Model Based Voice Decomposition Method

Masahide Sugiyama

Graduate School of Computer Science and Engineering, The University of Aizu
Aizu-Wakamatsu, Fukushima 965-8580 Japan

Abstract: This paper proposes a voice decomposition method based on voice model. Using correlation distance as a criterion frequency domain decomposition is formulated. For a pair of correlation coefficient vectors minimizing powers of spectra is characterized using quadratic programming problem with constraint, where values of powers are non-negative. The solution of the problem without constraint is calculated using a linear equation where the coefficients are determined by correlation coefficient vectors. A corresponding matrix for the linear equation has a non-negative determinant, and the sufficient and necessary condition for existence of reverse matrix is that given correlation coefficient vectors are linear independent. The solution with constraint is also given by solving a linear equation.

1 Introduction

This paper proposes model-based voice decomposition method. Human being can segregate and distinguish each component in two or more mixtured and overlapped voices. However, it is too difficult to recognize each component in the overlapped voices, and the current speech recognition technology treats non-target voices as noises. The fundamental idea of this paper is that human being knows and has typical voice patterns and decomposes composed voice using these model patterns.

Voice waveform is transformed to a correlation vector as its feature vector. Therefore, the decomposition in waveform domain is translated to the decomposition in the correlation vector domain. On the other hand, generally voice is assumed to be represented by a set of representative vectors which are generated using VQ technique. Speech under noisy environment is represented as two sets of correlation vectors; one is from speech and the other is from noise. Cocktail effect speech by two talkers is also represented as two sets of correlation vectors; one is from first talker and the other is from the second talker.

For given number of mixtures and a set of representative vectors corresponding to its component, the decomposition of the input vector is defined as searching a pair of representative vectors which minimizes the correlation distance between input vectors and the composition of a pair of representative vectors.

For speech recognition decomposition on frequency (spectral) domain is sufficient. The author studied voice segmentation and clustering problem from multiple signal sources[1, 2], and this study is one extension overlapped voice from multiple signal sources[3]. auditory segregation of sound stream in [4] mentioned their method did not use voice characteristics, on the other hand, this paper stands on the positive use of this. For speech recognition, [5] studied composition of HMMs which represent sound sources. They did not study the decomposition problem.

2 Formulation of Decomposition

Two speech waves, $x_1(t)$ and $x_2(t)$, composes and the periodogram of its composed wave $x_s(t) = x_1(t)+x_2(t)$ is given as follows;

$$X_s(\lambda) = |\sum_t x_s(t)e^{-jt\lambda}|^2$$
$$= \sum_t |x_1(t)e^{-jt\lambda}|^2 + \sum_t |x_2(t)e^{-jt\lambda}|^2$$
$$+2\sum_{t,t'} x_1(t)x_2(t')e^{-j(t-t')\lambda}. \quad (1)$$

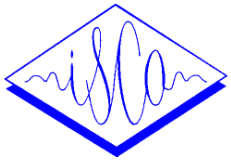
Here, suppose that $x_1(t)$ and $x_2(t)$ are independent (i.e., no correlation), Eq.(1) is simplified.

$$X_s(\lambda) = X_1(\lambda) + X_2(\lambda), \quad (2)$$

where, $X_1(\lambda), X_2(\lambda)$ are periodograms of $x_1(t), x_2(t)$, respectively. It is assumed that LPC spectra $f_s(\lambda)$ and $f_1(\lambda), f_2(\lambda)$ derived from X_s and X_1, X_2 assumed to have the same relationship.

$$f_s(\lambda) = f_1(\lambda) + f_2(\lambda). \quad (3)$$

When one of two spectra $f_i(\lambda)$ is uniquely known, the observation of $f_s(\lambda)$ determines the another spectra. This idea can be generalized as follows;



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

For given $f_s(\lambda)$, find $f_{1,j_1} \in V_1 = \{f_{1,j}\}$ and $f_{2,j_2} \in V_2 = \{f_{2,j}\}$ which minimize the following Eq.(4);

$$\min_{j_1,j_2} d(f_s, f_{1,j_1} + f_{2,j_2}). \quad (4)$$

Here, d is a spectral distance; LPC cepstrum distance or correlation distance. To get finite set of representatives there are two ways;

1. finite representatives $\{f_i(\lambda)\}$ to approximate spectral space
2. use constraint of LPC spectra

LPC cepstrum distance d between LPC spectra g_1, g_2 are calculated as follows;

$$d^2(g_1, g_2) = (c_{1,0} - c_{2,0})^2 + 2\sum_{n=1}^{N}(c_{1,n} - c_{2,n})^2 \quad (5)$$

Using power, u_i, and correlation coefficients, $r_{i,n}$, Eq.(3) is represented as follows;

$$f_1(\lambda) + f_2(\lambda) = (u_1 + u_2)\sum_{n=-\infty}^{\infty} \frac{u_1 r_{1,n} + u_2 r_{2,n}}{u_1 + u_2} e^{-jn\lambda}. \quad (6)$$

Therefore, power and correlation coefficients of $\hat{f}(\lambda) = f_1(\lambda) + f_2(\lambda)$ is calculated.

$$\begin{cases} \hat{u} & = u_1 + u_2 \\ \hat{r}_n & = \dfrac{u_1 r_{1,n} + u_2 r_{2,n}}{\hat{u}} \end{cases} \quad (7)$$

To calculate Eq.(5), LPC cepstrum coefficients $(\hat{c}_{j_1,j_2,n})$ $(n = 0, \ldots, N)$ are calculated from composed power and correlation coefficients in Eq.(7), and Eq.(4) is translated as follows;

$$d^2(f_s, f_{1,j_1} + f_{2,j_2}) = (c_{s,0} - \hat{c}_{j_1,j_2,0})^2 + 2\sum_{n=1}^{N}(c_{s,n} - \hat{c}_{j_1,j_2,n})^2$$

Therefore, for LPC cepstrum distance calculation, it is sufficient that representatives, $V_i = \{f_{i,j}\}$, have only information on its power and correlation coefficients.

3 Voice Decomposition Using Correlation Distance

In order to estimate power term the decomposition using correlation distance is described considering each model V_i is represented by the correlation coefficients. Using the correlation distance Eq.(4) is calculated as follows;

$$d(f_s, \sum_{i=1}^{I} f_i) = \sum_{n=-N}^{N}(u_s r_{s,n} - \sum_{i=1}^{I} u_i r_{i,n})^2. \quad (8)$$

The problem to obtain $u_i (i = 1, \ldots, I)$ which minimize Eq.(8) is called quadratic programming problem. Here, the problem has a constraint considering non negativeness of u_i.

Without the constraint the least square method gives the solutions by the linear equation which is derived from partial derivatives of Eq.(8) by u_i setting 0.

$$\sum_j u_j \sum_n r_{j,n} r_{i,n} = u_s \sum_{n=-N}^{N} r_{s,n} r_{i,n}. \quad (9)$$

The matrix $\boldsymbol{R} = (R_{ij})$ is symmetric and its determinant $\det(\boldsymbol{R}) \geq 0$, where i, j elements R_{ij} is defined by $\sum_n r_{i,n} r_{j,n} (i, j = 1, \ldots, I)$. $\det(\boldsymbol{R}) = 0$ is equivalent that $\exists i; \boldsymbol{r}_i$ exists on a hyper plane determined by $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_I$, where $\boldsymbol{r}_i$ is excluded. For example, in case of two variables, $\boldsymbol{r}_1 \neq \boldsymbol{r}_2$ implies $\det(\boldsymbol{R}) > 0$. u_1, u_2 is determined by the derived linear equation.

To simplify the discussion the solution with the constraint in two dimensional case is described. The above solution for two dimension is calculated as follows;

$$\begin{cases} u_1 & = u_s \dfrac{R_{s1}R_{22} - R_{s2}R_{22}}{R_{11}R_{22} - R_{12}R_{21}} \\ u_2 & = u_s \dfrac{R_{11}R_{s2} - R_{21}R_{s1}}{R_{11}R_{22} - R_{12}R_{21}} \end{cases} \quad (10)$$

Criterion (Eq.8) is equal to a quadratic function of u_1, u_2, the surface of the criterion is elliptic in three dimensional space and there is always solution which gives the minimal value. Here, if $u_2 < 0$, the solution is given when the elliptic curve is touched to u_1 axis. This is equal to the quadratic function of u_1 with $u_2 = 0$ in the following $D(u_1, u_2)$ has a multiple root.

$$D(u_1, u_2) = \sum_{n=-N}^{N}(u_1 r_{1,n} + u_2 r_{2,n} - u_s r_{s,n})^2$$

$$\frac{\partial D(u_1, 0)}{\partial u_1} = 2\sum_{n=-N}^{N}(u_1 r_{1,n} - u_s r_{s,n}) r_{1,n}$$

$$u_1 \sum_{n=-N}^{N} r_{1,n} r_{1,n} = u_s \sum_{n=-N}^{N} r_{s,n} r_{1,n}$$

$$u_1 = u_s \frac{R_{s1}}{R_{11}}. \quad (11)$$

From the above discussion, spectral representatives do not require their power terms as power terms can be estimated using Eqs.(9) and (11).

4 Experiments for Evaluation

4.1 Speech database

200 different combinations are generated from randomly chosen 400 words in ATR 5240 word set. Silence intervals are removed using the phoneme label information. The duration length for the longer word is reduced to the length with the shorter word. In experiment 4.2 100 pairs of 200 words in the phoneme balanced 216 words are used as the training speech data.

4.2 Difference of overlapping on wave and correlation domains

Let the power of $x_1(t)$ and $x_2(t)$ be u_1 and u_2. The ratio of these powers is defined as $\alpha = \sqrt{\frac{u_1}{u_2}}$. Let βdB be the ratio of the power levels when $x_1(t)$ and $x_2(t)$ are overlapped. Its ratio on the wave domain can be calculated as $\gamma = 10^{\beta/20}$. $x_s(t)$ is calculated as follows;

$$x_s(t) = x_1(t) + \alpha\gamma x_2(t). \tag{12}$$

On the other hand, considering the relation between the correlation coefficients $r_{1,n}$ and $r_{2,n}$ of x_1, x_2, the overlapping in the correlation domain is as follows;

$$\tilde{r}_{s,n} = \frac{r_{1,n} + \gamma^2 r_{2,n}}{1 + \gamma^2}. \tag{13}$$

The value of LPC cepstrum distance is converted into dB scale using $d_{\mathrm{dB}}(g_1, g_2) = 4.342\ d_{\mathrm{CEP}}(g_1, g_2)$.

The overlapping and decomposition processing on correlation and wave domains is compared. Fig.1 shows the diagram of the experiment. At first, for each frame the power calculation, pre-emphasis and hamming windowing are processed, and each power is normalized. In order to adjust powers to the given ratio, `WAVE-2` is weighted as Eq.(12), and the overlapping on the waves and correlation analysis is performed in Fig.1 (A). The other path is pre-emphasized, Hamming windowed, and after correlation analysis and power normalization, weighting and overlapping is done in Fig.1 (B) as shown in Eq.(13). After LPC Cepstrum analysis the distance is calculated.

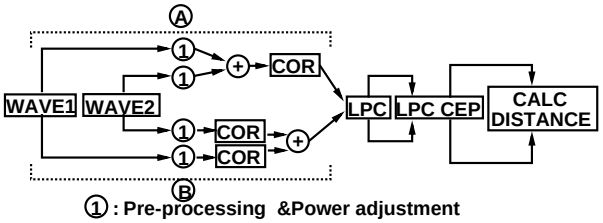


Figure 1: Procedure of Experiment

Comparing the complexity of mixed voice and single voice in spectral shapes, it seems that the former is larger than the latter. Based on the above assumption, LPC analysis order is varied from 12 to 24. Table 1 shows the experiment condition. As Fig.2 shows that the largest distance value in the result is of 0.8dB, where $\beta = 0$dB overlapping and 24th LPC analysis. The smallest distance value is 0.25dB, where $\beta = 40$dB overlapping and 12th LPC analysis. The result shows that the smaller the analysis order the smaller the distance value. The distance for two adjoining analysis orders, for example, 16th and 20th, in the same overlapping dB is about 0.02dB. The distance value in higher LPC analysis order becomes larger because higher order LPC analysis represents more precise frequency components. Lower order in LPC analysis is better, however, a higher order LPC analysis is applied in the following decomposition experiment because maintaining rich frequency components using higher order LPC analysis can carry the characteristics of input voice.

Table 1: Analysis conditions

speaker	1 male,1 female
pre-emphasis	$1 - 0.97z^{-1}$
auto correlation	24
LPC analysis	12,16,20,24
LPC cepstrum	36
window length	256 points(21.3ms)
window shift	256 points(21.3ms)
sampling frequency	12kHz

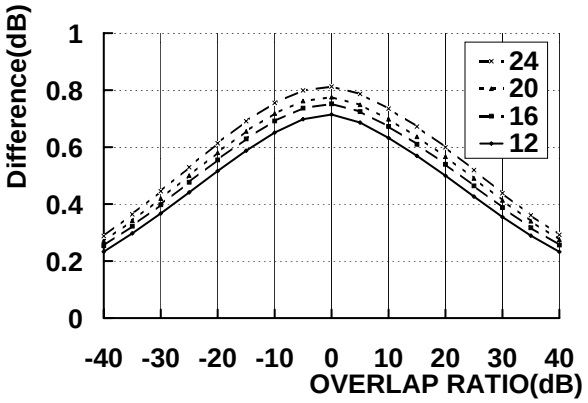


Figure 2: Effect of LPC analysis order

4.3 Accuracy of power estimation

The accuracy of the power estimation given by the least square method is examined. In this experiment the order of LPC analysis and the truncation order of LPC cepstrum are both 24.

Power estimation is evaluated, where a correlation coefficient vector is used the nearest in the codebook.

As the criterion, correlation distance and LPC cepstrum distance were compared. When LPC cepstrum distance is applied, the variance of estimated power value is large and sometimes the estimation failed. For correlation distance, power estimation works well within 10dB of power level. Power values sometimes become negative. Using correlation distance the fail rate of the power estimation is 3% for overlapped level β =0dB, and that is 17% for level β =10dB. The fail rate is always about 10% larger when the LPC cepstrum distance is used. Therefore, the correlation distance is better for the distance calculation in overlapped voice.

4.4 Decomposition using VQ codebook

The decomposition using VQ codebooks is evaluated. The distance is calculated after power estimation for all combination of correlation coefficients in two codebooks, and the optimal combination, f_{1i}, f_{2j}, is searched. The decomposed spectra are compared with the direct calculated spectra from voice. The comparison is done using LPC cepstrum distance. Analysis condition is the same as experiment 4.2 Fig.3 shows the result where the correlation orders are 16th and 24th. The difference is about 5.43dB for 24th correlation order and β =0dB overlapping level. In this condition the negative root case is 26.1%. Better result is achieved for 24th correlation order than 16th correlation order. It seems that the decomposition becomes easier for higher order because higher order spectra have much information. The other possible reason is that the accurate power estimation is possible for higher order. The reason why in Fig.3 the difference for `WAVE2` in 10dB overlapping level is smaller is that its power becomes 10 times larger and it is easier to catch the spectral feature and to decompose.

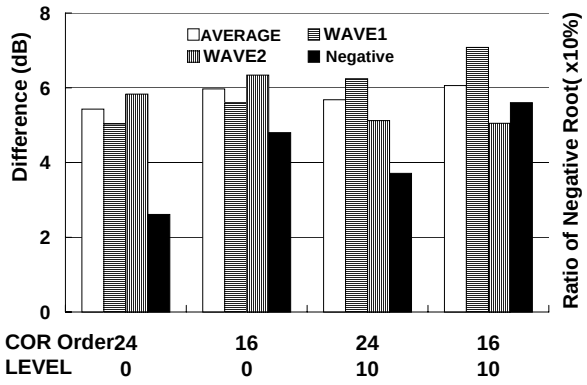


Figure 3: Result of decomposition experiment

5 Conclusion

The difference is calculated that distance of the processing in the correlation domain are examined and a wave domains. The experiment result demonstrated that processing on correlation domain is almost the same as one on wave domain. The power estimation from two correlation coefficients was examined. The difference between decomposed and original spectra is about 5dB. As the future studies, the solving method with constraint will be evaluated and the solution with time domain constraint will be formulated.

Acknowledgment

The author would like to thank to exciting discussion with K.Unayama, 1996 Graduation Research Student, who was doing experiments on this paper. On solution of quadratic problem, he also would like to thank to helpful comments from Prof.K.Funahashi, Univ. of Aizu.

References

[1] M.Sugiyama, et al, Speech Segmentation and Clustering Problem Based on An Unknown-multiple Signal Source Model, Trans. of IEICE, Vol.J76-D-II, No.12, pp.2477-2485 (Dec. 1993). (in Japanese)

[2] J.Murakami, et al, Study of Unknown-Multiple Signal Source Clustering Problem using Ergodic HMM, Trans. of IEICE DII, Vol.J78-D-II, No.2, pp.197-204 (Feb. 1995). (in Japanese)

[3] M.Sugiyama, Model Based Voice Decomposition Method, Tech. Rep. of Speech Research Committee, SP97-10, pp.1-8 (June 1997). (in Japanese)

[4] H.G.Okuno, et al, Evaluation of Sound Stream Segregation from the Viewpoint of Speech Recognition, IEICE Tech. Rep., SP95-86, pp.35-40 (Dec. 1995). (in Japanese)

[5] F.Martin, et al, Recognition of Noisy Speech by Using the Composition of Hidden Markov Models, Proc. of ASJ Fall Meeting, 1-7-10 (Oct. 1992).