2012 International Workshop on Nonlinear Circuits,
Communications and Signal Processing
NCSP'12, Honolulu, Hawaii, USA. March 4-6, 2012

An Experimental Study on Dynamic Features of Speech Structure

Shinya Shimizu, Masayuki Suzuki, Nobuaki Minematsu, Keikichi Hirose

Graduate School of Information Science and Technology,
The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
E-mail: {s_shimizu, suzuki, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

One of the biggest difficulties in automatic speech recognition (ASR) is how to deal with variations of speech signals caused by non-linguistic information, such as age, gender, etc. Various methods have been proposed to compensate for the variations and one of them is speech structure [1]. Speech structure, which extracts only contrastive features and discards absolute features, is proved to be transform-invariant mathematically and to be very robust with the non-linguistic variations experimentally [2]. Although the conventional speech structure extracts local and distant contrastive features, it did not extract dynamic features explicitly which are supposed to exist in the contrastive features. In this paper, we reformulate speech structure based on trajectory HMM and derive trajectory structure (TSR), in which dynamic and contrastive features can be defined and used in ASR. We carry out an experiment of n-best rescoring of isolated word recognition using trajectory structure and obtain 28.5% relative decrease in word error rate.

1. Introduction

Speech signals contain various kinds of information, such as linguistic messages, speaking styles, speaker identity, recording conditions, etc. When one tries to get some specific kinds of information from speech signals, one wants to extract the acoustic features that represent only the target information and are independent of the other kinds of information. ASR systems, which convert speech signals to texts, need the acoustic features that convey linguistic information only. However, mel-cepstrum-based features, which are most commonly used, are not independent at all of non-linguistic information. Therefore, researchers have developed various methods to compensate for non-linguistic variation in speech features. These methods are, for example, feature normalization, noise suppression, speech enhancement, and model adaptation. However these methods are reported to be ineffective in some applications, such as children's speech recognition [2].

To solve the problem, a method was proposed [1] to extract the acoustic features that are mathematically independent of the non-linguistic variations. The proposed representation is called speech structure. In the proposed representation, first, the speech feature sequence is converted to a sequence of distributions, from each pair of which a distance is calculated using f-divergence . The obtained distance matrix is adopted as a speech representation of the input utterance. f-divergence is mathematically proved to be invariant with any continuous and differentiable transformation, as which any non-linguistic speech variation can be characterized. These facts indicate that the f-divergence distance matrix can be regarded as invariant representation with non-linguistic variations. Speech structure has been applied to several applications and showed good results especially for pronunciation assessment where children's speech sometimes has to be compared to adult teachers' speech [3].

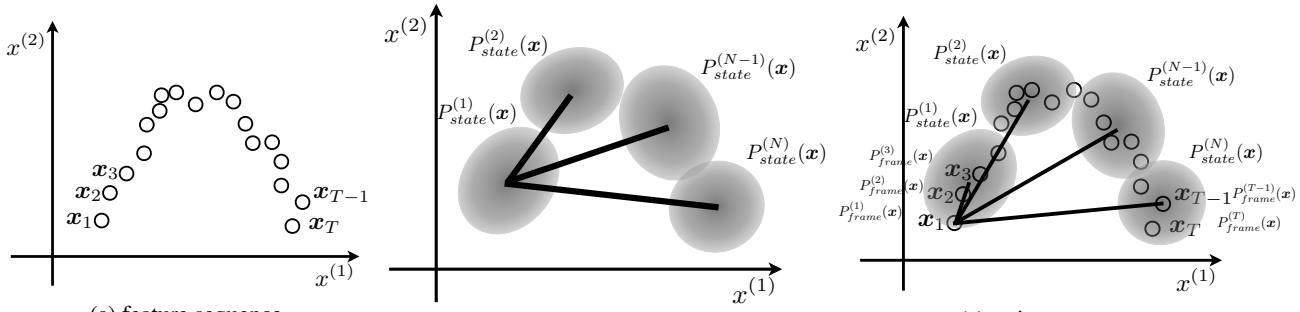
However, the computational implementation of the speech structure is still immature and can be sophisticated in some aspects. Since the speech structure is a f-divergence distance matrix among the distributions, temporal dynamics, which may be actually observed in a single distribution, has to be ignored completely. In this paper, we firstly derive a speech structure not based on the classical HMM but based on the trajectory HMM. Using the trajectory HMM, we can define a distance vector at each time. Next, we derive dynamic and contrastive features using first and second derivatives of distance vectors. We carry out an experiment of isolated word recognition and obtained 28.5% relative reduction in word error rate by using n-best rescoring based on trajectory structures.

2. Speech Structure Model

In speech science and technology, phonemic identity is often characterized as spectrum envelope and it is represented as a point in a cepstrum space. However, non-linguistic factors can change the coordinate values of the point easily. On the contrary, in a speech structure, only distances (contrasts) between two distributions, which often refer to phonemes, are calculated and absolute features are discarded instead. We use Bhattacharyya distance (BD), which is one kind of f-divergence , because it was found to work well by previous studies.

$$BD(p_i, p_j) = -\log \int \sqrt{p_i(\boldsymbol{x}) p_j(\boldsymbol{x})} d\boldsymbol{x} \qquad (1)$$

In ASR using speech structure, firstly we train an HMM with N states from an input utterance and obtain N output distributions, and secondly we calculate BDs between each pair,



(a) feature sequence (b) speech structure (c) trajectory structure

Figure 1: A feature sequence, its speech structure, and its trajectory structure

and finally adopt them as the features of the utterance.

3. Trajectory Structure Model

While in the conventional speech structure we trained an HMM with N states and obtained $\binom{N}{2}$ BDes between each pair of the N state distributions, in the trajectory structure model, we assume that each time frame has its own unique distribution. In other words, we can have a coarsely quantized distribution sequence from a classical HMM and a finely quantized sequence from a trajectory HMM. Using these two distribution sequences, we can calculate a BD between a fine distribution and a coarse distribution. Fig.1 shows an example of feature sequence, its speech structure, and its trajectory structure. In the standard approach, a feature sequence $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]$ itself is often used as a representation of the input utterance (Fig.1(a)). In the conventional speech structure, we train an HMM with N states, calculate BDes from every pair of the states, and obtain a sequence of distance vector $[\boldsymbol{s}_1, \boldsymbol{s}_1, \cdots, \boldsymbol{s}_N]$. $\boldsymbol{s}_n$ is a distance vector of the n-th state,

$$\begin{aligned}\boldsymbol{s}_n &= \Big[BD(P^{(n)}_{state}, P^{(1)}_{state}),\ BD(P^{(n)}_{state}, P^{(2)}_{state}),\\ &\quad \cdots, BD(P^{(n)}_{state}, P^{(N)}_{state})\Big]^{\mathrm{T}}. \end{aligned} \tag{2}$$

Where $P^{(n)}_{state}$ is the output distribution of the n-th state. Because Bhattacharyya distance is symmetric, we just pick up $\binom{N}{2}$ distances out of the N^2 distances. These $\binom{N}{2}$ distances are used as a representation of the input utterance (Fig.1(b)).

In trajectory structure model, after we train a classical HMM with N states, it is used to derive frame-dependent distributions. Then, we obtain a sequence of distributions $\left[P^{(1)}_{frame}(\boldsymbol{x}), P^{(2)}_{frame}(\boldsymbol{x}), \cdots, P^{(T)}_{frame}(\boldsymbol{x})\right]$, where T is the total number of frames and $P^{(t)}_{frame}(\boldsymbol{x})$ is the distribution of the t-th frame. Each $P^{(t)}_{frame}(\boldsymbol{x})$ is calculated from trajectory HMM [4], which derives the temporally changing distributions of static features by imposing the explicit relationship between static features and dynamic features. The detailed procedure to obtain $P^{(t)}_{frame}(\boldsymbol{x})$ is described later. By using T fine distributions and N coarse distributions, at time t, we calculate a distance vector whose i-th element is BD between the t-th fine distribution and the i-th coarse distribution. Since this vector can be obtained at each time, we have T distance vectors with their dimension being N. (Fig.1(c)). The distance vector at time t, $\boldsymbol{d}_t$, is given by

$$\begin{aligned}\boldsymbol{d}_t &= \Big[BD(P^{(t)}_{frame}, P^{(1)}_{state}),\ BD(P^{(t)}_{frame}, P^{(2)}_{state}),\\ &\quad \cdots, BD(P^{(t)}_{frame}, P^{(N)}_{state})\Big]^{\mathrm{T}}. \end{aligned} \tag{3}$$

Here, a sequence of the distance vectors $[\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_T]$ is used as a representation of the input utterance. In addition, since the distance vector is obtained at each time, we can calculate its Δ and Δ^2 features and concatenate them and $\boldsymbol{d}_t$.

So far, we have introduced the basic procedure to derive TSR model. In the rest of this section, we introduce the detailed procedure to derive $P^{(t)}_{frame}(\boldsymbol{x})$ based on trajectory HMM. Let $\boldsymbol{x}$ denote a concatenation of an input feature sequence $\left[\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \cdots, \boldsymbol{x}_T^{\mathrm{T}}\right]^{\mathrm{T}}$, each $\boldsymbol{x}_t$ is a concatenation of a static feature vector and its Δ and Δ^2 features, $\left[\boldsymbol{c}_t^{\mathrm{T}}, \Delta\boldsymbol{c}_t^{\mathrm{T}}, \Delta\boldsymbol{c}_t^{\mathrm{T}}\right]^{\mathrm{T}}$. Let M denote the dimension of $\boldsymbol{c}_t$. Δ and Δ^2 features are defined as weighted sums of adjacent static feature vector as follows.

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L}^{L} w^{(1)}(\tau)\boldsymbol{c}_{t+\tau}, \tag{4}$$

$$\Delta^2\boldsymbol{c}_t = \sum_{\tau=-L}^{L} w^{(2)}(\tau)\boldsymbol{c}_{t+\tau}, \tag{5}$$

where L is the length of window to calculate dynamic features, and $w^{(1)}(\tau)$ and $w^{(2)}(\tau)$ are coefficients for $\boldsymbol{c}_{t+\tau}$. Because each $\boldsymbol{x}_t$ is calculated by an linear transformation of $\boldsymbol{c}_t$, there exists a matrix W such that

$$\boldsymbol{x} = W\boldsymbol{c}, \tag{6}$$

where $\boldsymbol{c}$ is a concatenation of a static feature sequence $\left[\boldsymbol{c}_1^{\mathrm{T}}, \boldsymbol{c}_2^{\mathrm{T}}, \cdots, \boldsymbol{c}_T^{\mathrm{T}}\right]^{\mathrm{T}}$. Because the dimension of $\boldsymbol{c}$ is MT and that of $\boldsymbol{x}$ is $3MT$, W is a $3MT \times MT$ matrix. When we assume this feature sequence is generated by an HMM that has N states each of which has a single Gaussian as output distribution, the probability of $\boldsymbol{x}$ given alignment $\boldsymbol{q}$ and HMM

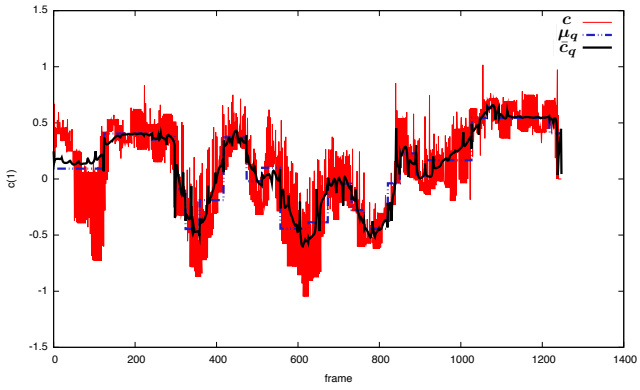


Figure 2: Observations of 1st-order mel-cepstrum, means of the original HMM, and means of the frame-dependent HMM.

parameter λ, $P(\boldsymbol{x}|\boldsymbol{q},\lambda)$ is calculated as follows.

$$P(\boldsymbol{x}|\ \boldsymbol{q},\lambda) = \prod_{t=1}^{T}\mathcal{N}(\boldsymbol{x}_t|\ \boldsymbol{\mu}_{q_t},\Sigma_{q_t}) \quad (7)$$

$$= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_q},\Sigma_{\boldsymbol{q}}) \quad (8)$$

where q_t is the index of the state to which $\boldsymbol{c}_t$ belongs. $\boldsymbol{\mu_q}$ and $\Sigma_{\boldsymbol{q}}$ are concatenations of sequences $\boldsymbol{\mu}_{q_t}$ and Σ_{q_t} respectively.

$$\boldsymbol{\mu_q} = \left[\boldsymbol{\mu}_{q_1}^{\mathrm{T}},\boldsymbol{\mu}_{q_2}^{\mathrm{T}},\cdots,\boldsymbol{\mu}_{q_T}^{\mathrm{T}}\right]^{\mathrm{T}} \quad (9)$$

$$\Sigma_{\boldsymbol{q}} = \mathrm{diag}\left[\Sigma_{q_1},\Sigma_{q_2},\cdots,\Sigma_{q_T}\right] \quad (10)$$

Because $\boldsymbol{x}$ satisfies Eq.6, there exist a mean vector $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$ and a covariance matrix $\boldsymbol{P_q}$ such that

$$P(\boldsymbol{x}|\ \boldsymbol{q},\lambda) = \mathcal{N}(\boldsymbol{W}\boldsymbol{c}|\ \boldsymbol{\mu_q},\Sigma_{\boldsymbol{q}}) \quad (11)$$

$$= K_{\boldsymbol{q}}\,\mathcal{N}(\boldsymbol{c}|\ \bar{\boldsymbol{c}}_{\boldsymbol{q}},\boldsymbol{P_q}) \quad (12)$$

where $K_{\boldsymbol{q}}$ is a constant that depends on alignment $\boldsymbol{q}$. An actual example of $\boldsymbol{c}$, $\boldsymbol{\mu_q}$, and $\bar{\boldsymbol{c}}_{\boldsymbol{q}}$ is shown in Fig.2. The covariance matrix $\boldsymbol{P_q}$ is not a diagonal matrix, but still a band matrix, in which only diagonal elements and their adjacent elements are non-zero. To obtain an independent distribution for each frame, we approximate $\boldsymbol{P_q}$ as a block diagonal matrix as follows

$$\boldsymbol{P_q} \approx \mathrm{diag}\left[\boldsymbol{p}_{\boldsymbol{q}}^{(1)},\boldsymbol{p}_{\boldsymbol{q}}^{(2)},\cdots,\boldsymbol{p}_{\boldsymbol{q}}^{(T)}\right]. \quad (13)$$

Finally, we obtain an independent distribution for each frame.

$$\mathcal{N}(\boldsymbol{c}|\ \bar{\boldsymbol{c}}_{\boldsymbol{q}},\boldsymbol{P_q}) \approx \prod_{t=1}^{T}\mathcal{N}(\boldsymbol{c}_t|\ \bar{\boldsymbol{c}}_{\boldsymbol{q}}^{(t)},\boldsymbol{p}_{\boldsymbol{q}}^{(t)}), \quad (14)$$

where $\bar{\boldsymbol{c}}_{\boldsymbol{q}}^{(t)}$ is a vector with its dimension of M that corresponds to the t-th time frame. From the above equation, we can define frame-dependent distribution $P_{frame}^{(t)}(\boldsymbol{c}_t)$,

$$P_{frame}^{(t)}(\boldsymbol{c}_t) = \mathcal{N}(\boldsymbol{c}_t|\ \bar{\boldsymbol{c}}_{\boldsymbol{q}}^{(t)},\boldsymbol{p}_{\boldsymbol{q}}^{(t)}). \quad (15)$$

4. N-best Rescoring Based on TSR Model

4.1. Procedure of N-best Rescoring

N-best rescoring based on speech structure is proposed in [5]. We carried out an experiment of N-best rescoring based on TSR models. The task is isolated word recognition. Let $P_{hmm}(\boldsymbol{x}|w_i)$ denote output probability of $\boldsymbol{x}$ given word w_i's HMM, and $P_{tsr}(\boldsymbol{c}|w_i)$ denote that of $\boldsymbol{c}$ given word w_i's TSR model. The rescored output probability $P_{res}(\boldsymbol{x}|\ w_i)$ is given by:

$$P_{res}(\boldsymbol{x}|w_i) = P_{hmm}(\boldsymbol{x}|w_i)P_{tsr}(\boldsymbol{c}|w_i)^{w_{tsr}} \quad (16)$$

where w_{tsr} is the weight of TSR model. The procedure to calculate TSR likelihood $P_{tsr}(\boldsymbol{c}|w_i)$ is shown in Fig. 3. To calculate $P_{tsr}(\boldsymbol{c}|w_i)$, a classical HMM trained for the input utterance is needed. For that, we first trained a speaker-independent HMM for each word, which is used as initial model. The parameters of this initial HMM are updated only by the input utterance and the resulting HMM is used to derive the TSR model. If we do not use speaker-independent word HMMs as initial and background models, the resulting HMM of an utterance and that of another utterance will show different alignment patterns between states and feature vectors even when the two utterances are of the same word. Because the feature vector in speech structures is composed of distances between HMM states, it is essential to satisfy a condition that state i in an HMM and state i in another HMM keep the same linguistic function when these two HMMs correspond to the same word. Using an utterance-specific but temporally aligned HMMs, we obtain a TSR vector sequence.

How to model statistically the TSR vector sequences? We can use some commonly-used sequential models like HMM, where plural alignment paths are allowed between a feature sequence and the state sequence of the HMM. In our case, however, because alignment between the TSR vector sequences and the retrained HMM is already determined, TSR vectors of a state of the HMM are modeled as Gaussian distribution. Finally the likelihood $P_{tsr}(\boldsymbol{c}|\ w_i)$ is given as

$$P_{tsr}(\boldsymbol{c}|\ w_i) = \prod_{t=1}^{T}\mathcal{N}(\boldsymbol{d}_t|\ \boldsymbol{\mu}_{w_i}^{(q_t^*)},\boldsymbol{\Sigma}_{w_i}^{(q_t^*)}), \quad (17)$$

where $\boldsymbol{d}_t$ is the t-th TSR vector, $\boldsymbol{\mu}_{w_i}^{(n)}$ and $\boldsymbol{\Sigma}_{w_i}^{(n)}$ are the mean vector and covariance matrix for the n-th state in w_i, and q_t^* is the state index to which the t-th frame belongs. We adopt the Viterbi path $\boldsymbol{q}^*$ instead of considering all the paths. It is possible to consider all the paths but it is very costly.

4.2. Experimental Conditions

We used Tohoku University and Panasonic isolated spoken word database [6], which contains 212 kinds of japanese words spoken by 60 speakers. The word length varies from 3 morae to 7 morae. We used the utterances by 30 speakers as a training data set and ones by the other 30 speakers as an evaluation data set. In the training, the parameters of TSR model

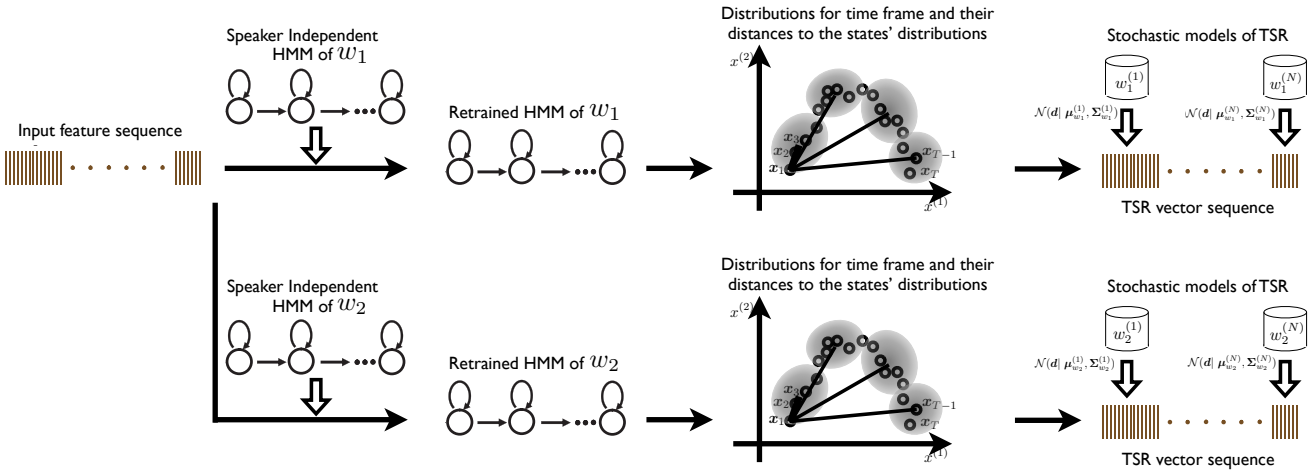


Figure 3: Procedure to calculate TSR likelihood for each candidate

Table 1: Conditions

Sampling	16bit/16kHz
Window	25ms Blackman Window & 1ms Shift
Feature	mel-cepstrum 18 dim., its Δ and Δ^2
Delta Window Length	20 frames
Word HMM	8-mixture with diagonal matrix
	25 states and 23 distributions

Table 2: Rescoring by TSR decreases the word error rate

Scoring method	Word error rate
HMM	1.37%
HMM + TSR	1.05%
HMM + TSR + Δ TSR + Δ^2 TSR	0.98%
N-best Oracle	0.04 %

$\boldsymbol{\mu}_{w_i}^{(n)}$ and $\boldsymbol{\Sigma}_{w_i}^{(n)}$ are estimated to maximize Eq.17. We set TSR weight w_{tsr} in Eq.16 as 1.0×10^{-11} by preliminary experiments. For rescoring, 10-best words are used as candidates. Other conditions are shown in Table 1.

4.3. Results

We compared three methods, HMM only, HMM rescored by TSR, and HMM rescored by TSR with its Δ and Δ^2. The results are shown in Table.2. As shown in the table, rescoring by TSR decrease the word error rate by 23.3% relative and rescoring by TSR with its Δ and Δ^2 decrease the word error rate by 28.5% relative. The results show that TSR works effectively in the isolated word recognition.

5. Conclusions

Due to coarse quantization in time, fine and dynamic features are not well modeled in the conventional implementation of speech structures. In this paper, we reformulated speech structure using trajectory HMMs and we successfully derived temporally-fine speech structure. Using the new speech structure, we introduced dynamic features of the speech structure. We carried out an experiment of N-best rescoring of isolated word recognition and obtained 28.5% decrease relative in word error rate.

References

[1] N. Minematsu, "Yet another acoustic representation of speech sounds," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–585.

[2] N. Minematsu, S. Asakawa, M. Suzuki, and Y. Qiao, "Speech structure and its application to robust speech processing," *New Generation Computing*, vol. 28, no. 3, pp. 299–319, 2010.

[3] M. Suzuki, N. Minematsu, Dean Luo, and K. Hirose, "Sub-structure-based estimation of pronunciation proficiency and classification of learners," in *Automatic Speech Recognition Understanding, 2009. ASRU 2009. IEEE Workshop on*, 13 2009-dec. 17 2009, pp. 574 –579.

[4] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the hmm as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173, 2007.

[5] M. Suzuki, G. Kurata, M. Nishimura, and N. Minematsu, "Continuous digits recognition leveraging invariant structure," *Proc. INTERSPEECH*, pp. 993–996, 2011-8.

[6] Shozo Makino, Niyada Katsuyuki, Mafune Yasuo, and Kido Kin'iti, "Tohoku university and panasonic isolated spoken word database," *Acoustical Science and Technology*, vol. 48, no. 12, pp. 899–905, 1992-12-01.