Automatic Recognition of Connected Vowels Only Using Speaker-invariant Representation of Speech Dynamics

Satoshi Asakawa[1], *Nobuaki Minematsu*[1], *Keikichi Hirose*[2]

[1]Graduate School of Frontier Sciences, The University of Tokyo
[2]Graduate School of Information Science and Technology, The University of Tokyo
{asakawa, mine, hirose}@gavo.t.u-tokyo.ac.jp

Abstract

Speech acoustics vary due to differences in gender, age, microphone, room, lines, and a variety of factors. In speech recognition research, to deal with these inevitable non-linguistic variations, thousands of speakers in different acoustic conditions were prepared to train acoustic models of individual phonemes. Recently, a novel representation of speech dynamics was proposed [1, 2], where the above non-linguistic factors are effectively removed from speech as if pitch information is removed from spectrum by its smoothing. This representation captures only speaker- and microphone-invariant speech dynamics and no absolute or static acoustic properties such as spectrums are used. With them, speaker identity has to remain in speech representation. In our previous study, the new representation was applied to recognizing a sequence of isolated vowels [3]. The proposed method with a single training speaker outperformed the conventional HMMs trained with more than four thousand speakers even in the case of noisy speech. The current paper shows the initial results of applying the dynamic representation to recognizing continuous speech, that is connected vowels.
Index Terms: speech dynamics, robust invariance, structure

1. Variable substances, invariant dynamics

Many speech sounds are produced as standing waves in a vocal tract and acoustic properties of the waves depend on the shape of the vocal tube. Different shapes cause different timbre. No two humans have the same tube and then, speech acoustics vary. Many speech sounds are produced as voiced with vibrations of a vocal cord and *F0* of the sounds depends on the length, tension, mass of the cord. Shorter and lighter cords vibrate more rapidly. No two humans have the same cord. Speech acoustics vary again.

Pitch is physically characterized by *F0* and its dynamic pattern is often visualized. Difference in the length or mass translates the pattern to be higher or lower globally. Dynamic changes of *F0* are invariant and, due to this invariance, we can easily find the equivalence between two *F0* patterns of the same linguistic content although they are absolutely different. This is the case with music. Transposition of a musical piece does not change its melody. Many people verbalize a piece and its transposed version as the same sequence of syllable names. Here, perception of Do, the tonic sound, occurs completely irrespective of sound substances. They perceive a scale structure in the melody and, within the structure, a certain sound will recall an internal voice of Do. They *hear* voices of Do, Re, Mi, etc.

As is well-known, a process of producing a vowel is very similar to that of producing a sound with a wind instrument. A vocal tract is an instrument and, by changing its shape, sounds of different timbre are generated, called speech sounds. Music is composed of dynamic changes of pitch and speech is composed of dynamic changes of timbre. The former dynamics are easily formulated to be invariant among different instruments and speakers. What about the latter dynamics? For a variety of non-linguistic factors, is it possible to formulate them as invariant? Figure 2 shows a piano sound sequence of CDEFG and a speech sound sequence of /aiueo/ with the Japanese vowel chart. The two dynamic patterns are illustrated in phase spaces, where pitch is defined physically as one-dimensional feature of *F0* and timbre is tentatively defined as two-dimensional feature of *F1* and *F2*. Cepstrum coefficients can also be used to expand a 10- to 20-dimensional phase space. As is shown in the vowel chart, it is often said in phonetics that the vowel structure of male speech can be translated to become that of female speech. If this is correct enough, the timbre dynamics can be easily formulated to be invariant because speaker difference only translates the sound structure. However, every speech engineer knows that this idea is too simple to apply to real world speech data.

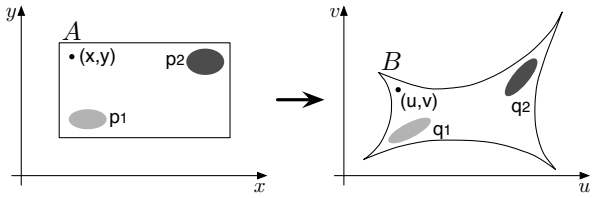

Figure 1: Linear or non-linear mapping between two spaces

What sort of function can map the acoustic space of speaker A into that of speaker B? Linear or non-linear? This question has been frequently raised in the research of speaker adaptation in speech recognition and speaker conversion in speech synthesis. Figure 1 shows two acoustic spaces of speakers A and B. Acoustic events of p_1 and p_2 of A are mapped to those of q_1 and q_2 of B, respectively. It is easily supposed that a mapping function of A's entire space into B's entire space has to be very complicated. Further, the form of the function will depend on both the source and target speakers. These indicate that, if one wants to focus on invariance in the timbre dynamics, he has to derive some invariant acoustic observations with respect to any form of mapping function. Is the *robust* invariance possible?

2. Robust and structural invariance

The answer is definitely yes if the two spaces have one-to-one correspondence [4]. Point (x, y) in space A is uniquely mapped to (u, v) in space B and vice versa. In the following, a two-dimensional space is used for explanation but it does not reduce the generality. Every event is characterized as distribution.

$$1.0 = \iint p_i(x,y)dxdy, \quad 1.0 = \iint q_i(u,v)dudv$$

Here, we consider functions of f and g for the mapping.

$$x = f(u,v), \qquad y = g(u,v)$$

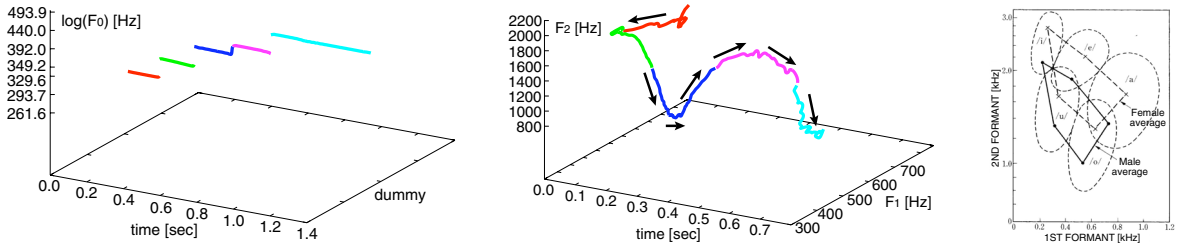

Figure 2: Dynamic changes of pitch in CDEFG and those of timbre in /aiueo/ with the Japanese vowel chart

f and g can be non-linear. Even when they cannot be represented by any known analytical expressions, the following discussion is effective. By f and g, any integral operation in space A can be rewritten as its counterpart in space B.

$$\begin{aligned}\iint \phi(x,y)dxdy &= \iint \phi(f(u,v),g(u,v))|J(u,v)|dudv \\ &= \iint \psi(u,v)dudv,\end{aligned}$$

where

$$\psi(u,v) = \phi(f(u,v),g(u,v))|J(u,v)|.$$

$J(u,v)$ is Jacobian. Any p_i in A can be mapped into q_i in B.

$$q_i(u,v) = p_i(f(u,v),g(u,v))|J(u,v)|.$$

Physical properties of p_i are different from those of q_i. p_1 may represent /a/ of speaker A and q_1 may represent /a/ of B. What can be robustly invariant between a set of p_is in space A and a set of q_is in space B? Let us consider Bhattacharyya distance, one of the distance measures between two distributions.

$$\begin{aligned}BD(p_1,p_2) &= -\log \iint \sqrt{p_1(x,y)p_2(x,y)}dxdy \\ &= -\log \iint \sqrt{p_1(f(u,v),g(u,v))|J| \cdot p_2(f(u,v),g(u,v))|J|}dudv \\ &= -\log \iint \sqrt{q_1(u,v)q_2(u,v)}dudv = BD(q_1,q_2)\end{aligned}$$

BD between two events in space A and BD between their corresponding two events in space B cannot be changed. Events can change easily but difference between the events cannot change by any transformation. This invariance does not require any calculation or formulation of functions f and g and Jakobian $J(u,v)$. If distributions of events can be estimated correctly, we can easily find the invariance of distances. This invariance is also satisfied with other distance measure, such as Kullback-Leibler distance. In this paper, BD was adopted because preliminary experiments showed its superiority.

The shape of a triangle is determined uniquely if the length of the three segments is given. The shape of n points in a geometrical structure is determined uniquely if the length of all the ${}_nC_2$ diagonal segments is given. In other words, if a distance matrix is given for n points, the matrix determines the shape of the n-point structure uniquely. As told above, BD is robustly transformation-invariant. Given n distributions, a BD-based distance matrix derives its robustly-invariant structure.

What type of transformation can effectively specify non-linguistic speech distortions? They are often classified into three types; additive, convolutional, and linear transformational distortions. The last two types will be the focus of this paper because the first type is not inevitable. Microphones and rooms are typical causes of convolutional distortion. If a speech event is represented by cepstrum vector c, this distortion changes c into $c'=c+b$. Vocal tract length difference is a typical example of linear transformational distortion. It is often modeled as frequency warping of the log spectrum and it can be appropriately modeled as $c'=Ac$ [5]. Various distortion sources are found in speech communication but the total distortion due to the *inevitable* sources, A_i and b_i, is eventually modeled as $c'=Ac+b$, known as affine transformation. Figure 3 schematizes the spectral distortions due to A_i and b_i, corresponding to horizontal and vertical ones, respectively. Although this model is linear and the simplest, the BD-based structure is invariant with more complicated transformations such as non-linear ones.

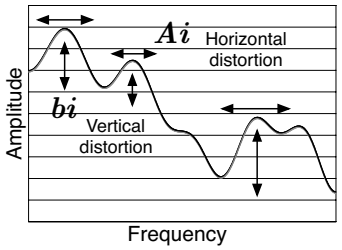

Figure 3: Spectral distortions caused by A_i and b_i

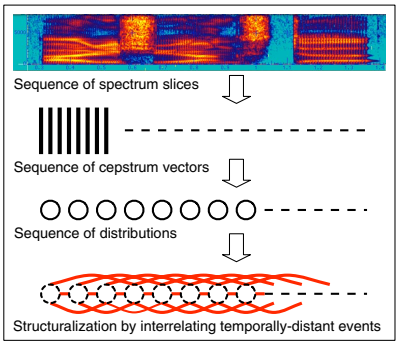

Figure 4: BD-based robustly-invariant structure of speech

From a spoken utterance, it is possible to extract its invariant structure, shown in Figure 4. After converting the utterance into a sequence of distributions, all the phonic contrasts between any two distributions are calculated. Here, long-distance contrasts are also considered. As BD is interpreted mathematically as correlation between two distributions, a BD-based distance matrix represents a full set of interrelations between any two of the acoustic events, including temporally-distant ones. If speech dynamics are represented in an $m+1$ dimensional phase space, as in Figure 2, its invariant structure is obtained after converting the speech trajectory into a sequence of distributions and projecting the distributions onto the m dimensional phase space (F_1/F_2 plane in Figure 2). Then, a BD-based structure can be formed from the projected distributions. In previous studies, speech dynamics or trajectories were often characterized as a series of local dynamic features such as delta cepstrum (See Figure 2 and suppose that the space is a cepstrum-based space).

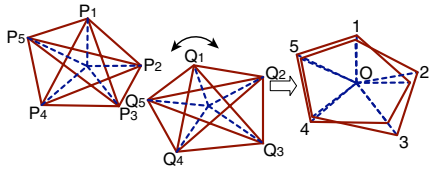

Figure 5: Structural matching through shift and rotation

Table 1: Acoustic conditions for the analysis

sampling	16bit / 16kHz
window	25 ms length and 4 ms shift
parameters	mel cepstrum (1 to 12) + Δ (1 to 12)
distribution	1-mixture Gaussian with a diagonal matrix

The fact that any transformation cannot change the structure of speech dynamics indicates that any transformation works geometrically as either of the two operations, rotation and shift. In the case of $c'=Ac+b$, A and b are rotation and shift, respectively. This mathematically claims that the direction of speech dynamics has to depend on the vocal tract length, i.e., speaker's age [1]. This is why we capture the speech dynamics only based on their scalar quantities which are robustly invariant.

3. Structure-based speech recognition

3.1. Task of the recognition experiments

In order to compare with the results obtained in our previous study [3], a continuous version of the task adopted in that study was used here. The task is recognizing connected vowels and the number of vowels in an utterance is 5; V_1-V_2-V_3-V_4-V_5, where $V_i \neq V_j$. Since Japanese has five vowels, PP is ${}_5P_5$ (120).

3.2. Framework of the structure-based recognition

Once an utterance is represented as structure, it will be matched with reference structures stored in a template database. Figure 5 shows acoustic matching between two structures P and Q. One of the two structures is shifted and rotated so that the two can be overlapped the best. Then, the structure-based distance is calculated as the minimum of the total distance between the corresponding two points after the two geometrical operations. In [1], it was shown that the minimum distance D can be approximately obtained as euclidean distance between the two distance matrices, where the upper-triangle elements form a vector;

$$D(P,Q) = \sqrt{\frac{1}{n}\sum_{i<j}(p_{ij}-q_{ij})^2}. \quad (1)$$

p_{ij} is a (i,j) element of P and n is the number of distributions.

The overall framework is shown in Figure 6. The left side shows the procedure to extract the structure from an input utterance. To convert a frame sequence to a distribution sequence, the MAP-based training of HMMs was adopted because all the distributions had to be estimated from only a single utterance. After that, a distance matrix was obtained from the distributions, and the upper-triangle elements were used as a feature vector. The acoustic conditions used are shown in Table 1.

The right side is a reference template database. Here, each of the 120 words was modeled as structure statistically, which was a multivariate Gaussian distribution. Distance between an input utterance (an upper-triangle vector) and a template (a multivariate distribution) was calculated as Mahalanobis distance.

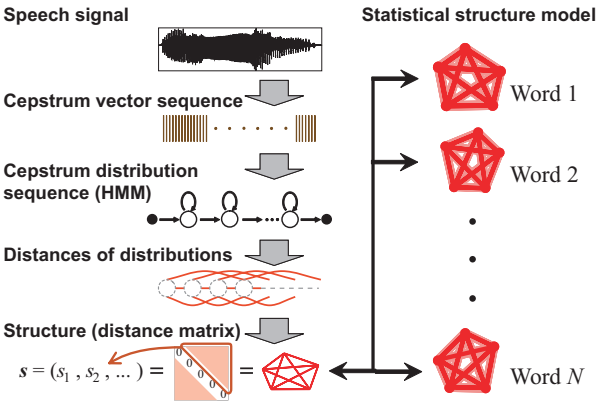

Figure 6: Framework of the structural recognition

3.3. A problem of time-alignment between utterances

Switching from isolated vowels to connected vowels, we had a big problem, which was time-alignment between utterances. It is assumed in Equation 1 that two distribution sequences have good correspondence or alignment between them. In recognizing isolated vowel sequences, since they had a clear boundary between consecutive vowels, the good correspondence was easily guaranteed. But it is not in the current task and the number of distributions may be different between two utterances.

To solve this problem, DTW between two distribution sequences was examined as preprocessing. Here, DTW was implemented in two different ways. One was using acoustic substances (cepstrum distributions), and the other was not. Even in the first case, they were used only for this time-alignment and the structural matching was done only with the two matrices after equalizing the number of distributions using the alignment.

With the substances, the local distance measure between P_t (t-th distribution in sequence P) and Q_s was defined as

$$d_{sub}(P_t,Q_s) = \sqrt{BD(P_t,Q_s)}. \quad (2)$$

The other measure was obtained without substances as

$$d_{str}(P_t,Q_s) = \sum_{m=1}^{M} |p_{tm}-q_{sm}|\,. \quad (3)$$

This is regarded as total difference of interrelations of two given events to each of all the events. M is the length of the sequence.

3.4. Another problem of too strong invariance

With any mapping function, the structural invariance is satisfied. This strong invariance will probably cause a critical problem, where a word and another linguistically different word will be observed as identical. This should decrease the recognition performance easily. Some constraints should be introduced to restrict allowable geometrical transformations and we considered that articulatory constraints should be used. However, we did not have good knowledge on relating possible articulatory variations to allowable geometrical operations. Therefore in this paper, purely geometrical constraints were examined.

We focused on rotation, any of which always satisfies the structural invariance. However, if a structure is projected into a sub-space, the projected structure will change by transformation. By hypothesizing that the structural invariance is also satisfied in sub-spaces, geometrically speaking, the allowable operations are restricted. This hypothesis is easily introduced into the structural matching procedure by considering a cepstrum

stream as multiple independent sub-streams. The total distance between two structures was calculated by accumulating structural sub-distances obtained in the individual sub-spaces. To verify this hypothesis, a parameter vector had to be divided adequately into sub-vectors. In this paper, uniform division was tentatively examined. All the elements of a vector were divided into a group of sub-vectors of the same number of dimensions.

3.5. Experimental set-up

8 male and 8 female adult speakers joined the recording and five utterances were recorded for each of the 120 words. The total number of utterances was 9,600. The samples from 4 males and 4 females were used for training and the others for testing. In the previous study, only a single speaker was used for training [3]. In this work, however, as the required number of utterances was so large, multiple speakers were used for training.

In the case of using DTW, the statistical templates were trained also using DTW as preprocessing. When an input structure was matched after DTW with substances, the templates also needed acoustic substances. However, the structural matching was done only with the two matrices. The number of distributions in an input and that in the individual templates were equalized to be 10, 15, 20, 25, and 30. A speech stream was treated as two separate streams of cepstrum and its Δ, meaning that two kinds of structures were always considered here.

The parameter division was further carried out to reduce the invariance, where the numbers of division were 1 (no division), 2, 3, 4, 6, and 12 for each of the two streams. In this experiment, the number of distributions was fixed to 25 and DTW was *not* done, meaning that no acoustic substances were used at all.

For comparison, two sets of speaker-independent HMMs were tested for the same task; 260-speaker tied-state HMMs and 4,130-speaker tied-mixture HMMs [6], both of which were trained with MFCC and CMN. CFG allowing only the testing 120 words was used as language model.

3.6. Results and discussions

Figure 7 shows the results without the parameter division. The best performance of the proposed method without DTW was 66.6% in 25 distributions, 73.4% with non-substance DTW and 92.6% with substance DTW in 30 distributions. Basically, the larger the number of distributions is, the better the performance is. In our previous study [3], one distribution was sufficient for an isolated vowel but, for connected vowels, more distributions are needed naturally. Comparing the recognition performance with/without DTW, it is obvious that applying DTW improves the performance and that the improvement with substances is much larger. These results clearly indicate that DTW resolved the alignment mismatch and that the resolution owed much to acoustic substances. Figure 7 also shows the results of the HMMs; 82.1% and 97.4% for 260- and 4,130-speaker HMMs, respectively. Considering these results, we have to admit that the speech recognition only with speech dynamics seems hopeless. Conversely speaking, this may imply the too strong invariance induced by abstraction.

With some constraints, however, completely different results were obtained and they are shown in Figure 8 as function of the number of division. It should be noted that all the experiments were done without DTW. With a larger number of division, the better performance was obtained and, in the current experiments, the best performance was 92.6%. If some optimization is done for the division, the performance should improve. Why so large improvement in comparison with the

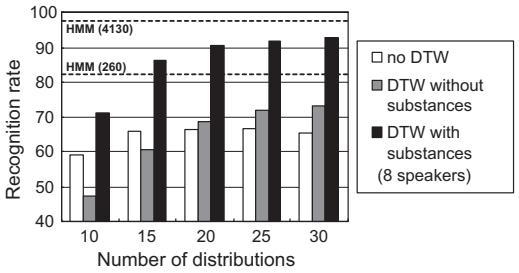

Figure 7: Recognition performance with/without DTW

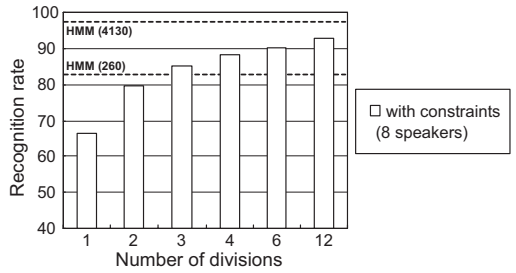

Figure 8: Recognition performance with constraints

results in Figure 7? Since distributions are estimated based on the algorithm of training HMMs, the alignment mismatch in distributions is smaller than that in frames. Although Figure 7 claims that the mismatch in distributions is not small enough, we consider firstly that this mismatch can be cancelled to some degree by dealing with speech as multiple independent streams. Further, as in the above discussion, the multiple stream strategy also induces some appropriate transformational constraints.

4. Conclusions

This paper showed the initial results of applying the speaker-invariant representation of speech dynamics to recognizing continuous speech. The MAP-based HMM training algorithm was used to structuralize an input utterance. Using some geometrical constrains realized as the multiple stream strategy, the proposed method only with 8 training speakers outperformed 260-speaker HMMs and showed the rather comparable performance to 4,130-speaker HMMs. Since the proposed method only extracts speech dynamics, it cannot identify any separate sounds. This is directly opposite to the conventional methods because they were based on identifying individual frames or sounds, and that after collecting an enormous amount of data. This strategic difference can be interpreted as holism vs. reductionism [4].

5. References

[1] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *Proc. ICASSP*, pp.889–892, 2005.

[2] N. Minematsu, *et al.*, "Theorem of the invariant structure and its derivation of speech Gestalt," *Proc. SRIV*, pp.47–52, 2006.

[3] T. Murakami, *et al.*, "Japanese vowel recognition using external structure of speech," *Proc. ASRU*, pp.203–208, 2005.

[4] N. Minematsu *et al.*, "Linear and non-linear transformation invariant representation of information and its use for acoustic modeling of speech," Proc. Spring Meeting Acoust. Soc. Jpn., pp.147–148, 2007.

[5] M. Pitz, *et al.*, "Vocal tract normalization equals linear transformation incepstral space," *IEEE Trans. Speech and Audio Processing*, 13, 5, pp.930–944, 2005.

[6] T. Kawahara, *et al.*, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," *Proc. ICSLP*, pp.3069–3072, 2004.