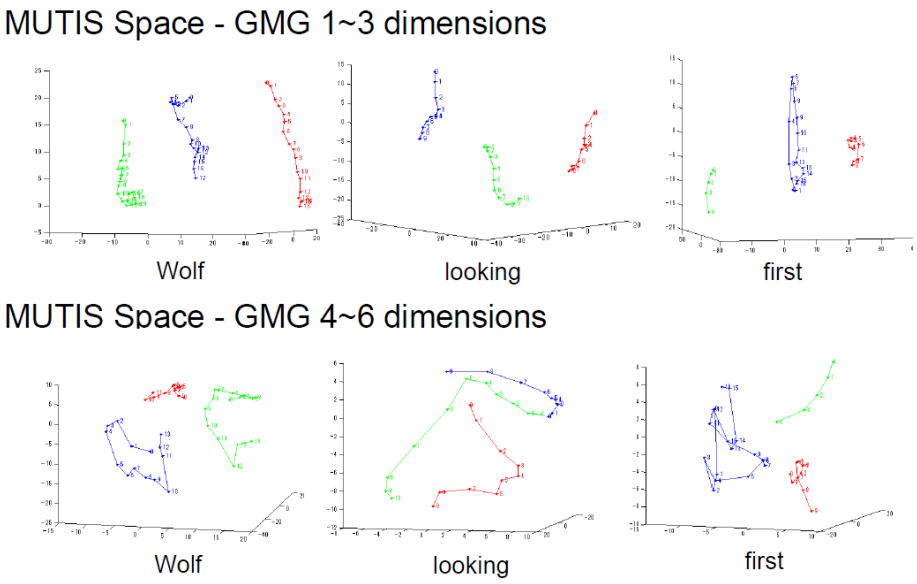
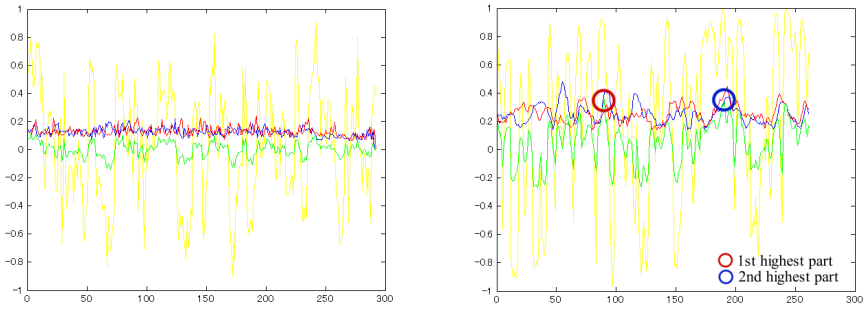


Fig. 4: Plotted Tongue Trajectories of Each Word.



DL → KL Stream NM → KL Stream

Fig. 5: Plotting correlate coefficient in DTW. Horizontal and vertical axis mean time(frame) and local distance via DTW. Blue line is comparison of non-native speaker A and native speaker B, red one is comparion of A and native speaker C, yellow one is correlate coefficient between red and blue line, and green one is products of blue and yellow line. Through NM→KL Stream, green line approaches 0.4 point at some parts.

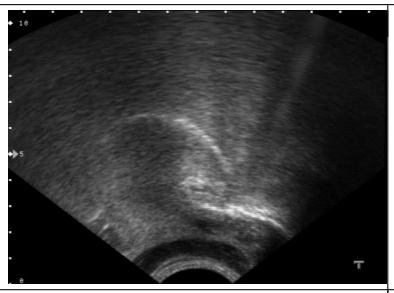
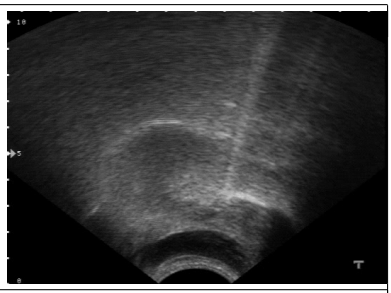
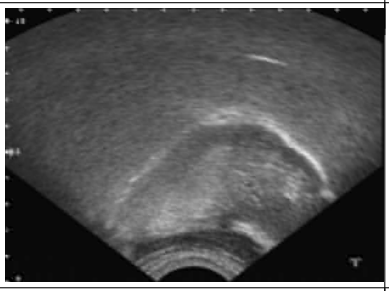
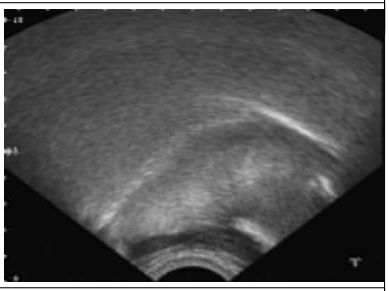
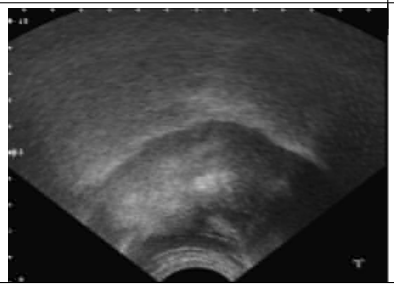
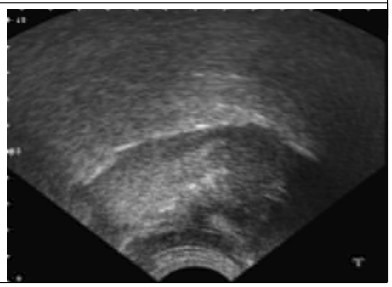
	NM → KL Stream (In Fig.5)	
Participant	1st highest part ("shepherd")	2nd highest part ("dark")
A		
B		
C		

Fig. 6: Trackback of Comparison via NM→KL Stream. At first highest part, participants said the word "shepherd", At second highest part, they said the word "dark". A's tongue shape is clearly different from others. This figure shows that we could find non-native speaker's speaking error in red circle in Fig.5 with comparison via NM→KL stream.