Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

A data reduction method to estimate vowel distributions and its use in comparing two formant estimation methods

Tadashi Sakata (1), Yuichi Ueda (1) and Akira Watanabe (2)

(1) Graduate School of Science and Technology, Kumamoto University, Kumamoto, Japan
(2) Kumamoto University, Kumamoto, Japan

PACS: 43.72.-p

ABSTRACT

Speech features such as formants of vowels uttered by many talkers are considered to form a normal distribution in each phoneme on a feature space. However, those features may apparently show the different dispersions peculiar to the estimation methods. Therefore, if the correct distributions can be found by a credible method, it will make clear the definition of feature estimation errors so that the comparative evaluations between the feature estimation methods will become possible. In this paper, we first propose the data reduction method to estimate true formant distributions of vowels. In the method, we apply the principal component analysis to the formant data of each vowel on a F_1-F_2 space to search an average value and a three-sigma ellipse. If the average and the ellipse are searched iteratively after removing the outside data of the ellipse regarded as errors, they finally converge. The proportion of the data samples within the final ellipse to all data will be different in the formant estimation methods. We consider that the estimation method of larger proportion is higher in the accuracy because of the high trust. IFC (Inverse Filter Control) method, in which formants are estimated from zero-crossing information, has been compared with LPC method under the above criteria. As a result of the analysis using vowels in words, it has been shown that the IFC method is superior to the LPC in the proportion and the ratio of area in the final ellipse to that in initial one. The proportions of data within the final ellipses are 90-96% in IFC and 84-95% in LPC, which are obtained from five Japanese vowels uttered by 20 males. The intuition obtained by observing the states of distributions supports the numerals in the analysis. Based on the results, we conclude that the formant estimation using zero-crossing information (IFC) is more effective than that by spectral shapes (LPC).

I. INTRODUCTION

Many formant estimation methods have been proposed. Most of them estimate the formants by the matching between spectral shapes of speech and a defined system function [1], [2]. For that, we previously developed a method to estimate formants from zero crossing information of components that were obtained by inverse filtering of speech signals [3]. In estimating formants from synthetic speech, the latter method, which we call IFC (Inverse Filer Control) method, is superior to LPC method as a representative of the former for using spectral shapes [4]. However, it seems that there is no criterion to evaluate by comparing the estimation accuracies of them for real speech of comprehensive features. This is based on that it is difficult to know true formant distributions of vowels in real speech. In this paper, we propose a method to estimate the true distributions, which is called the data reduction method. After that, based on the method, we compare accuracies of the two formant estimation methods, IFC and LPC.

II. FORMANT DISTRIBUTIONS ESTIMATED BY TWO METHODS

1. Speech database

We use 216 phonetically balanced words including in the ATR Japanese speech database. Speakers were 40 native speakers of Japanese (20 males and 20 females). All speech data have phoneme labels.

2. Formant estimation methods

Speech signals are quantized with 12 kHz of a sampling frequency. Analysis is carried out in 20 ms of a frame length and 10 ms of a frame shift. Under these conditions, we use two different types of formant estimation methods as follows:

(a) Inverse filter control (IFC) method [3]

The IFC method estimates formant frequencies from zero-crossing information of decomposed signals by controlling mutually 32 basic inverse filters. The inverse filter is characterized by a pair of complex conjugate zeros to eliminate a pair of complex conjugate poles as a formant. The basic inverse filters are connected to obtain separated formant components, which are ideally single resonant waves. In the system, in the mutual control of zero frequencies of the inverse filters with fixed bandwidths by the system output parameters, if the parameter is adequate, the zeros converge to formant frequencies and speech waves are finally separated into a set of pseudo-single resonant waves. Thus, formant frequencies can be estimated from a spectrum or zero-crossing information of the pseudo-single resonant waves in the filter outputs. The IFC method uses a spectral parameter only in the first control to obtain rough estimates of formants, and from the

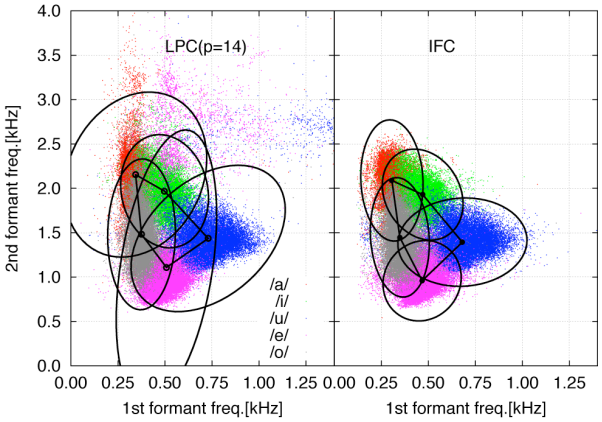
second control, the parameter is transferred to zero-crossing information. In this paper, we estimated the number of formants at six and five in the signal band (0-6 kHz) for male and female voices respectively, based on the relationship between their vocal tract lengths and neutral vowels.

(b) Linear predictive coding (LPC) method [2]

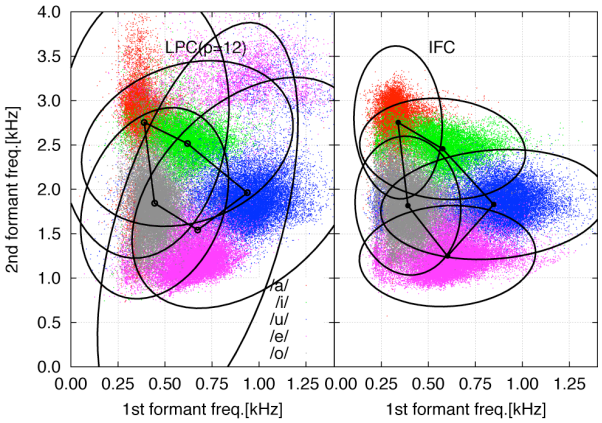
The LPC method is one of the most popular methods to estimate formant frequencies. This method is ultimately based on the best matching between a spectrum to be analysed and a synthesized one so that formant frequencies are estimated using spectral shapes. In the LPC in this study, linear predictive coefficients are first extracted by the autocorrelation method and next, formants are estimated by solving the polynomial equation represented using the predictive coefficients. The orders of LPC model used in this study are 12 and 14 for male voice, and 10 and 12 for female. We have selected poles only whose bandwidths are below 1000 Hz and regarded them as formants.

3. Estimated formant distributions

Figure 1 shows the formant (F_1-F_2) distributions of vowels in 216 words that are extracted from 50% of frames in the center of each vowel part. In the Figure, three-sigma ellipses are also depicted, which are estimated by the principal component analysis using all vowel data. Number of data frames in this analysis is described in Table 1.



(a) Male voice



(b) Female voice

Figure 1. The formant distributions of vowels in words extracted using the IFC and the LPC methods: The ellipses are obtained from all data using the principal component analysis. (Initial estimate for the data reduction method)

According to the results of analysis in Figure 1, we notice that LPC data apparently show the more widely dispersion than IFC data and may not represent features of normal distributions. If a talker's group is limited by sex and age such as adult males, adult females or children, the formants of each vowel will form the different normal distributions in the restrictive talker's category. We consider that this assumption can naturally be accepted from the hearing judgement. Therefore, the formants that do not belong to each suitable normal distribution can be regarded as estimation errors that are mainly caused by mismatches in the analysis conditions. Thus, it is necessary for deciding errors to estimate the true normal distributions.

Table 1. Number of processed data frames in each vowel.

	/a/	/i/	/u/	/e/	/o/
Male	23034	14737	21532	14768	27896
Female	24147	15490	22488	15547	28744

III. DATA REDUCTION METHOD FOR ESTIMATING TRUE DISTRIBUTIONS

As described in the previous section, if the true distributions can be found by a credible method, it will make clear the definition of feature estimation errors so that the comparative evaluation between the two formant estimation methods will be possible. Therefore, we propose the data reduction method to estimate true normal distributions of each vowel. In the method, we first obtain a center of distribution and a 3σ-ellipse from all data of a peculiar vowel, which are extracted in the pronounciations of each talker's group. Next, data put outside of the 3σ-ellipse are removed because of high possibility to be errors, and then the center as the weighted mean and the ellipse are recomputed again. Successively, the above processing is repeated until the center converges, that is, until the ellipse converges. In this case, the computation of 3σ-ellipse is carried out by restoring the removed data after updating the center by the weighted mean using the data of inside of the 3σ-ellipse. The outline of computations is shown as follows:

(1) Use of principal component analysis:

The equation of ellipse is represented as

$$\left(\frac{(x-x_c)\cos\theta+(y-y_c)\sin\theta)}{a}\right)^2+\left(\frac{-(x-x_c)\sin\theta+(y-y_c)\cos\theta}{b}\right)^2=1, \quad (1)$$

where (x_c, y_c) is the coordinate of center, a is 3σ-length of the major axis, b shows it of the minor axis, and θ is an argument between x-axis and the major axis of the ellipse. These parameters are obtained by the principal component analysis.

(2) Computation of weighted mean:

The center of the distribution is computed by the weighted mean of M data as follows:

$$x_c=\sum_{i=1}^{M}W(x_i,y_i)x_i\Big/\sum_{i=1}^{M}W(x_i,y_i)$$
$$y_c=\sum_{i=1}^{M}W(x_i,y_i)y_i\Big/\sum_{i=1}^{M}W(x_i,y_i) \quad (2)$$

where the weights are indicated by

$$W(x_i,y_i)=1-\left\{\left(\frac{(x_i-x_c)\cos\theta+(y_i-y_c)\sin\theta)}{a}\right)^2+\left(\frac{-(x_i-x_c)\sin\theta+(y_i-y_c)\cos\theta}{b}\right)^2\right\}^{\frac{1}{2}} \quad (3)$$

The weighting function, $W(x, y)$ shapes an elliptic cone whose base represents the ellipse in (1) and whose height is a maximum at the center. That is, the weight decreases linearly from 1 at the center to 0 at the edge of the ellipse. (The image of the elliptic cone as the weighting function is shown in Figure 2.) This elliptical cone function is effective to reasonably remove data distant from the center of distribution and to estimate a true center of normal distribution.

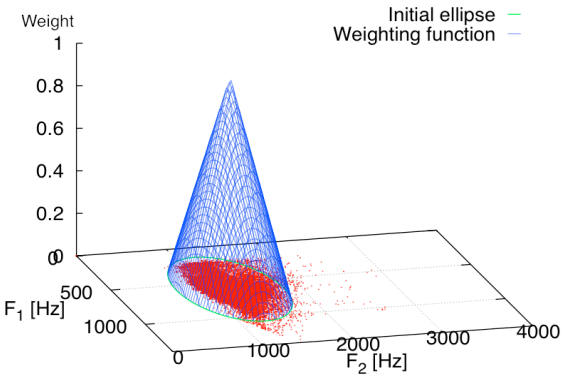


Figure 2. The image of the elliptic cone as the weighting function

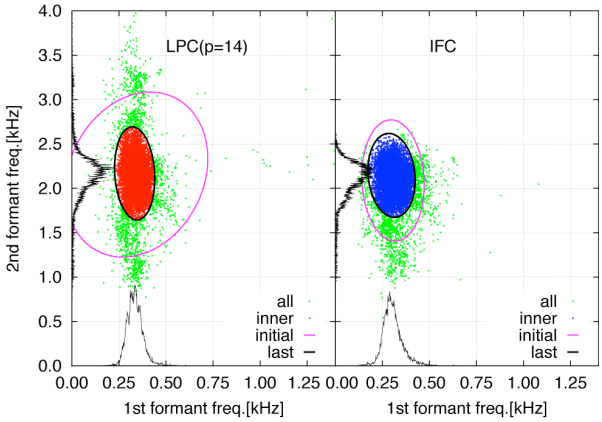
Decision of the parameters by the principal component analysis and the shift of the center in a coordinate by the weighted mean are executed iteratively until the shift is within 1.0 [Hz] in the Euclid distance. We investigated the center and the 3σ-ellipse after convergence. We consider that these features represent approximations of the true formant distributions estimated using the various methods.

IV. ANALYSIS RESULTS AND DISCUSSIONS

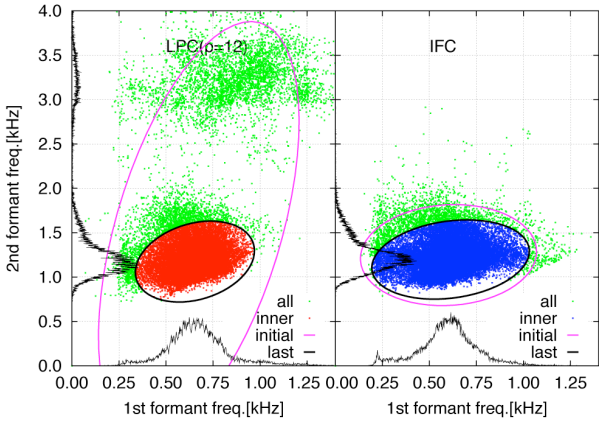
1. Change of 3σ-ellipses by applying the data reduction method

We applied the data reduction method to formant distributions extracted by the LPC and the IFC methods. Figure 3 shows the initial and final ellipses of /i/ by male and /o/ by female voices together with the F_1-F_2 diagrams. F_2 estimated by the LPC method spread widely, particularly in female voice, which can apparently be separated into two areas. Thus, the area of the initial ellipse by the LPC method is much larger than that by the IFC method. However, when applying the data reduction method to these data, we obtain the small areas as final ellipses after convergence in both IFC and LPC, in which points are dense. In the case of female /o/, the center of the distribution shifts as shown in Figure 4. As a result, we can point out the restricted area with high density.

Figure 5 shows the ratios of the LPC to the IFC elliptic areas in every vowel. Although the initial ratios of the ellipses by both formant estimations are large, the final ratios approximate 1.0. This tendency means that even if the original data include errors, the final distributions estimated by the data reduction method approximate the same one. In Figure 6, the coordinates of the center points after convergence are shown on the F_1-F_2 diagram. Those coordinates obtained by IFC and LPC is close to each other between male voices or female voices. Moreover, according to the relationship between the center points of male and female voices, those exist on the linear function through the origin and form analogous pentagons. This fact proves that resonant frequencies are inversely proportional to vocal tract lengths [5]. From Figures 5 and 6, we consider that the results of the data reduction method provide the true distributions whatever formant estimation methods are used if they work in the accuracy to some extent.



(a) Male voice; /i/



(b) Female voice; /o/

Figure 3 The initial and final ellipses extracted by the data reduction method.

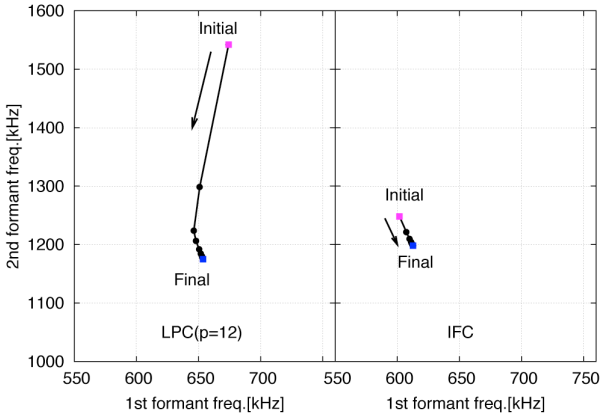


Figure 4. Examples of center's shift in the result of the data reduction method (Female /o/)

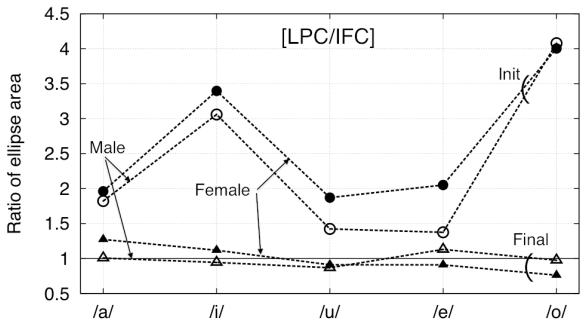


Figure 5. Ratios of two ellipse-areas (LPC/IFC)

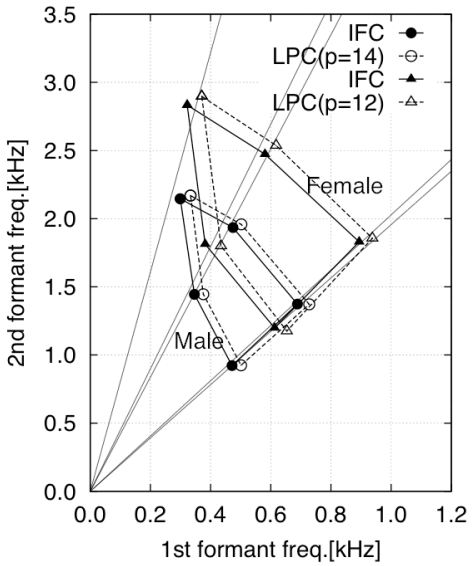


Figure 6. Centers of formant distributions by male and female voices: Five Japanese vowels

2. Comparative evaluations of formant estimation methods

From the results of previous section, we define the formant estimation errors by data points in the outside of the final ellipses. We evaluate comparatively the two formant estimation methods. Figure 7 shows changes of ellipses caused by the data reduction method using estimates by IFC and LPC respectively. The ratios of the initial to the final ellipses are larger in LPC than in IFC despite that the final ellipses are approximately equal in LPC and IFC as shown in Figure 5. This indicates that although many data points by LPC occupy a small area of the normal distribution in a vowel, we can also see gross errors to a certain degree. On the other hand, such points are rare in IFC.

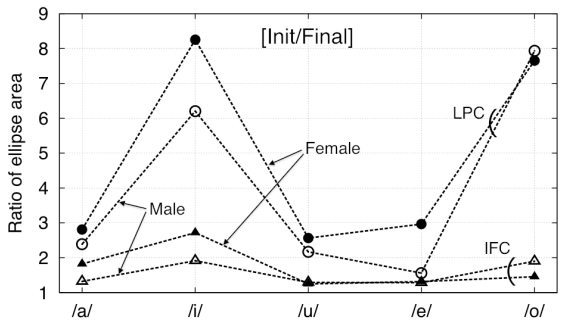


Figure 7. Ratios of two ellipse-areas (Initial/Final)

Table 2. Proportions of data samples within the final ellipses

(a) Male voice (unit:[%])

	/a/	/i/	/u/	/e/	/o/
IFC	95.70	89.96	95.87	95.51	90.11
LPC(p=14)	92.85	88.44	93.99	95.20	84.03
LPC(p=12)	91.61	87.67	90.64	94.08	65.98

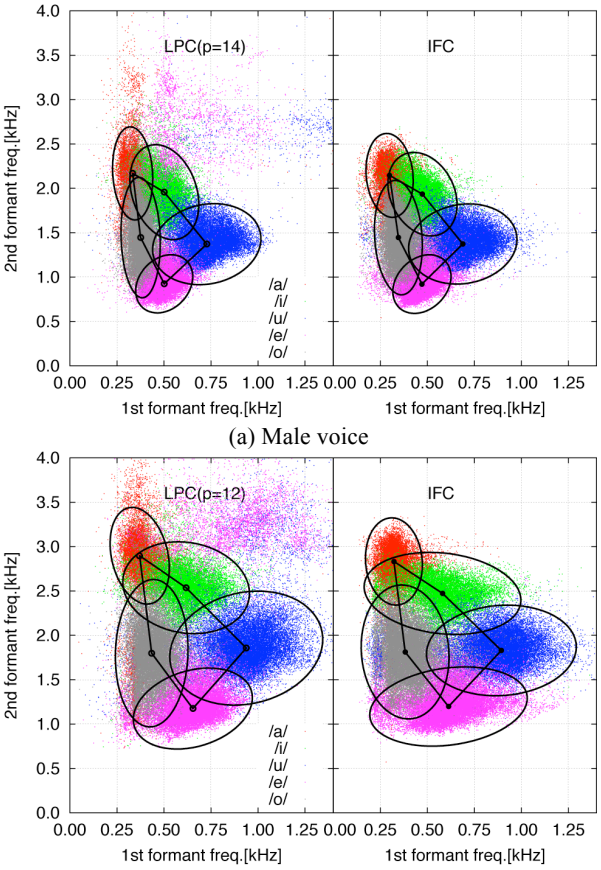
(b) Female voice (unit:[%])

	/a/	/i/	/u/	/e/	/o/
IFC	90.40	86.64	96.43	96.11	92.06
LPC(p=12)	87.84	78.02	93.58	91.19	76.62
LPC(p=10)	81.76	85.91	91.51	92.78	76.00

To express the accuracy of formant estimation, we propose using the proportions of the data within the final ellipses shown in Table 2. According to Table 2, the proportions in IFC are higher than it in LPC in all cases. The proportions are 90.0-95.9% and 86.6-96.4% in IFC for male and femal voices, respectively, and similarly 84.0-95.2% and 76.0-93.6% in LPC. Thus, we conclude that the IFC method is superior to the LPC method in formant estimation accuracy.

As a general rule, the estimation results by LPC are dependent on the analysis order and limits of bandwidths to select formants from extracted poles. Therefore, we used two orders that seemed to be adequate for analysing male and female voices respectively. The limit of bandwidths influences the judgement of excess or lack of poles. We estimated formants using the several bandwidths in the range of 300-3000 Hz, and as a resut, we chose 1000 Hz to be adequate on the average. A weak point in LPC method, we think, is that optimal analysis conditions change according as individual speech materials. This unstability in LPC does not appear in IFC method because the formant frequencies are estimated by zerocrossing frequencies.

Finally, in Figure 8, we show the final ellipses after convergence that are depicted on the formant distributions.



(a) Male voice

(b) Female voice

Figure 8. The formant distributions of vowels in words extracted using the IFC and the LPC methods: The ellipses show final estimates in the data reduction method.

V. CONCLUSION

If we can know the true formant area of real speech in a frequency space, it is possible to comparatively evaluate formant estimation accuracy by various methods. In this paper, we proposed the data reduction method to estimate true formant distributions of vowels. In this method, we assume that true formant data are in 3σ ellipse in each vowel. By executing computations of the ellipse's parameters by the principal component analysis and the weighted mean using weights of the elliptic cone iteratively, we obtained the convergent ellipses. When using data estimated by IFC and LPC methods, 76-96% of data points are put in the ellipses that are compact

with almost the same size and the same mean. Thus, we judged the data reduction method to be effective to estimate the true formant areas.

Next, we have compared the two formant estimation methods, IFC and LPC, by evaluating the areas of the ellipses and the proportions of data included in them. As the results, we conclude that the IFC method is superior to the LPC method in the accuracy for estimating formant frequencies. This phenomenon essentially indicates that direct formant information is contained more in zero-crossing parameters obtained by inverse filtering than in specral shapes. This conclusion is also supported by the analysis result of synthetic vowels [4].

REFERENCES

1 C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens and A. S. House, “Reduction of speech spectra by analysis-by-synthesis techniques,” *J. Acoust. Soc. Amer.*, **33**, 1725-1736 (1961)

2 J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech* (Springer-Verlag, Berlin, 1976).

3 A. Watanabe, “Formant estimation method using inverse-filter control,” *IEEE Trans. Speech Audio Process.*, **9**, 317-326 (2001)

4 T. Sakata, Y. Ueda and A. Watanabe, ”Zero-crossing-based formant estimation method: Its features and accuracy,” *Acoust. Sci. & Tech.*, **31**, 2, 195-198, (2010)

5 A. Watanabe and T. Sakata, ”Reliable methods for estimating relative vocal tract lengths from formant trajectories of common words,” *IEEE trans, Audio Speech Language Process.*, **14**, 1193-1204 (2006)