AN ITERATIVE TRAJECTORY REGENERATION ALGORITHM FOR SEPARATING MIXED SPEECH SOURCES

S. W. Lee[1], Frank K. Soong[1, 2], and P. C. Ching[1]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Microsoft Research Asia, Beijing, China
{yswlee, pcching}@ee.cuhk.edu.hk, frankkps@microsoft.com

ABSTRACT

Hamonicity and continuity are two important perceptual cues for separating mixed speech sources. This paper focuses on the separation of two speech sources with a single-microphone input. An iterative, least-squares (LS) based trajectory regeneration algorithm is proposed to estimate the magnitude spectrum of each source. Time-derivatives of the spectrum, or the dynamic spectral information, is used as a constraint in solving the resultant weighted normal equations. Each estimated spectral trajectory, as a result, exhibits similar temporal variations as the original source. Asymptotically, we also prove that the regenerated trajectory yields the same time variations as the given dynamic information. When cascaded with our previously proposed harmonic filtering algorithm to separate mixed voiced signals, the new trajectory regeneration is shown to be very effective to reduce mean squared errors by 82.2% and 69.5%, relatively, with ideal and approximated dynamic information, respectively.[1]

1. INTRODUCTION

In real-life situations, sound signals often reach our ears as a mixture of target signal and background noise or competing speech. While human attends each individual component sounds quite easily even with only one ear [1, 2], the performance of most speech processing systems are easily degraded in this adverse condition. It is critical to extract individual sound sources from the input mixture prior to any further processes. This is referred as the source separation problem. This paper focuses on the single-microphone speech source separation. Source separation has been one of the popular research topics [2 - 5]. Two major approaches, independent component analysis (ICA) and computational auditory scene analysis (CASA), have received lots of attention. ICA utilizes the statistical properties between sources and the availability of several different input mixtures; while CASA studies the perceptual organization and mimics how human listeners segregate concurrent sounds. Hence, CASA is always possible to have the number of microphones less than the number of source signals.

In human perception, the input mixture is going through an auditory scene analysis (ASA) [2], which separates individual sound sources by looking at the regularity found in the input mixture. Sound components, which are likely to come from the same origin, are grouped together as one single source. There are four major regularities used, namely harmonicity, continuity, coherent changes taken in sound components and common onset and offset. For voiced speech, components having identical fundamental frequency (F0) are regularly spaced in frequency domain. This is the harmonicity regularity. Continuity refers to the phenomenon that a single sound changes its properties smoothly and slowly. Furthermore, abruptly changed sounds tend to perceptually exhibit closure, provided the properties before and after the discontinuity are matched. Coherent changes describe the finding that components from a specified source vary in amplitude and frequency in a coherent manner. As it is unlikely that distinct sources start and end at the same time, the auditory system tends to group components having identical onset and offset time together. It is believed that the auditory system not only uses acoustical information, but also high-level knowledge for actual sound segregation. The high-level knowledge includes feedback from recognition and predictions from what will be expected to hear [6].

A speech segregation system for mixtures with two sources has been previously proposed, which exploits the harmonic structure in voiced speech [7]. It is a recursive algorithm that finds an optimal pitch prediction error filter given either the input mixture or any periodic signal. At the first iteration, one of the F0s of two voiced sources is estimated (denotes it as $F0^1$) and by filtering the input mixture, the energy at corresponding harmonic frequencies is significantly suppressed. The output residual becomes the estimate source associated with $F0^2$. At the second iteration, this residual, which roughly contains one periodic signal (of $F0^2$) only, is then used to derive a pitch prediction error filter again and the output residual will be the other source associated with $F0^1$. The process iterates until estimates converge. From the experimental results, this harmonicity segregation system works well for synthetic speech; for real speech, the performance is not satisfactory that the output estimate still contains some residual from the interfering source. This may due to the properties below: (1) real speech is neither ideally periodic nor driven by impulse train and the way that energy concentrated in harmonic structure is wider like a kernel, rather than an impulse in synthetic speech; and (2) energy is not exactly located at multiples of F0, especially in high frequency region.

In this paper, an iterative trajectory recovery algorithm is proposed by using both continuity and expectation from dynamic information. The spectral trajectory represents the continuity

[1] This work is partially supported by a grant awarded by the Hong Kong Research Grants council. The work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

examined and the dynamic information applied is the delta coefficient calculated on the log spectral magnitude. After passing through the harmonic filtering in [7], the spectral trajectories of output estimate will then be recovered by following the expected delta shape in an appropriate manner. The proposed regeneration algorithm is supported by both mathematical proof and experimental results. In particular, the performance with imperfect delta coefficient (using piecewise-constant approximation or three-level quantization) is also studied and most of the residual of interfering speech is removed by inspecting the mean-square-error (MSE) in estimation.

2. TRAJECTORY REGENERATION

Dynamic spectral information, namely the velocity and acceleration features, has been used successfully for automatic speech or speaker recognition [8, 9]. The dynamic spectral information is obtained over a time window by linear regression [10]. As a result, the dynamic information of a clean source signal should be relatively continuous and smoothly varying. Giving this dynamic information, the spectral movement and the corresponding direction are known. This represents the expected spectral trajectory. In the following, the detailed descriptions of the proposed trajectory regeneration algorithm using dynamic information will be given.

2.1. Formulation

The input mixture signal *x(n)* is related to the source signals $x_1(n)$ and $x_2(n)$ as

$$x(n) = x_1(n) + x_2(n) \tag{1}$$

Let $x_1'(n)$ and $x_2'(n)$ denote the two intermediate estimated source signals given at the output of the harmonic filtering system. Fig. 1 illustrates the block-diagram of the overall separation system, where $x_1''(n)$ and $x_2''(n)$ are the resultant estimates.

The trajectory regeneration process is carried in the log magnitude spectral domain. Same procedure is applied to both $x_1'(n)$ and $x_2'(n)$. Let $x_i'(n)$ be one of the intermediate estimate ($i\in$[1, 2]). $x_i'(n)$ is first windowed into frames and their corresponding short-time complex spectra $X_i'(k, n)$ at frequency bin *k* and frame *n* are computed. The magnitude is further converted to,

$$y_{ik}(n) = 10\log\left(\left|X_i'(k,n)\right|\right) \tag{2}$$

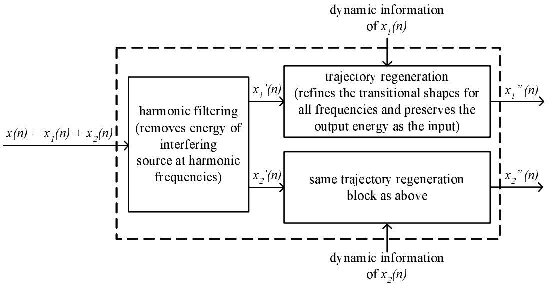
dynamic information
of $x_1(n)$
$x(n) = x_1(n) + x_2(n)$
harmonic filtering (removes energy of interfering source at harmonic frequencies)
$x_1'(n)$
$x_2'(n)$
trajectory regeneration (refines the transitional shapes for all frequencies and preserves the output energy as the input)
same trajectory regeneration block as above
$x_1''(n)$
$x_2''(n)$
dynamic information
of $x_2(n)$

Fig. 1. Block-diagram of the overall separation system. $x_1'(n)$ and $x_2'(n)$ enter two identical trajectory regeneration blocks apart.

For each frequency bin, perform the following. Let $\mathbf{y_{ik}} = [y_{ik}(1), y_{ik}(2),\ldots, y_{ik}(N)]^T$ be the observed magnitude trajectory at frequency bin *k*, for *k* = 1, 2, …, *K*, where *K* and *N* represent the numbers of bins and frames respectively. *T* denotes the transpose operation. It is assumed that the ideal delta coefficient $\Delta\mathbf{y_{ik}}$ ($\Delta\mathbf{y_{ik}} = [\Delta y_{ik}(1), \Delta y_{ik}(2),\ldots, \Delta y_{ik}(N)]^T$) is given, but not derived by $\mathbf{y_{ik}}$. In practice, this assumption should be relaxed and the performance with imperfect $\Delta\mathbf{y_{ik}}$ will be reported in later section. As the delta coefficient is associated to the static counterpart by linear regression [10], we have

$$\begin{aligned}&\begin{bmatrix}\mathbf{I}\\ \mathbf{W}\end{bmatrix}\left[y_{ik}'(1)\ y_{ik}'(2)\ \ldots\ y_{ik}'(N)\right]^T\\ &=\left[y_{ik}(1)\ y_{ik}(2)\ \ldots\ y_{ik}(N)\ \Delta y_{ik}(1)\ \Delta y_{ik}(2)\ \ldots\ \Delta y_{ik}(N)\right]^T\end{aligned} \tag{3}$$

where $y_{ik}'(n)$, **I** and **W** represent the unknown static trajectory components to be found, the identity matrix and the linear regression coefficient matrix respectively. In a matrix form,

$$\begin{bmatrix}\mathbf{I}\\ \mathbf{W}\end{bmatrix}\mathbf{y_{ik}}'=\begin{bmatrix}\mathbf{y_{ik}}\\ \Delta\mathbf{y_{ik}}\end{bmatrix} \tag{4}$$

The least squares solution is,

$$\begin{aligned}\mathbf{y_{ik}}'&=\left\{\begin{bmatrix}\mathbf{I} & \mathbf{W}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{I}\\ \mathbf{W}\end{bmatrix}\right\}^{-1}\begin{bmatrix}\mathbf{I} & \mathbf{W}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{y_{ik}}\\ \Delta\mathbf{y_{ik}}\end{bmatrix}\\ &=\left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\begin{bmatrix}\mathbf{I} & \mathbf{W}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{y_{ik}}\\ \Delta\mathbf{y_{ik}}\end{bmatrix}\end{aligned} \tag{5}$$

Both static and dynamic trajectories impose constraints in the estimation of $\mathbf{y_{ik}}'$. Referring to Equation (4), $\mathbf{y_{ik}}'$ has to be unchanged as the observed $\mathbf{y_{ik}}$, but attains the transitional shape as $\Delta\mathbf{y_{ik}}$ simultaneously. It is believed that the coarse power levels of individual sources have been appropriately adjusted by the preceding harmonic filtering, as energy of interfering source at harmonic frequencies is greatly reduced. This is maintained by the constraint from $\mathbf{y_{ik}}$. Nevertheless, the separation process in harmonic filtering is done in a frame basis. Adjacent frames are estimated alone and hence, for continuity, the spectral trajectory needs to be refined with the given $\Delta\mathbf{y_{ik}}$.

The above recovery scheme is iterated by substituting $\mathbf{y_{ik}}'$ back in Equation (5) as $\mathbf{y_{ik}}$ to obtain a new estimate, until $\mathbf{y_{ik}}'$ converges or a certain number of iterations. By converting $\mathbf{y_{ik}}'$ to linear scale and combining with phase in $X_i'(k, n)$, the output speech is finally reconstructed by the overlap-add method.

2.2. Proof

The above iterative trajectory regeneration process is investigated to find: as the iteration proceeds, if the proposed method leads to identical trajectory as the given $\Delta\mathbf{y_{ik}}$ or not. Equation (5) shows the estimated $\mathbf{y_{ik}}'$ at the first iteration. At the second iteration,

$$\begin{aligned}\mathbf{y_{ik}}'&=\left\{\begin{bmatrix}\mathbf{I} & \mathbf{W}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{I}\\ \mathbf{W}\end{bmatrix}\right\}^{-1}\begin{bmatrix}\mathbf{I} & \mathbf{W}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\begin{bmatrix}\mathbf{I} & \mathbf{W}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{y_{ik}}\\ \Delta\mathbf{y_{ik}}\end{bmatrix}\\ \Delta\mathbf{y_{ik}}\end{bmatrix}\\ &=\left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-2}\begin{bmatrix}\mathbf{I} & \mathbf{W}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{y_{ik}}\\ \Delta\mathbf{y_{ik}}\end{bmatrix}+\left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{W}^{\mathrm{T}}\Delta\mathbf{y_{ik}}\end{aligned} \tag{6}$$

At the third iteration,

$$\begin{aligned}\mathbf{y_{ik}}'&=\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-3}\begin{bmatrix}\mathbf{I} & \mathbf{W}^{T}\end{bmatrix}\begin{bmatrix}\mathbf{y_{ik}}\\ \Delta\mathbf{y_{ik}}\end{bmatrix}+\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-2}\mathbf{W}^{T}\Delta\mathbf{y_{ik}}\\ &+\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-1}\mathbf{W}^{T}\Delta\mathbf{y_{ik}}\end{aligned} \tag{7}$$

Hence, the proposition below is proposed for the m^{th} iteration,

$$\mathbf{y_{ik}}' = \left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-m}\left[\mathbf{I}\quad \mathbf{W}^{\mathrm{T}}\right]\begin{bmatrix}\mathbf{y_{ik}}\\ \Delta\mathbf{y_{ik}}\end{bmatrix} + \left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-(m-1)}\mathbf{W}^{\mathrm{T}}\Delta\mathbf{y_{ik}}$$
$$+\ldots+\left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-2}\mathbf{W}^{\mathrm{T}}\Delta\mathbf{y_{ik}} + \left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{W}^{\mathrm{T}}\Delta\mathbf{y_{ik}} \quad (8)$$
$$= \left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-m}\mathbf{y_{ik}} + \left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-m}\mathbf{W}^{\mathrm{T}}\Delta\mathbf{y_{ik}} + \ldots + \left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{W}^{\mathrm{T}}\Delta\mathbf{y_{ik}}$$

In the following, this proposition is proved by using mathematical induction. The case for *m* = 1 is identical to Equation (5). Assume that when *m* = *p*,

$$\mathbf{y_{ik}}' = \left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-p}\left[\mathbf{I}\quad \mathbf{W}^{\mathrm{T}}\right]\begin{bmatrix}\mathbf{y_{ik}}\\ \Delta\mathbf{y_{ik}}\end{bmatrix} + \left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-(p-1)}\mathbf{W}^{\mathrm{T}}\Delta\mathbf{y_{ik}} \quad (9)$$
$$+\ldots+\left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-2}\mathbf{W}^{\mathrm{T}}\Delta\mathbf{y_{ik}} + \left(\mathbf{I}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{W}^{\mathrm{T}}\Delta\mathbf{y_{ik}}$$

At the (*p* + 1)th iteration,

$$\mathbf{y_{ik}}' = \left\{\left[\mathbf{I}\ \mathbf{W}^{T}\right]\begin{bmatrix}\mathbf{I}\\ \mathbf{W}\end{bmatrix}\right\}^{-1}\left[\mathbf{I}\ \mathbf{W}^{T}\right]\begin{bmatrix}\mathbf{y_{ik}}'\textbf{(in equation (9))}\\ \Delta\mathbf{y_{ik}}\end{bmatrix}$$
$$= \left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-(p+1)}\left[\mathbf{I}\ \mathbf{W}^{T}\right]\begin{bmatrix}\mathbf{y_{ik}}\\ \Delta\mathbf{y_{ik}}\end{bmatrix} + \left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-p}\mathbf{W}^{T}\Delta\mathbf{y_{ik}} \quad (10)$$
$$+\ldots+\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-2}\mathbf{W}^{T}\Delta\mathbf{y_{ik}} + \left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-1}\mathbf{W}^{T}\Delta\mathbf{y_{ik}}$$

The proposition in Equation (8) is proved. With this general expression, the following examine how $\mathbf{y_{ik}}'$ behaves if $m \to \infty$.

$$\mathbf{y_{ik}}' = \left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-m}\mathbf{y_{ik}} + \left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-m}\mathbf{W}^{T}\Delta\mathbf{y_{ik}}$$
$$+\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-(m-1)}\mathbf{W}^{T}\Delta\mathbf{y_{ik}} + \ldots \quad (11)$$
$$+\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-2}\mathbf{W}^{T}\Delta\mathbf{y_{ik}} + \left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-1}\mathbf{W}^{T}\Delta\mathbf{y_{ik}}$$

Using sum of geometric progression, Equation (11) becomes:

$$\mathbf{y_{ik}}' = \left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-m}\mathbf{y_{ik}} + \left[\mathbf{I}-\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)\right]^{-1}\left[\mathbf{I}-\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{m}\right]\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-m}\mathbf{W}^{T}\Delta\mathbf{y_{ik}} \quad (12)$$
$$= \left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-m}\mathbf{y_{ik}} + \left[-\mathbf{W}^{T}\mathbf{W}\right]^{-1}\left[\left(\mathbf{I}+\mathbf{W}^{T}\mathbf{W}\right)^{-m} - \mathbf{I}\right]\mathbf{W}^{T}\Delta\mathbf{y_{ik}}$$

When $m \to \infty$,

$$\mathbf{y_{ik}}' = \left[-\mathbf{W}^{T}\mathbf{W}\right]^{-1}\left[-\mathbf{I}\right]\mathbf{W}^{T}\Delta\mathbf{y_{ik}} = -\left[-\mathbf{W}^{T}\mathbf{W}\right]^{-1}\mathbf{W}^{T}\Delta\mathbf{y_{ik}} \quad (13)$$

We assume the linear regression coefficient matrix **W** is symmetric, which is approximately true except for a few boundary frames,

$$\mathbf{y_{ik}}' = -[-\mathbf{W}^{-2}]\mathbf{W}\Delta\mathbf{y_{ik}} = \mathbf{W}^{-1}\Delta\mathbf{y_{ik}} \quad (14)$$
$$\mathbf{W}\mathbf{y_{ik}}' = \Delta\mathbf{y_{ik}} \quad (15)$$

As the condition number of **W** is usually very large, $\mathbf{W}^{-1}$ cannot be accurately found. As a result, as shown in Equation (15), when iteration goes to infinity, the estimated $\mathbf{y_{ik}}'$ from the proposed algorithm generates a spectral trajectory which has its spectral dynamic information the same as $\Delta\mathbf{y_{ik}}$.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

x₁(n)	content: /ɑː//iː//uː//e//ɔː/ (in international phonetic alphabet)
	average power: 45 dB
	sampling frequency: 8 kHz
x₂(n)	content: We were away a year ago.
	average power: 45 dB
	Sampling frequency: 8 kHz

Table 1. Details of the speech samples.

The proposed trajectory regeneration algorithm is first verified with the ideal dynamic information in Section 3.1. In Section 3.2, imperfect dynamic information is used to check the regeneration performance under practical situations. Real speech samples are recorded for the experiments and their details are listed in Table 1.

3.1. Trajectory regeneration with ideal dynamic information

By using the source signal $x_i(n)$, the ideal $\Delta\mathbf{y_{ik}}$ is found. Fig. 2 shows an example of the recovered magnitude trajectories for $x_2(n)$. This trajectory is taken from frequency around 1560 Hz (bin 50 out of 256 bins in total). The number of iteration is 50.

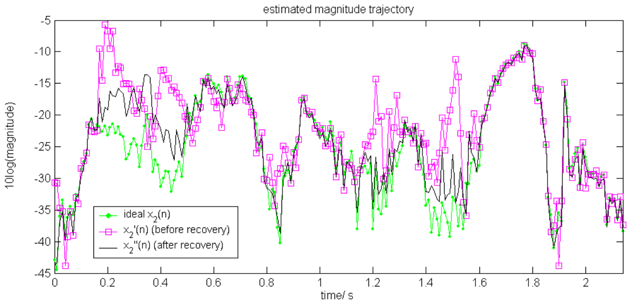

Fig. 2. Magnitude Trajectories of ideal $x_2(n)$, $x_2'(n)$ (before recovery) and $x_2''(n)$ (after recovery).

The MSE in log magnitude at various stages of the proposed separation process are measured and listed in Table 2.

stage of separation process		MSE (dB²)
input stage (before any separation process)		39.30
after harmonic filtering		28.28
after trajectory regeneration	*m* = 1	25.77
	m = 5	15.59
	m = 50	6.99

Table 2. MSE before and after a given number of trajectory regeneration iterations (with ideal dynamic information).

Adjusting the spectral transition according to the ideal delta coefficients, the MSE is reduced by 34.4% from 39.3 dB² to 28.28 dB² and 25.77 dB² after harmonic filtering and the first iteration respectively and further attained to 82.2% (6.99 dB²) after 50 iterations. The result obtained just after harmonic filtering represents the expected performance from the system in [7] for this real speech sample. Referring to Fig. 2, the shape of the estimated trajectory is highly similar to the ideal trajectory. This shows the proposed regeneration algorithm helps to generate output speech varies in a manner similar to what pure source signals behave and with good continuity. Since $\mathbf{y_{ik}}'$ has to maintain the power before and after regeneration alike and achieve an appropriate transitional shape similar to $\Delta\mathbf{y_{ik}}$ at the same time, it is certainly that even trajectory has been properly steered, there are occasionally some samples quite different from the ideal values (between 0.2 to 0.5 s). This illustrates the intrinsic property of dynamic information that only transitions are concerned, but not the actual magnitude levels. Thus, it is necessary to have both the coarse estimation from harmonic filtering and refinement from trajectory regeneration. Concerning the output speech quality, it is highly natural and no distinctive distortions can be heard.

3.2. Trajectory regeneration with imperfect dynamic information

Ideal dynamic information is not always available, especially in adverse environments. Using imperfect delta coefficients can illustrate how the proposed trajectory regeneration algorithm performs under more practical situations. Two kinds of imperfect dynamic information have been tried. The ideal $\Delta\mathbf{y_{ik}}$ is approximated by either (1) a mean value of $\Delta\mathbf{y}_{ik}$ for each segmented trajectory or (2) one of three quantization levels after clustering. The details of the two approximations will be given below.

Piecewise-constant approximation: The observed $\mathbf{y}_{ik}$ is passed through an auto-segmentation module and partitioned into a given number of homogeneous segments. All the ideal delta coefficients within a segment are replaced by the corresponding mean value. The auto-segmentation module uses dynamic programming to select the optimal partition, which generates the smallest distortion resulted from replacing magnitude values with the associated means. The distortion is defined as the Euclidean distance and the number of frames within a segment must be larger than a minimum segmental length. With the mean and variance statistics from the segmentation result, maximum-likelihood estimation (MLE), so as to make use of both the mean approximation and the variances of different segments, is used to find out $\mathbf{y}_{ik}'$.

Three-level quantization: As both delta and acceleration coefficient are the time difference of the trajectory samples, they are zero-mean. In addition, the sign given in the dynamic value plays an important role, as it controls whether the trajectory should go up or down. Hence, the ideal $\Delta\mathbf{y}_{ik}$ is clustered into three groups by k-means clustering and each ideal $\Delta\mathbf{y}_{ik}$ value is then quantized to one of the three centroids. Finally, the magnitude trajectory is estimated by Equation (5) with these quantized $\Delta\mathbf{y}_{ik}$.

stage of separation process		MSE (dB^2)
input stage (before any separation process)		39.30
after harmonic filtering		28.28
after trajectory regeneration	*m* = 1	23.18
	m = 5	17.40
	m = 50	17.19

Table 3. MSE before and after a given number of trajectory regeneration iterations (after piecewise-constant approximation).

stage of separation process		MSE (dB^2)
input stage (before any separation process)		39.30
after harmonic filtering		28.28
after trajectory regeneration	*m* = 1	26.07
	m = 5	20.67
	m = 50	11.99

Table 4. MSE before and after a given number of trajectory regeneration iterations (after three-level quantization).

Table 3 and 4 show the MSE in log magnitude of the two approximations. Comparing the results in Table 2 and 3, the iteration process in MLE with piecewise-constant approximation saturates soon and only 56.3% of the total MSE is eliminated after 50 iterations. It is found that, however, the MSE obtained after the first iteration is the lowest. This demonstrates the superiority of MLE over the standard least-square solution when $\mathbf{y}_{ik}$ or delta coefficients have different reliabilities. More weights (small variances) are put on those reliable data. Besides, the performance of the auto-segmentation module is critical that the segmentation boundaries are highly sensitive to the total number of segments and the minimum segmental length. In order to have accurate estimate of variances, each segments cannot be too short.

Regarding the estimation after three-level quantization, the MSE in log magnitude is satisfactory that it is reduced to 11.99 dB^2 (69.5%) finally. Comparing with Table 2, the MSE values are slightly higher, but the transitional direction preserved after quantization is shown to be an effective cue for trajectory recovery. Comparing the results in Table 3 and 4, the three-level quantization is much better than the piecewise-constant approximation and a lower MSE can be achieved, if more iterations are allowed. The output speech from piecewise-constant approximation contains some distortion, while there is hardly any distortion perceived in the output speech from the former approach. Furthermore, no parameters are involved in three-level quantization.

4. CONCLUSIONS

A trajectory regeneration algorithm for separating mixed speech sources in one single microphone is proposed. The complete system includes: harmonic filtering and trajectory regeneration. Trajectory regeneration refines the spectral power obtained in harmonic filtering. The dynamic spectral information is used as a constraint in regenerating a continuous spectral trajectory iteratively. The dynamic information represents the expected spectral changes along time and this can be one of the feedback information from speech recognition. The iterative regeneration process is proved to converge asymptotically to a trajectory which has the same dynamic information. Experimental results also show that significant amount of interfering energy is removed, even with imperfect dynamic information.

5. REFERENCES

[1] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Amer.*, vol. 25, pp. 975-979, Sep. 1953.

[2] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, London: The MIT Press, 1990.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York: John Wiley & Sons, Ltd, 2001.

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Application*, England: John Wiley & Sons, Ltd, 2002.

[5] G. J. Brown and M. Cooke, “Computational auditory scene analysis,” *Computer Speech and Language*, vol. 8, pp. 297-336, Oct. 1994.

[6] M. Weintraub, “A theory and computational model of auditory monaural sound separation,” Ph.D. dissertation, Stanford University, 1985.

[7] S. W. Lee, F. K. Soong, and P. C. Ching, “Harmonic filtering for joint estimation of pitch and voiced source with single-microphone input,” *Proc. Eurospeech*, pp. 309-312, Sep. 2005.

[8] S. Furui, “Speaker-independent isolated word recognition using dynamic features of speech spectrum,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 52-59, Feb. 1986.

[9] F. K. Soong and A. E. Rosenberg, “On the use of instantaneous and transitional spectral information in speaker recognition,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 871-879, Jun. 1988.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.1)*, Cambridge University, 2001.