Reliable Methods for Estimating Relative Vocal Tract Lengths From Formant Trajectories of Common Words

Akira Watanabe and Tadashi Sakata

Abstract—**This paper describes reliable methods for estimating relative vocal tract lengths from speech signals. Two proposed methods are based on the simple principle that resonant frequencies in an acoustic tube are inversely proportional to the tube length in cases where the configuration is constant. We estimated the ratio between two speakers' vocal tract lengths using first and second formant trajectories of the same words uttered by them. In the first method, which is referred to as "strict estimation method," we sought instances at which the gross structures of two vocal tracts are analogous by applying dynamic time-warping to formant-trajectories of common words that were uttered at different speeds. In those instances, which were found from among more than 100 common words by two speakers, an average formant ratio proved to be an excellent estimate (about ±0.1% in errors) for a reciprocal of the vocal tract length ratio. Next, we examined a simplified method for estimating those ratios using all corresponding points of two formant-trajectories: it is the "direct estimation method." Estimation errors in the direct estimation were evaluated to be about ±0.3% at equal utterance-speeds and ±2% at most, within 2.0 of the ratios of "fast" to "slow." Finally, we estimated relative vocal tract lengths for four Japanese speaker groups, whose members differed in terms of age and gender. Experimental results showed that the average vocal tract length of adult females and that of 7–10-year-old boys and girls are 21%, 27%, and 30%, respectively, shorter than adult males'.**

Index Terms—**Acoustic tube, dynamic time warping, formant frequency, vocal tract length.**

I. Introduction

REPLACING voice characteristics that are peculiar to a speaker with those of another speaker is a useful technique for speaker adaptation or normalization in speech recognition and for voice conversion between two speakers in speech synthesis. Resonance differences caused by various vocal tract lengths directly influences those voice characteristics. For that reason, knowing the vocal tract length (VTL) of a speaker gives important clues to assess that speaker's voice characteristics. Many studies have applied VTL normalization to speech recognition. Methods in those studies are classifiable into two groups: those doing recognition after estimating factors to normalize influence of VTL on formants [1], [2], and those seeking normalizing factors for VTL in a model to maximize likelihood [3], [4]. However, most of them aim to find a factor increasing the recognition rates rather than accurately estimate relative VTL.

Recent development of magnetic resonance imaging (MRI) technique has allowed the direct observation of fine vocal tract configurations [5]–[8]. Irrespective of technique, it is necessary to determine a standard point in each utterance for comparing vocal tract lengths because they change during the utterance. However, it is usually impossible to record speech sounds together with an image because of machine noise in MRI measurement [9]. Thereby we infer that obscurity in judgment of the standard point is unavoidable in MRI measurement. For this reason, measuring many vocal tract lengths under a strictly unified condition will be extremely laborious and difficult with MRI. On the other hand, for many applications such as speech recognition or synthesis, we usually do not require absolute vocal tract lengths, but rather relative ones. For that reason, we propose simple and reliable methods to obtain relative vocal tract lengths from speech signals based on the pharyngeal-oral tract model [10], which generates formants. The proposed methods are much more readily applicable to actual speech processing than the MRI technique because they require only speech signals.

Estimation of vocal tract length itself from speech signals was undertaken to normalize formant frequencies of vowels in the previous study by Wakita [11]. For that estimation, the vocal tract area function was first estimated from an utterance of a vowel using linear predictive analysis [12]. Subsequently, the VTL was estimated using an iterative method to minimize a certain error criterion based on the area function. The method was used to estimate absolute vocal tract lengths, which differ for each vowel.

In this paper, we propose to simply estimate a ratio between two speakers' vocal tract lengths using their natural utterances of the same words. Although the ratio is estimated from formant trajectories of word-utterances, differences of articulatory movements in individual words almost never influence the estimates. The formant estimation method using inverse filter control (IFC) [13] has made it possible to obtain reliable formant trajectories from continuous speech uttered by various speakers. Reliable formant trajectories are indispensable for estimating relative vocal tract lengths using the proposed methods. This paper describes the methods, reliability of estimates, the estimated distributions of relative vocal tract lengths, and practical utility of the methods.

Manuscript received August 20, 2004; revised January 1, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Thierry Dutoit.

A. Watanabe is with the Kumamoto Prefectural College of Technology, Kumamoto Prefecture, 869-1102, Japan (e-mail: watanabe@cs.kumamoto-u.ac.jp; watanabe@kumamoto-pct.ac.jp).

T. Sakata is with the Department of Computer Science, Kumamoto University, Kumamoto City, 860-8555, Japan (e-mail: tadashi@cs.kumamoto-u.ac.jp).

Digital Object Identifier 10.1109/TSA.2005.860829

II. Strict Estimation Method and Experimental Results

A. Background

The acoustic theory of a uniform tube, which is a model of the vocal tract uttering a neutral vowel, posits that resonant (formant) frequencies are provided by the following equation [10]:

$$F_n = \frac{(2n-1)c}{4L}. \tag{2.1}$$

In that equation, F_n is the nth formant frequency [Hz], c is sound velocity [m/s], and L is the vocal tract length [m].

Equation (2.1) for uniform tubes states that formant frequencies are inversely proportional to vocal tract length. Therefore, from (2.1), we can derive the following relation between formant frequencies (F_{nA}, F_{nB}) and vocal tract lengths (L_A, L_B) for speakers A and B

$$\frac{L_A}{L_B} = \frac{F_{nB}}{F_{nA}}. \tag{2.2}$$

If this relation holds for two vocal tracts of any analogous configuration, we can assume fundamentally that the ratio of A to B in a specific formant (F_n) is approximately constant in a phoneme that is in the same position in a syllable or a word in a topic sentence that is uttered by two speakers. However, since comprehensive factors affect parameter variations in real speech, we infer that (2.2) is correct only under some limits. In this study, we imposed three restrictions on speech parameters and materials to be used. First, (2.2) is calculated using only the lowest two formant-frequencies (F_1 and F_2). When two speakers utter identical vowels, only gross structures of their vocal tracts are similar in the respective vowels. The gross structures, which are defined by a tongue hump position and height, generate F_1 and F_2 that make it possible to identify the vowels, whereas fine structures of vocal tracts as individual differences affect F_3 or higher formants determining mainly sound qualities. Therefore, if two vocal tracts have analogous gross-structures with each other, their F_1 and F_2 will satisfy (2.2), the relation with their respective vocal tract lengths. Nevertheless, F_3 or higher formants may disturb relation (2.2), even in an identical vowel with analogous gross-structures, because of the fine structural differences in vocal tracts. As an example of the evidence, we mention the following observations: When observing the relation between vocal tracts and acoustical data, upright and supine postures uttering an identical vowel reportedly affect the pharyngeal and laryngeal cavity shapes without changing the VTL and the gross structure, thereby causing shifts only in F_3 or higher formants (or in a spectrum above 1.5 kHz in a male voice) [14]. We use only the lowest two formant-frequencies for the reasons outlined above. Second, we choose "isolated words" as a speech unit and use many varieties of isolated words for estimation. "Isolated vowels" are easily treated for practical processing. However, fixed articulatory biases as individual differences of speakers may occur in isolated vowels that are uttered under strong consciousness to correct articulation. In contrast, when estimating formant ratios from many varieties of isolated words, we expect that various coarticulation effects on vowels in different phoneme environments reduce, as a mean for identical vowels in words, articulatory biases that may appear in utterances of isolated vowels and those in a fixed context. Third, we compare formant trajectories of identical words uttered by two speakers. Because two utterances of an identical word have an almost identical context-dependent coarticulation effect in their respective utterances, it is proper to calculate (2.2) at corresponding points between two sets of formant trajectories of each word. Thus, we consider that three restrictions described above enhance the estimation accuracy from (2.2). To sum up, we estimate the VTL ratios using the lowest two formant-frequencies from many varieties of identical words uttered by two speakers.

B. Procedure and Its Basis

A speaker's vocal tract length and configuration vary naturally with time during the utterance of a word. If two speakers utter a word synchronously in rhythm and intonation, their vocal tracts shorten and lengthen in much the same proportion at the same time; moreover, those configurations are approximately analogous during utterances. Therefore, the ratio between their vocal tract lengths can be estimated using a formant ratio at any point from (2.2). However, if the utterance speed differs between two speakers, vocal tract configurations are influenced by the different extent of the coarticulation effect, which causes distinctions in formant frequencies. Hence, when using (2.2), it is desirable to strictly estimate the vocal tract length (VTL) ratio between two speakers using sets of the lowest two formant-frequencies at the instants their vocal tract configurations are regarded as analogous. In this paper, the term, "analogous configurations" does not mean perfect similarity in vocal tract area functions. Instead it means configurations that are analogous in terms of gross structure, satisfying (2.2) in both of the lowest two formant-frequencies of utterances by A and B, as

$$\frac{F_{1B}}{F_{1A}} = \frac{F_{2B}}{F_{2A}} = \frac{L_A}{L_B} = \mu \text{ (constant)}. \tag{2.3}$$

Based on considerations explained above, we adopt the following procedure to obtain fundamental data common to two proposed estimations: We first estimate reliable formant trajectories in word utterances using inverse filter control (IFC) method [13]. The formant estimation system operates at 12 kHz sampling. The frame length and the frame shift are 20 ms and 10 ms, respectively. Second, after automatically removing unvoiced and silent parts, for which pitch is not detected, each formant trajectory is reconnected and then normalized as 0 and 1 in mean and variance, respectively. Third, we determine corresponding points between the two utterances of each word by applying dynamic time warping (DP matching) to the normalized formant trajectories, in which constraints for initial and final points are relaxed [15]. In this case, we must define a distance between two frames at each grid point on the formant-trajectory diagram to determine a minimum-distance route. The distance is defined as Euclidean distance on a two-dimensional space of the normalized F_1 and F_2 at each grid point. Finally, original formant trajectories by two speakers are relocated along the unified minimum-distance route. In this manner, we obtain modified formant-trajectories that have arranged the corresponding points in two utterances of an identical word.

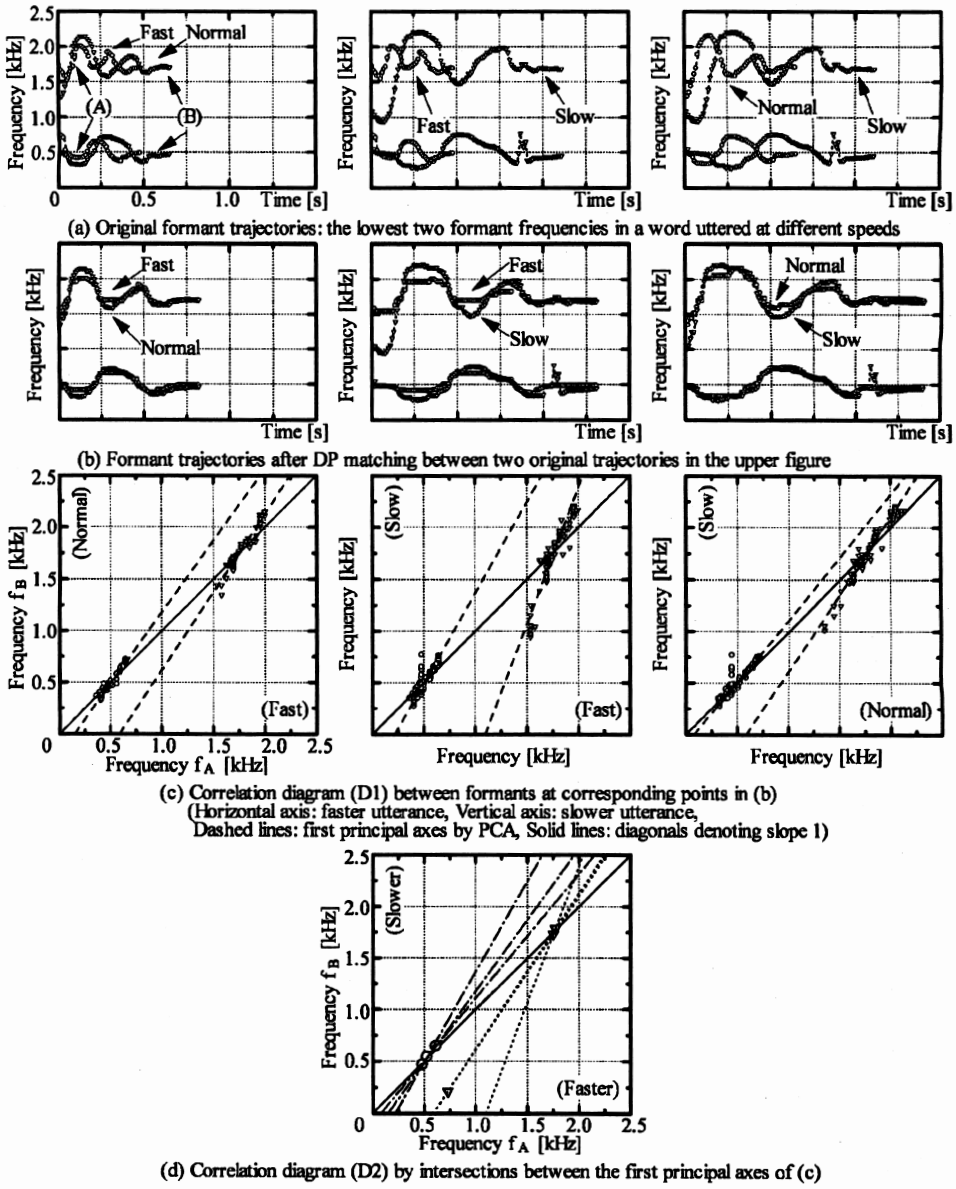


Fig. 1. Explanatory figure of strict estimation principle. (Example from utterances of a word at different speeds by a single speaker).

Next, we explain the strict estimation procedure using an example of formant trajectories uttered by one male speaker. A subject uttered a Japanese word, /toriaezu/([toɾiaezɯ]"for the time being" in English), three times at different speeds: "fast," "normal," and "slow," Fig. 1(a) and (b) shows original and DP-matched formant-trajectories, respectively, which are represented as a set of different utterance speeds like "fast" and "normal." Fig. 1(c) indicates correlation diagrams (D1) of formants for each pair of utterances after DP matching. By applying principal component analysis (PCA) [16] to individual correlation diagrams of F_1 and F_2, we obtain the two first principal axes; they are indicated with two dashed lines in each panel of Fig. 1(c). The first principal axis represents a linear equation that minimizes the sum of squared distances from the data points of F_1 or F_2. (An isolated word forms a block that speakers can use to hold speed constant easily during an utterance. Accordingly, we can estimate first principal axes with certainty for most utterances.) A solid line in each panel of Fig. 1(c) and (d) shows a diagonal whose slope is one, representing a single speaker's VTL ratio. Next, the six first principal axes are determined in three correlation diagrams of Fig. 1(c) and treated together in Fig. 1(d), in which it may look like five lines in total because two lines in F_2 are very close to each other. In this case, if intersections of three lines for individual formants satisfy (2.3) with a small dispersion, they indicate estimates of formant frequencies at instants of identical (or analogous) vocal tract configurations. Thereby, we can obtain an average slope μ (the reciprocal of the VTL ratio) from the intersections. Effectiveness of the proposed method in this example is verified by the fact that most intersections appear to approximate the linear function that passes through the origin with a slope of one, denoting the true slope for a single speaker. From that result, we infer an important con-

jecture: "When seeking first principal components of correlation diagrams (D1) between two utterances at different speeds, intersections among those components indicate formants in a fixed configuration that is unaffected by changes of utterance speed, but that is peculiar to a word. Articulatory movements in faster utterances are neutralized around the fixed configuration." Existence of the fixed configurations is confirmed using many words in Section II-C. Although one intersection of two similar lines in F_2 shown in Fig. 1(d) engenders a large error, we can omit it as a point estimated by extrapolation. That is, we have used only intersections that exist in the range of formant changes.

The above procedure is also justified for utterances of two speakers whose vocal tract lengths (L_A and L_B) differ from one another. If the ratio of their lengths is the constant $L_A/L_B = \mu$, we can produce approximately the same situation as a single speaker's case using correlation between F_{kB} and μF_{kA} (formant number: $k = 1, 2$). Therefore, intersections of individual first principal components are first computed from identical words that two speakers utter at various speeds. Then, intersections of many words comprise a new correlation diagram (D2). Subsequently, seeking the first principal component (a linear equation) that passes through the origin in the correlation diagram (D2), a slope (μ) of the linear equation is exposed as an estimate for an inverse ratio of vocal tract lengths. Reliability of the obtained slope will be guaranteed by a small dispersion of samples, which is defined as a standard deviation of distances from the linear equation to samples. We newly define the analysis to find the first principal component on the condition that it passes through the origin. We denote it as "conditional principal component analysis," or conditional PCA. (See the Appendix for the specific mathematical expression. Using linear regression instead of the conditional PCA is inappropriate because it cannot satisfy the necessary relation, $\mu_1\mu_2 = 1$ in the Appendix.)

C. Strict Estimation Experiments and Results

1) Acquisition Methods for Utterances at Three Different Speeds: We recorded word utterances by the following method to roughly normalize utterance speeds among speakers. Three adult speakers [two males (M001, M002) and one female (F001)] uttered each of 216 Japanese words from the ATR database [17] at three speeds in the order of normal, fast, and slow. In this case, just after hearing each word-utterance at normal speed, as spoken by a specific speaker (adult male) in the database, each speaker was asked to utter the same word at the same speed as the sound that was heard, then successively at both "considerably faster" and "considerably slower" speeds than the first. Total time from the normal speed utterance to the slow one, including silence intervals, is limited to within 6 s in every word. Results for the three speakers showed that average speeds per word in "fast," "normal," and "slow" utterances were 8.3–9.9, 6.1–6.3, and 3.5–4.2 [mora/s] respectively, with standard deviations of 1.6–1.8, 1.1–1.3, and 0.7–1.0 [mora/s].

2) Experimental Results in the Single-Speaker Case: In cases of estimating VTL ratios using single speakers' utterances, we can confirm the proposed method by verifying that the estimated slopes (μ) or VTL ratios are nearly equal to unity. In the explanatory figures (Fig. 1), we explain the strict estimation procedure using faster utterance samples for the horizontal axis (A) and slower ones for the vertical axis (B) in correlation diagrams. However, if all utterances are assigned to each of the two axes in the same way as the above, it may give rise to the estimation errors with a constant tendency. Therefore, it is desirable for the unbiased estimation to append samples in an inverse combination: faster for B and slower for A. To realize this method, we first divided 216 words in the ATR database into two groups of 108 words: odd numbered words (W1) and even numbered words (W2). Next, the procedure in Section II-B was applied to both faster (A) and slower (B) utterances of W1 and slower (A) and faster (B) of W2 to collect intersections between principal axes. Thereby, we obtained correlation diagrams with small dispersions (SD = 40–50 Hz) shown in Fig. 2(a) for all combinations of utterance speeds using word groups W1 and W2. The sufficiently small dispersions of intersections suggest that the method to mix two different assignments for the axes is adequate to estimate the slopes (μ). Because VTL ratios can be estimated as reciprocals of the slopes (μ) in Fig. 2(a), the estimation error rates are 0.04%, −0.04% and 0.10% for speakers M001, M002, and F001, respectively. These error rates indicate excellent results: errors by the strict estimation method are smaller than 0.2 mm if an adult male's VTL is assumed to be 17.5 cm. (We estimate 2 mm per 1%, at most, as a rough absolute-standard for errors.)

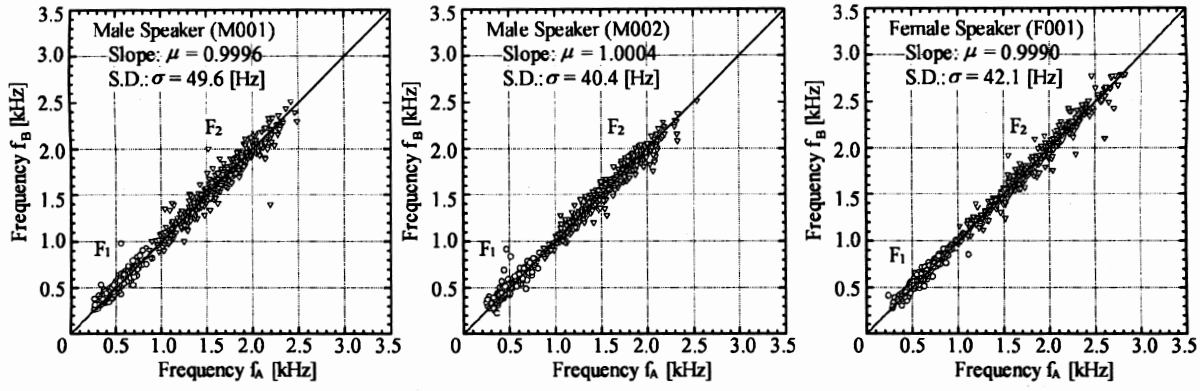
Fig. 2(b) shows convergent processes for each estimate when keeping word-balances in utterances of W1 and W2. Average slopes and confidence intervals at 95% in more than 120 words have been estimated to be 0.9994 ± 0.0008, 1.0004 ± 0.0014, and 0.9990 ± 0.0013 for the three speakers. Variations in the estimations after 120 words are within $\pm 0.14\%$.

3) Two-Speaker Case: Fig. 3(a) shows convergent processes of slope (μ) between two speakers, which we obtained using almost the same procedure as that used for the single-speaker cases. In a two-speaker case, we need not divide 216 words into two groups because two speakers utter all words at three different speeds. Using those utterances, balanced combinations of the utterance speeds can be implemented for two speakers who are fixedly assigned to each of two axes. The figure shows that confidence intervals for slope estimation are very small also in the two speakers' case. Two examples of the correlation diagrams (D2) in Fig. 3(b) show sufficiently small dispersions of samples (SD = 53–61 [Hz]) that guarantee (2.3). Thereby, we obtained vocal tract length ratios between two speakers as $L_{(M001)}/L_{(M002)} = 0.9876$, $L_{(M001)}/L_{(F001)} = 1.1683$, and $L_{(M002)}/L_{(F001)} = 1.1833$.

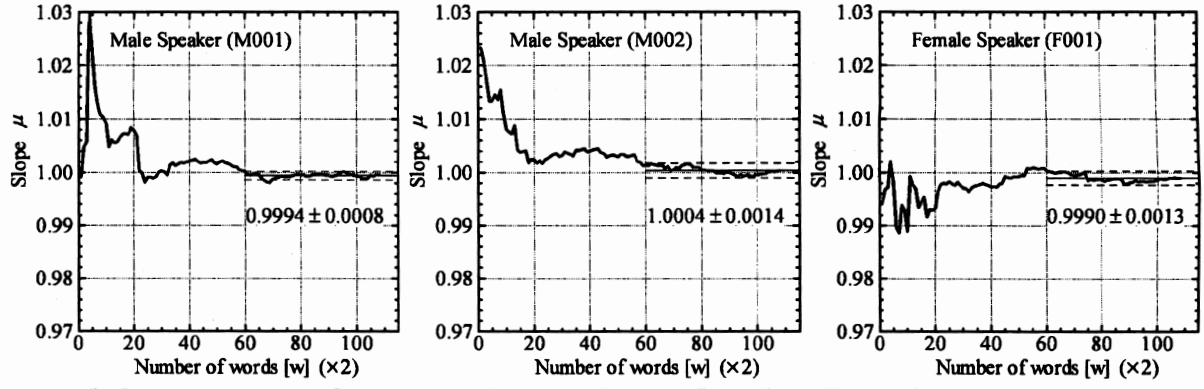
The three estimates were calculated using direct DP matching between all pairs of utterances of 216 common words by two speakers. Despite that, the above estimates support the following relation:

$$\frac{\left(\frac{L_{(M001)}}{L_{(F001)}}\right)}{\left(\frac{L_{(M001)}}{L_{(M002)}}\right)} \cong \frac{L_{(M002)}}{L_{(F001)}}. \tag{2.4}$$

That is, an indirect estimate using an intermediate speaker (M001) on the left of (2.4) is nearly equal to an estimate by di-



(a) Correlation diagrams (D2) and slopes estimated with 216 words



(b) Convergent processes of slopes (Numerals: average slope ± confidence interval at 95% from 120 to 216 words)

Fig. 2. Strict estimation results of average slopes from utterances by single speakers. (True slope: $\mu = 1.0$, VTL ratio: reciprocal of average slope).

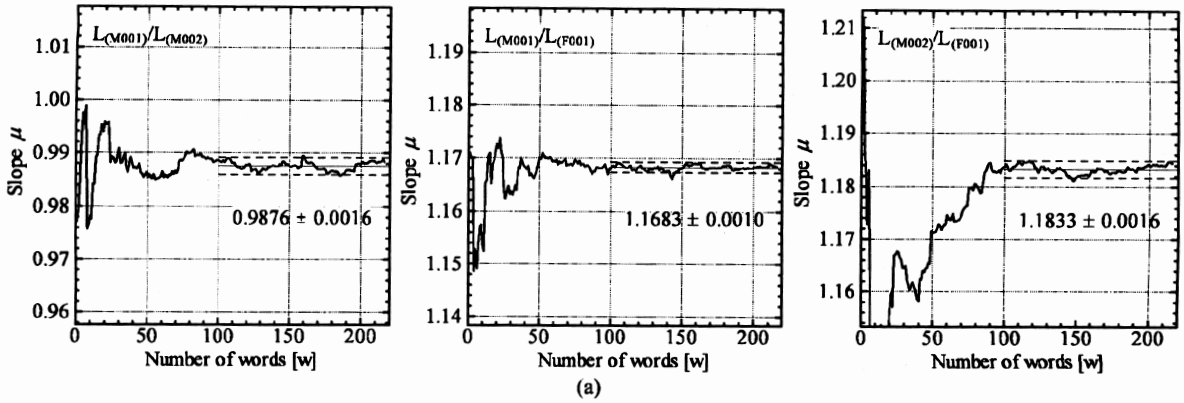
rect DP matching between two speakers (M002 and F001) on the right. Difference between the estimates by direct and indirect combinations of speakers is very small (below ±0.03%). Consequently, not only the result in the single speakers' case, but also the estimates of μ in (2.3) with a small dispersion and the relation (2.4) support that the proposed method is reliable for exact estimation of VTL ratios.

III. Direct Estimation Method for Simplification

A. Validity of the Method

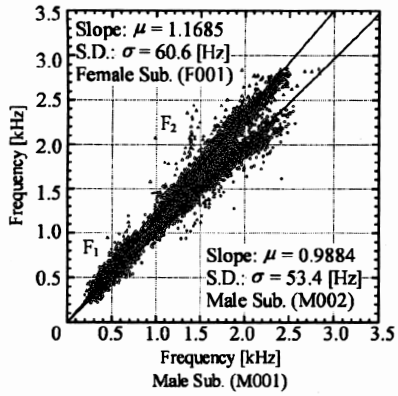
The strict estimation method requires each speaker to utter the same word a few times at different speeds. Conventional databases are often composed of a group of words or sentences that are uttered only once by a speaker. For that reason, we next investigated a technique to estimate VTL ratios using all frames of words uttered only once at a moderate speed. Three examples of Fig. 1(c), which were uttered at different speeds by a speaker, illustrate that the slopes of the first principal components in slow utterance to the fast one are clearly greater than one (unity) because the coarticulation effect narrows the width of formant changes in faster utterances. However, the slope of the conditional first principal component that is common to F_1 and F_2 appears intuitively as unity. This intuition is based on the following facts. First, if two utterances have similar speed, correlation diagrams of F_1 and F_2 distribute near the diagonal axis (the conditional first principal component whose slope is nearly equal to one). Second, when two utterance-speeds differ, correlation diagrams rotate slightly in each fulcrum of intersections between the diagonal axis and two first-principal axes for F_1 and F_2. As an utterance becomes faster, vocal tract movements are neutralized around a fixed configuration at the instant of passing the fulcrum's formants; consequently, the widths of formant changes become narrower. Thereby, if utterance A is faster than B, the correlation diagrams (D1) in Fig. 1(c) rotate counterclockwise, and *vice versa*. Third, intersections exist near the center of gravity of each correlation diagram in respective formants. Roughly speaking, the conditional first principal component always appears to be positioned such that it maintains the balance of each correlation diagram for F_1 and F_2 in utterances at moderately different speeds.

Slopes of the conditional first principal component may be somewhat unstable in a small number of word-utterances because individual differences of formant frequencies are relatively large in voiced sounds with zeros such as nasals, even in common words. Nevertheless, the balance will become stable if the proportion of vowels increases through the use of many varieties of words. If the above conjecture is true, a stable slope of the conditional first principal component is obtainable from many utterances. Thereby, we have estimated the speech length that is necessary for slope convergence after confirming that the slope converges. Three examples of convergent processes in the same set of speakers as that of Fig. 3(a) are shown in Fig. 4(a). That figure depicts the changes of slope



(a)

Convergent processes of slopes (Numerals: average slope ± confidence interval at 95% from 100 to 216 words)



(b)

Correlation diagrams and slopes estimated with 216 words

Fig. 3. Strict estimation results of average slopes from utterances by two speakers. (VTL ratio: reciprocal of average slope).

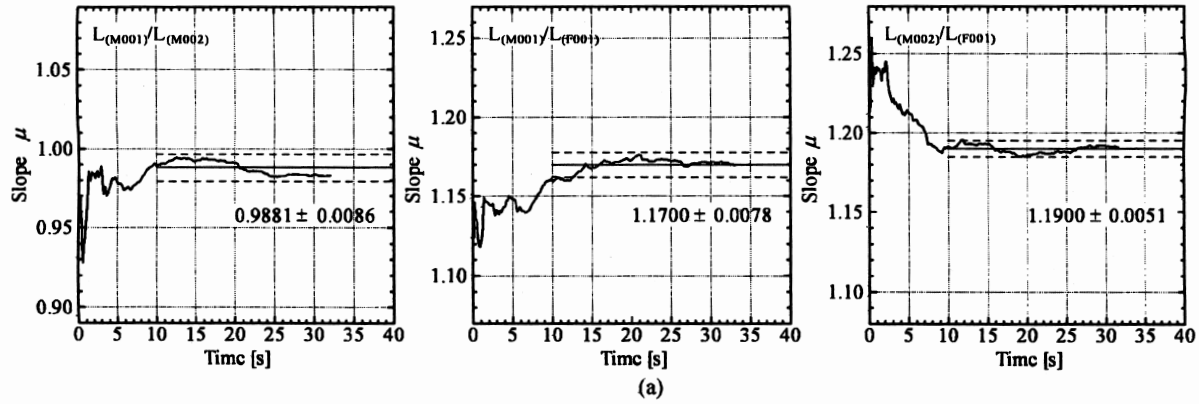
of the conditional first principal component by adding words. Results including other cases demonstrate that the variation of estimates is sufficiently small (less than 0.9%) in the confidence interval at 95% after collecting 1000 data-frames. Relation (2.4), shown in the strict estimation results, also holds in this simplified method. Therefore, it is understandable that the slope converges almost perfectly if speech data of more than 1000 frames (10 s) are used for estimation. As a rough estimate, 10 s of speech is satisfied with about 15 Japanese word utterances; we used 30–40 words to ensure convergence in this research. Correlation diagrams [D1 of Fig. 1(c)] that were obtained from all frames in normal-normal speed utterances are shown in Fig. 4(b) for comparison with another (D2) showing intersections only [Fig. 3(b)]. Standard deviations (65–76 [Hz]) of distances from the conditional first principal axis to samples are somewhat larger than those (53–61 [Hz]) in Fig. 3(b) because of estimation from all corresponding points. Standard deviations are larger for greater utterance-speed ratios. This sample dispersion influences estimation errors for VTL ratios in the simplified method.

Nevertheless, this method may be practical because utterance-speed ratios are usually limited. Error rates will be estimated in experiments of Section III-B. We refer to this simplified method as the "direct estimation method."

B. Experimental Results

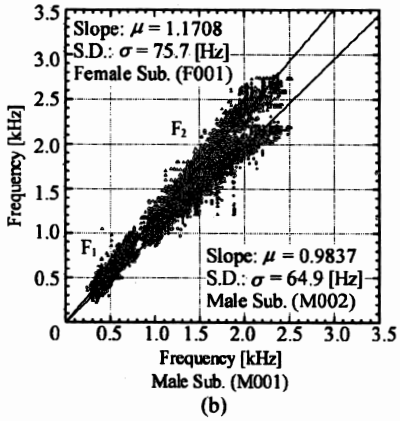
1) Single-Speaker Case: To examine the validity of the direct estimation method, we directly computed conditional first principal components using all frames of F_1 and F_2 trajectories from three groups of 36 word-data from each of the two word-groups (W1 and W2) uttered at different speeds by each of the speakers: M001, M002, and F001. Fig. 5(a) shows VTL ratios and errors estimated in the relation with utterance speed ratios. Context-dependent variations are represented using bars representing SD. In Fig. 5(a), individual differences among speakers are visible along with the correlations between estimation errors and utterance speed ratios. Individual differences among speakers in error magnitudes reflect the extent of consciousness of trying to pronounce words clearly at any utterance speed: changes of coarticulation effects caused by a change in utterance speed are smaller in utterances by speaker M002 than in others.

Next, we introduce Fig. 5(b), which directly represents the relation between estimation errors and utterance speed ratios, by reforming Fig. 5(a). Seeking the regression functions based on the least-mean-square-error (LMSE) method in Fig. 5(b), we can investigate errors when applying the direct estimation method to equal-speed utterances that remove the influence of utterance speeds to estimates. The figure shows that all regres-



(a)

Convergent processes of slopes (Numerals: average slope ± confidence interval at 95% from speech between 10 and 32 s)



(b)

Correlation diagrams and slopes estimated with 30 words

Fig. 4. Direct estimation results of average slopes from utterances by two speakers. (VTL ratio: reciprocal of average slope).

sion functions pass the origin, and that the intercepts, that is, the estimation errors in equal-speed utterances are ±0.25% at most. This figure also indicates that if utterance speed ratios are limited to an appropriate range, direct estimation is effective. As a rough conclusion for single speakers' case from Fig. 5(b), if a speed ratio of "faster" to "slower" is controlled to below about 2.5, errors are regarded as less than ±2%.

2) Two-Speaker Case: Next, we estimated VTL ratios between two speakers using the direct method from all combinations of utterances at three different speeds. Fig. 6 shows the relation between estimated error rates and utterance speed ratios in the two-speaker case. The average error rates are represented together with standard deviations of variation caused by seven different word-groups, each comprising 30 words. The individual errors are defined as deviations from the strict estimates.

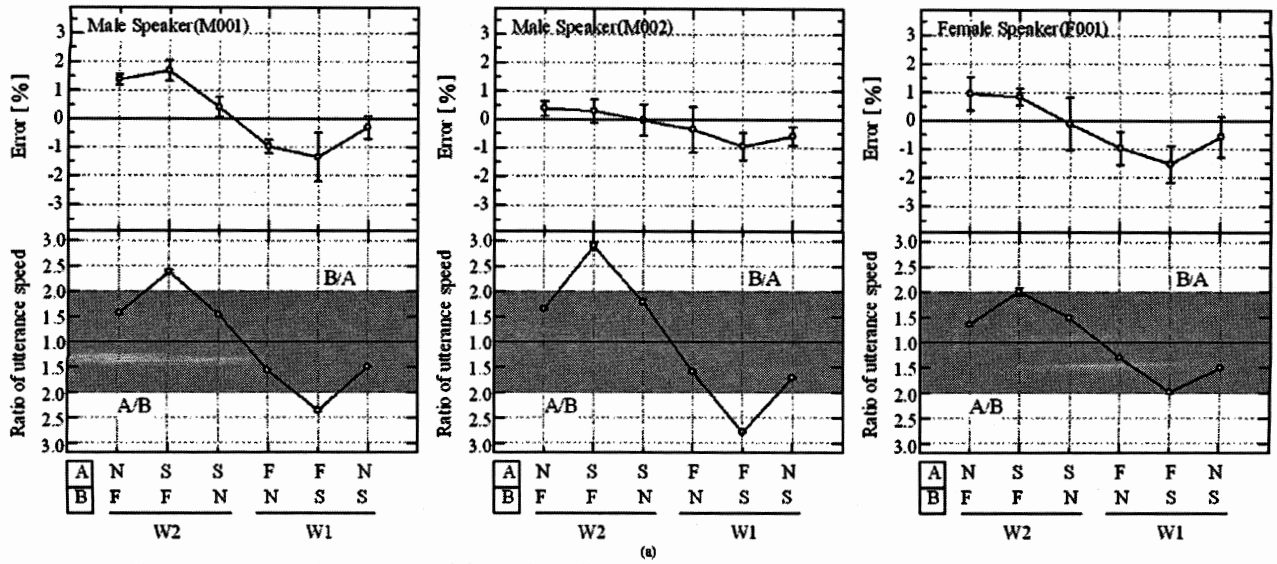
As in the single speakers' cases, we can seek direct estimates of VTL ratios in equal-speed utterances by gaining a regression function applying LMSE method to data points of Fig. 6. Estimation errors in equal speed are smaller than ±0.3%, as shown in Fig. 6: the direct estimates in equal-speed utterances almost exactly correspond with the strict estimates. According to analogies with the single speakers' cases in Section II-C2) and Section III-B1), we inferred that both strict estimates and direct estimates for equal-speed utterances closely approximate the true ratios also in two speakers' cases. Furthermore, we infer from Fig. 6 that errors in direct estimation between two speakers are smaller than ±2% if the average ratio of utterance speeds is within 2.0.

Fig. 6 also shows a margin for utterance speeds necessary to estimate vocal tract length ratios within ±2% of the error rate. When imitating utterances based on their auditory images, usual speech with a greater than 2.0 speed ratio of "faster" to "slower" is rare. Therefore, if speakers utter words or sentences naturally, the utterance speed will be controlled within the margin to hold error rates low.

IV. Estimating Relative Vocal Tract Lengths of Speakers in Some Databases Containing Different Words

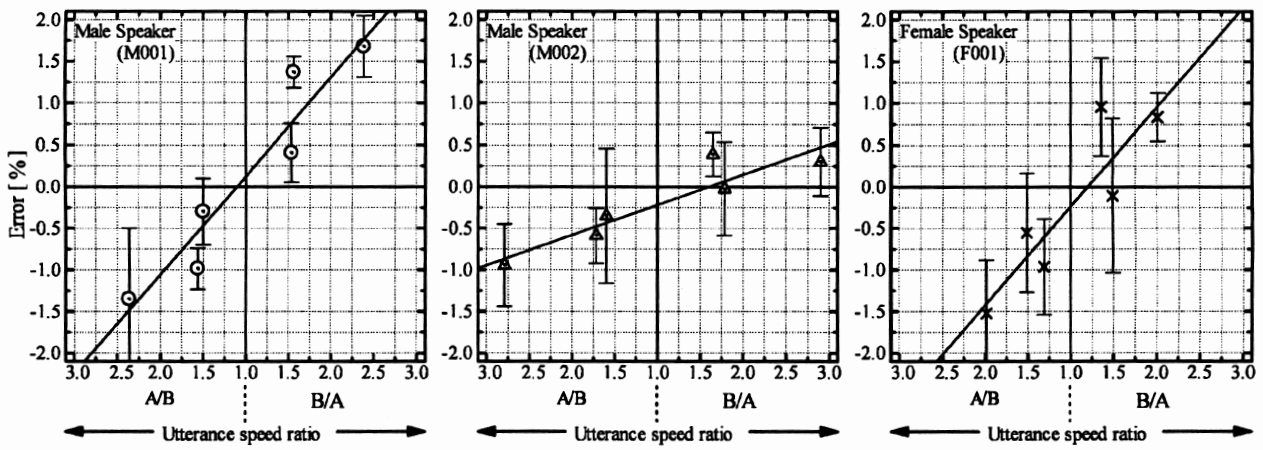
A. General Procedure

When estimating relative vocal tract lengths of speakers in a specific database, we first arbitrarily choose a standard speaker out of all speakers in the database. Next, the ratio of vocal tract lengths between the standard speaker and every other one is estimated by the direct estimation method using sets of the same word's utterances. By assuming a standard speaker's vocal tract length to be unity, we can obtain a relative value for any speaker. Subsequently, it is easy to represent all estimates by shifting the standard to an average, for example, to that of males' vocal tract lengths.



(a)

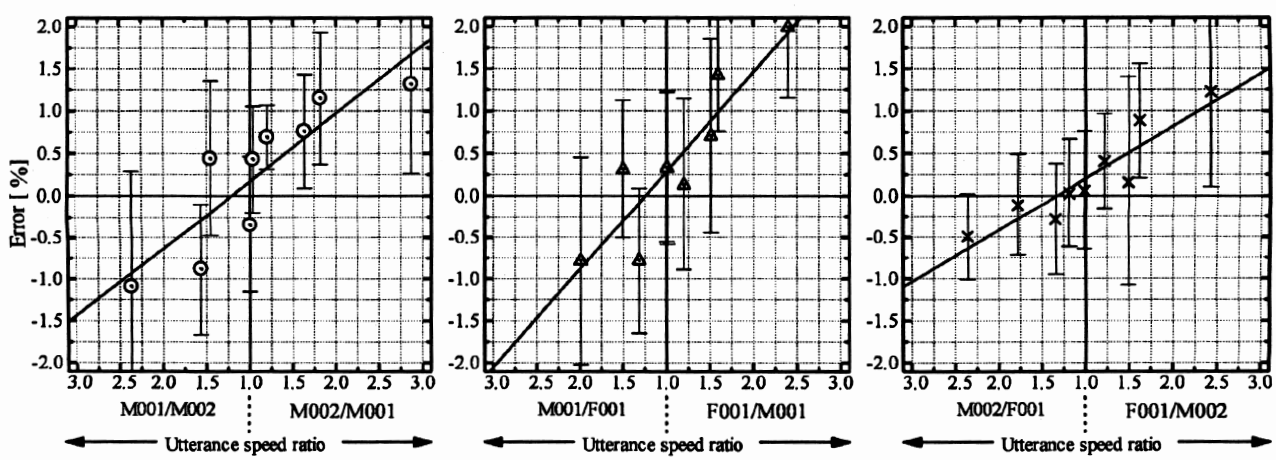
(1) A: utterance assigned to horizontal axis in correlation diagram, B: vertical axis (2) F: fast utterance speed, N: normal, S: slow
(3) W1: odd numbered words of 216 words in ATR database, W2: even numbered words



(b)

A: utterance assigned to horizontal axis in correlation diagram, B: vertical axis

Fig. 5. (a) Changes of error rates in direct estimation caused by the difference of utterance speeds. (When VTL ratio is 1.0, error is absolutely 0). (b) Relation between utterance speed ratios and error rates by direct estimation. (Single-speaker case: Regression functions based on LMSE criterion pass the origin).



[Male speaker]: M001, M002, [Female speaker]: F001

Fig. 6. Relation between utterance speed ratios and error rates by direct estimation (Two-speaker case: Regression functions based on LMSE criterion pass the origin).

TABLE I
ANALYSIS OF VARIANCE FOR THE ESTIMATES

Speaker group	Sum of squared deviations				F	Significance level at 0.05[%]	
	Within a class		Between classes				
<Database 1>		(d.f.)		(d.f.)			
(Male)	0.003074	(220)	1.035361	(19)	3900.0*	>>	2.19
(Female)	0.002037	(220)	0.251086	(19)	1427.0*	>>	2.19
<Database 2>							
(Male)	0.002460	(220)	0.148454	(9)	737.7*	>>	2.81
(Female)	0.001935	(220)	0.148454	(9)	747.1*	>>	2.81
<Database 3>							
(Boy)	0.034388	(920)	1.125095	(39)	771.8*	>>	1.67
(Girl)	0.025711	(920)	1.317149	(39)	1208.5*	>>	1.67

Variance ratio: F=(Variance between classes)/(Variance within a class)
Degrees of freedom: (d.f.)

When applying this procedure to some different databases that consist of different words uttered by different speakers, we require an intermediate speaker who utters all words used for estimation in those databases. First, relative vocal tract lengths of all speakers to the standard speaker are estimated in individual databases. Next, VTL ratios of standard speakers of all databases to the intermediate speaker are sought as in the case within a database.

Those results can suggest a simple method for computing relative vocal tract lengths with a unified standard. That is, when the conversion coefficient (slope) of formant frequencies from an intermediate speaker M to a standard speaker A_1 in the database A is a a_1 and the one from A_1 to arbitrary speaker A_k is a_k $(k = 1, 2, \ldots N)$, the VTL ratio between M and A_k is represented as $L_{Ak}/L_M = 1/(a_1 \cdot a_k)$. Similarly, we can obtain the ratio between M and B_j $(j = 1, 2, \ldots, L)$ of another database B as $L_{Bj}/L_M = 1/(b_1 \cdot b_j)$, using coefficient b_1 from M to standard speaker B_1 and b_j from B_1 to B_j of database B. Consequently, we obtain relative vocal tract lengths of all speakers belonging to databases (A and B) in the standard of that of the intermediate speaker M. Subsequently, if we obtain an average VTL ratio (L_{AV}) for some speakers under the standard L_M, we can change the standard vocal tract length from L_M into L_{AV} by multiplying each ratio by L_M/L_{AV}.

B. Estimation Experiments From Three Different Databases

1) Databases: Databases used for estimation experiments are the following.

1) Database 1: 30 words uttered by 20 adult males and 20 adult females selected from the ATR Japanese speech database comprising 216 phonetically-balanced words [17].
2) Database 2: 30 words uttered by 10 adult males and 10 adult females selected from the Tohoku University and Panasonic isolated spoken-word database containing 212 words [18].
3) Database 3: 40 words familiar to children, uttered by 40 boys and 40 girls of 7–10 years old from our original database.

Most words in the constructed databases are 4–6 syllable words because short words do not usually show clear formant changes.

2) Estimation Results and Reliability of Estimates: Based on the direct estimation procedure, each word uttered by the standard speaker of each database was matched directly by dynamic programming with the same word-utterances by all other speakers belonging to the database. Consequently, VTL ratios between the standard speaker and every other speaker were estimated in each database. Next, we obtained ratios between an intermediate speaker and each of three standard speakers. For that reason, the choice of standard speaker and intermediate speaker may influence estimation results. Finally, we investigated the sensitivity of estimates by changing standard speakers and an intermediate speaker as well as the estimates themselves.

We arbitrarily chose one male and one female from each of the three databases as standard speakers. In addition, we adopted two males and one female, whose ages were 22–24 years, as intermediate speakers. Estimations were conducted in all combinations of the standard speakers and the intermediate ones.

In databases 1 and 2, adult speakers pronounce words naturally during database construction. For that reason, all average ratios of utterance speeds, which are defined as faster/slower for every word, are within 2.0. That is, differences of utterance speeds between standard speakers and every other speaker are relatively small, so error rates of estimation can be regarded as $\pm 2\%$ at most. Moreover, those between intermediate speakers and standard ones suggest small errors of the same extent. With database 3, which consists of boys' and girls' utterances, after removing one speaker whose average speed ratio is beyond 2.0, estimation errors can be guaranteed to be within $\pm 2\%$.

After estimating ratios between a standard speaker and every other one, we obtain normalized vocal tract lengths of individual speakers so that an average length for all adult male speakers will become unity. Fig. 7 represents estimation results as an average value and a standard deviation for each speaker. The speaker's number of Fig. 7 is assigned with distinction of sex; children are divided into four groups by age. The standard deviation of each speaker indicates variations of estimation caused by differences in standard speakers and intermediate speakers. Standard deviations are below 0.4% and 0.8% of the averages in adults and children, respectively. Regarding Fig. 7, in which the averages and the standard deviations are represented for individuals, we infer that variations of one speaker are much smaller than those between speakers. This intuition is supported by analysis of variance, as shown in Table I. As a conclusion for Japanese speakers, we estimated that averages of relative vocal-tract-length in adult males, adult females, and 7–10-year-old boys and girls as 1, 0.79, 0.73, and 0.70, respectively. Moreover,

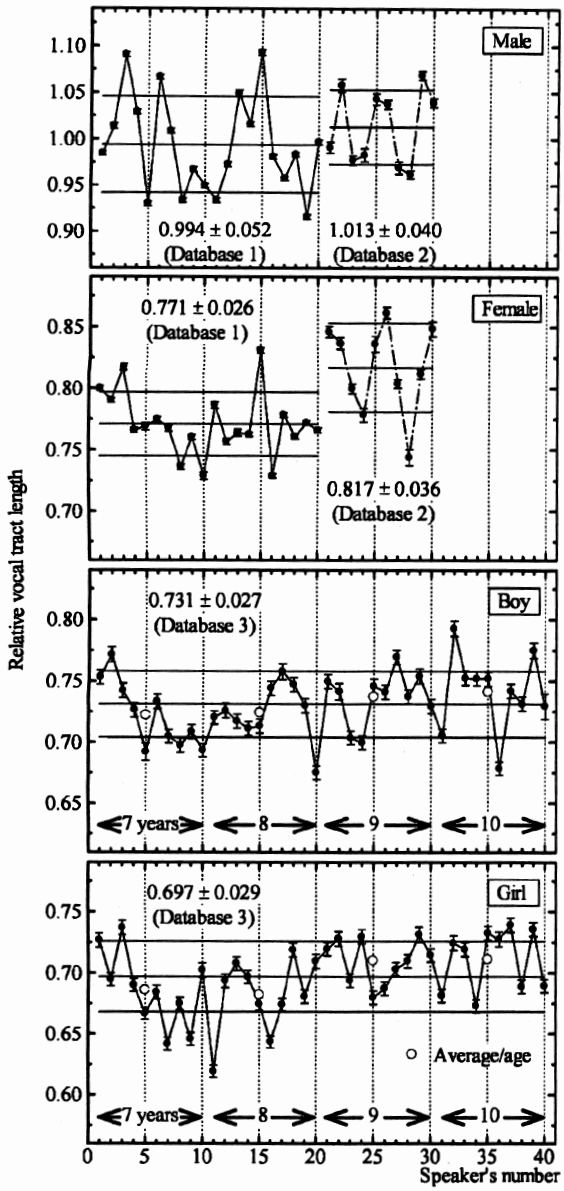


Fig. 7. Relative vocal tract lengths of Japanese speakers different in terms of age and gender.

variations of individual vocal-tract-lengths within each speaker-group have proved to be 4–5% of average VTL in the standard deviation. We have appended averages and standard deviations as numerals for every speaker-group (adult males, adult females, boys or girls) to Fig. 7.

C. Formant Normalization Effect by Relative Vocal Tract Lengths and Application To Word Recognition Tests [21]

Fig. 8 shows a change in the F_1–F_2 contour of five Japanese vowels before and after formant normalization according to relative vocal tract lengths. For adults' data and children's data, we used all utterances in database 1 (216 words by 20 males and 20 females) and database 3 (40 words by 40 boys and 40 girls) in Section IV-B1). We carried out phonemic segmentation and labeling in each word by visual inspection using "Speech Analysis and Representation Tool" (Speech ART) [19]: That software applies a speech visualization method [20] that can clearly indicate phonemic segments. Formant frequencies were extracted from 25% long of central parts in vowels, which are longer than 100 ms, in words. We obtained normalized formants by multiplying original formants of each speaker by their respective relative vocal tract lengths shown in Fig. 7. Fig. 8 represents 2σ-ellipses in each vowel distribution, which is defined as twofold standard deviations (2σ) in each of the first and second principal components. The figure suggests that normalization greatly improves the compactness of the distribution within each vowel and separations between vowels.

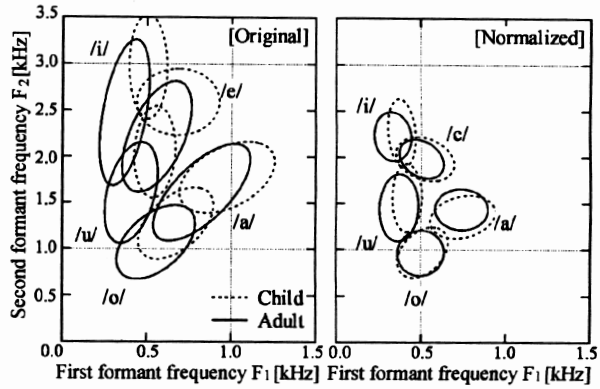


Fig. 8. Formant normalization effect on five Japanese vowels by relative vocal tract lengths (Normalized to average vocal tract length of adult males).

Next, we applied parameter-normalization by the estimation results of VTL ratios to word recognition tests using phoneme templates [21]. The VTL ratios were used to produce a template that is normalized for a specific (standard) speaker by linearly warping frequency scales of parameters. All phoneme templates were constructed while maintaining a balance between speaker groups, using all utterances of database 1 and half of database 3. Regarding recognition tests, 30 adult males and 30 adult females (Ad: adults), who are different from the speakers in database 1, and the remaining half (20 boys and 20 girls, Ch: children) of database 3 joined as speakers. Prior to recognition, they uttered 30 key words during about 30 s to calculate their individual normalizing factors (VTL ratios) that make their parameters approach to the specific speaker's. When using the normalized single template that is common to adults and children, the word recognition rates with a dictionary of 5000 items improved 4.2(Ch)-11.6(Ad)% compared with use of a nonnormalized template, and were 1.6% (Ad & Ch) higher than when using the advantageous one of two nonnormalized templates that were constructed individually using adults' and children's utterances. The word recognition system using the normalized template common to adults and children showed good performance (75.0(Ad)-77.5(Ch%) in recognizing words uttered by speakers of unknown age and gender. Consequently, we infer that a small number of key word utterances by each speaker quickly and easily reveals idiosyncratic vocal characteristics to the speaker-independent speech recognition system and improves its performance. This fact demonstrates the feasibility of the proposed method.

V. Discussions: Formant Ratios for Vowel Normalization

Two perfectly analogous vocal tracts produce a constant ratio in formant frequencies that are inversely proportional to their respective vocal tract lengths. That is one of the unaffected assumptions, which are first derived from the theory of acoustic tubes. Therefore, we consider that the VTL normalization function between two speakers should be linear with frequency. However, Fant suggested in a well-known study [22] that formant normalization between adult male, female and child's speakers should be made nonuniform in every vowel based on average formant data of various languages. Its physiological explanation was also provided as "only one variant of many possible combinations" [23]. Afterwards, based on the explanation, a nonlinear warping function for the VTL normalization was proposed using a factor that was estimated from F_3 for individual speakers [1]. We next describe briefly why these studies reached different conclusions.

In most traditional studies, formant distributions were investigated from utterances of isolated (sustained) vowels [22], [23] or those in a fixed context (vowel V in /hVd/) [24] because of limits in formant estimation techniques of those days (1952–1973). Apart from the formant estimation errors especially in high pitch voice, the following factors might have influenced the conclusion. When speakers pronounce a small number of vowels, they will make efforts to produce clear articulation for each vowel being conscious of the differences from the other ones. As a result, the articulatory space may be more exaggerated than that of vowels in words. Moreover, it is no wonder that adult speakers are more serious and more skilful than children at pronouncing correctly designated vowels. As a more essential reason for the nonuniform formant normalization, it may be adequate to mention the nonuniform nature of vocal tract growth [22], [25]. However, in any case, we consider that not only differences in VTL, but also those in vocal tract configurations between speakers affect the formant ratios in the case where estimated from utterances of isolated vowels (or those in a fixed context).

In contrast, we estimated the formant ratios from utterances of many varieties of words. Thereby, speakers do not need to be nervous about pronouncing individual vowels correctly. Besides, as stated in II, we obtained a VTL ratio in the strict estimation as a reciprocal of the average formant ratio from data of many instances for which vocal tract configurations are analogous between two speakers. The estimated individual formant-frequencies distribute with a small dispersion around a linear function, whose slope is the average formant ratio. This phenomenon is evidence that linear warping with frequency is adequate for VTL normalization. In addition, the direct estimates also approximate the strict ones.

Consequently, we can confirm the normalization effect by the result in Fig. 8. The normalized formant distributions in Fig. 8 have not been computed from centroids of formant distributions in individual vowels within the speaker groups, but from VTL ratios between individual speakers and a standard speaker. Fig. 8 shows that the normalization effect is satisfactory. We consider that this result justifies the utility of the proposed VTL normalization method. If the nonlinear function [1] instead of our linear warping function by a VTL ratio is used, it will be difficult to obtain good effects of normalization at least in F_1 and relatively low F_2, judging from the functional form and the result in Fig. 8. Furthermore, the reliable formant estimation method [13] contributes greatly to the useful normalization.

VI. Conclusion

We have proposed simple and reliable methods to estimate relative vocal tract lengths of speakers using first and second formant-trajectories of common words. In the first method, which is called "strict estimation method," we seek instances of analogous configuration in two vocal tracts whose lengths are estimated as a ratio. At those instants, the ratio of first (and second) formant frequencies between two utterances is theoretically constant: that ratio indicates a reciprocal of the VTL ratio. To find those instances, we applied the DP matching method to formant trajectories extracted from two utterances of the same word. When collecting ratios of formant frequencies at those instances from more than 100 words, the average ratio that was estimated by conditional principal component analysis proved to be an excellent estimate of the inverse ratio between two vocal-tract lengths (about ±0.1% error rate).

Next, as a simplification of the method, we attempted direct estimation of the vocal tract length ratio by applying conditional principal component analysis to all corresponding points of formant trajectories. Estimation errors by the simplified method were about ±0.3% at equal utterance speed and ±2.0% or less in cases where the ratio of utterance speed was within 2.0.

Finally, we applied the direct estimation method to three databases containing different words uttered by speakers of different age and gender. Adopting an intermediate speaker who utters all words in all databases, we estimated relative vocal tract lengths of 140 speakers in all. These results suggested three new and important points. 1) When average vocal tract length of adult male speakers is standardized as unity, those of adult females, 7–10-year-old boys, and 7–10-year-old girls are estimated as 0.79, 0.73, and 0.70, respectively. 2) Individual differences of vocal tract lengths among speakers within a group indicate approximately 4–5% of each average value, in the standard deviation. 3) The influence on estimates of replacing standard speakers in the databases or an intermediate speaker with other ones is very small.

The estimated VTL ratios have not been confirmed directly by physiological observations with MRI, because it is extremely laborious to do for many subjects. We hope that manageable high-speed imaging and recording high-quality speech sounds together will be realized in the near future and the proposed method will be confirmed by direct observations.

As an application for investigating the practical utility of the direct estimation method, we attempted word recognition tests using parameters normalized by the estimated vocal tract lengths. Thereby, we confirmed that the proposed method is effective to improve speech recognition rates when recognizing, with a single template, words uttered by speakers of unknown age and gender. Thus, we conclude that the proposed methods are not only simple and reliable for estimation of relative vocal tract lengths; they are also applicable to speaker normalization for speech recognition or probably to voice-characteristic conversion for speech synthesis.

APPENDIX
CONDITIONAL PRINCIPAL COMPONENT ANALYSIS

We simply describe a two-dimensional case of analysis used directly in this paper. Through analysis, we obtain a linear equation ($y = \mu x$) that passes through the origin and has the least-mean-square-error distance from a group of points ($(x_i, y_i)\ i = 1, 2 \ldots N$) on the x-y plane. An intersection (x_0, y_0) between $y = \mu x$ and its perpendicular line $y = -(1/\mu)x + b$ is represented as

$$x_0 = \frac{b}{\left(\mu + \frac{1}{\mu}\right)} \text{ and } y_0 = \mu \cdot \frac{b}{\left(\mu + \frac{1}{\mu}\right)}. \tag{A1}$$

Therefore, using (A1) and $b = \mu^{-1}x_i + y_i$, the square distance between (x_i, y_i) and (x_0, y_0) is

$$\begin{aligned} D_i^2 &= (x_i - x_0)^2 + (y_i - y_0)^2 \\ &= x_i^2 + y_i^2 - \frac{\left(x_i^2 + 2\mu x_i y_i + \mu^2 y_i^2\right)}{(\mu^2 + 1)}. \end{aligned} \tag{A2}$$

From $d\left(\sum D_i^2\right)/d(\mu) = 0$ and $\mu > 0$, the slope μ to minimize $\sum D_i^2$ is represented as

$$\mu = \frac{\left[(\beta - \gamma) + \{(\beta - \gamma)^2 + 4\alpha^2\}^{1/2}\right]}{2\alpha} \equiv \mu_1. \tag{A3}$$

where $\alpha = \sum x_i y_i$, $\beta = \sum y_i^2$, and $\gamma = \sum x_i^2$. Thus, we obtain the slope μ from (A3).

On the other hand, if we replace β and γ mutually in (A3), then

$$\frac{\left[-(\beta - \gamma) + \{(\beta - \gamma)^2 + 4\alpha^2\}^{1/2}\right]}{2\alpha} \equiv \mu_2. \tag{A4}$$

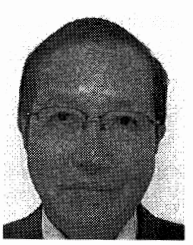
Therefore, from (A3) and (A4), $\mu_1\mu_2 = 1$. ($\because \mu_2 = 1/\mu_1$). This relation is reasonable for estimation of relative vocal tract lengths because we may assign each of two speakers to either axis on an x-y plane.

ACKNOWLEDGMENT

The authors gratefully acknowledge the valuable comments of Dr. Y. Ueda and an anonymous reviewer on this manuscript.

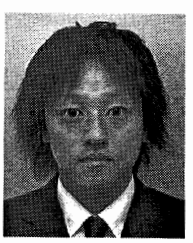
REFERENCES

[1] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP*, vol. 1, 1996, pp. 346–348.
[2] A. Faria, "Pitch-based vocal tract length normalization," in *Proc. Int. Comput. Sci. Inst.*, 2003, TR-03-001, pp. 1–14.
[3] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans., Speech Audio Process.*, vol. 6, no. 1, pp. 49–60, Jan. 1998.
[4] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 415–425, Sep. 2002.
[5] T. Bear, J. C. Gore, L. C. Gracco, and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *J. Acoust. Soc. Amer.*, vol. 90, pp. 799–828, 1991.
[6] S. S. Narayanan, A. A. Alwan, and K. Haker, "An articulatory study of fricative consonants using magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 98, pp. 1325–1347, 1995.
[7] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 100, pp. 537–554, 1996.
[8] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1511–1522, 1999.
[9] B. H. Story and E. A. Hoffman, "The relationship of vocal tract shape to three voice qualities," *J. Acoust. Soc. Amer.*, vol. 109, pp. 1651–1667, 2001.
[10] J. M. Pickett, *The Sounds of Speech Communications*. Baltimore, MD: University Park, 1980, pp. 46–49.
[11] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 2, pp. 183–192, Apr. 1977.
[12] ——, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, no. 5, pp. 417–427, Oct. 1973.
[13] A. Watanabe, "Formant estimation method using inverse-filter control," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 317–326, May 2001.
[14] T. Kitamura, H. Takemoto, K. Honda, Y. Shimada, I. Fujimoto, Y. Syakudo, S. Masaki, K. Kuroda, N. Oku-uchi, and M. Senda, "Difference in Vocal Tract Shape Between Upright and Supine Postures Observations by an Open-Type MR Scanner," IEICE, Tech. Rep. SP2004-29(2004-6), 2004.
[15] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993, pp. 229–232.
[16] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. London, U.K.: Academic, 1979, pp. 213–254.
[17] H. Kuwahara, Y. Sagisaka, K. Takeda, and M. Abe, "Construction of ATR Japanese Database As a Research Tool,", ATR Tech. Rep, 1989.
[18] S. Makino, K. Niyada, Y. Mafune, and K. Kido, "Tohoku university and panasonic isolated spoken word database" (in Japanese), *J. Acoust. Soc. Jpn.*, vol. 48, pp. 899–905, 1992.
[19] A. Watanabe, H. Ikeda, T. Ikeda, and Y. Ueda, "Development of speech ART 2000," in *Proc. Autumn Meet. Acoust. Soc. Jpn.*, 2001, pp. 359–360.
[20] A. Watanabe, S. Tomishige, and M. Nakatake, "Speech visualization by integrating features for the hearing impaired," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 454–466, Jul. 2000.
[21] N. Ikeda, T. Sakata, T. Hirayama, Y. Ueda, and A. Watanabe, "Effects of speaker normalization based on vocal tract length ratios on word recognition using compound parameters" (in Japanese), *IEICE Trans. Information and Systems of Electronics Communications and Computer Sciences, D-II*, vol. J87-D-II, pp. 1418–1427, 2004.
[22] G. Fant, *Proc. Non-Uniform Vowel Normalization*, 1975, STL-QPSR, 2-3, pp. 1–19.
[23] ——, *Speech Sounds and Features*. Cambridge, MA: MIT Press, 1973, pp. 88–93.
[24] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175–184, 1952.
[25] L. Menard, J.-L. Schwartz, L.-J. Boe, S. Kandel, and N. Vallee, "Auditory normalization of french vowels synthesized by an articulatory model simulating growth from birth to adulthood," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1892–1905, 2002.



Akira Watanabe received the B.E. degree in electrical engineering in 1962, and the M. E. and D. E. degrees in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1964 and 1968, respectively.

He was with the Faculty of Engineering, Kumamoto University, Kumamoto, Japan, from 1967 to 2003. He investigated a real-time formant estimation method in 1978–1979 as a Guest Researcher at the Royal Institute of Technology (KTH), Stockholm, Sweden. He is now a President of Kumamoto Prefectural College of Technology and also a Professor Emeritus of Kumamoto University. His research interests include speech analysis, processing and coding for the hearing impaired.



Tadashi Sakata received the B.E. and M.E. degrees in computer science from Kumamoto University, Kumamoto, Japan, in 1998 and 2000, respectively.

He is a Research Associate in the Department of Computer Science, Kumamoto University. His research interests include speech analysis and synthesis to efficiently transmit speech through vision and residual hearing for the hearing impaired.