AN INFORMATION THEORETIC APPROACH FOR SPEECH SOURCE ENUMERATION

David Ayllón[1]*, Roberto Gil-Pita*[1]*, Manuel Rosa-Zurera*[1] *and Hamid Krim*[2]

[1]Department of Signal Theory and Communications, University of Alcala, Spain
[2]Department of Electrical and Computer Engineering, North Carolina State University, NC

ABSTRACT

The solution of speech related problems such as source location or separation relies on a prior estimation of the number of sources. In this paper we propose a method for speech source enumeration based on the different relative delays that sources at different locations register at two microphones. The Probability Density Function (PDF) of the estimated delays exhibits peaks associated with each source. The Minimum Description Length (MDL) criterion is applied to the prediction error of a linear model fitted to the delay estimates. The method is validated for the estimation of different number of sources and different mixtures.

***Index Terms*—** Source enumeration, Array signal processing, Microphone array, Speech source separation.

1. INTRODUCTION

Speech source enumeration is a problem that remains largely open and unsolved. Determining the number of speech sources is a critical first step when solving other speech-related problems such as source location or the so-called Blind Source Separation (BSS), where many of the proposed algorithms assume the number of sources known in advance.

There are many different approaches to signal enumeration, and those based on information theoretic criteria have largely been used in array signal processing [1]. Two such criteria for order estimation of an observed process are the Akaike Information Criterion (AIC) [2] and the Rissanen's Minimum Description Length (MDL) principle [3], which have inspired many algorithms to solve the aforementioned problem, for instance, [4, 5, 6]. Most of those algorithms have been applied to problems where the relative bandwidth of the signals is low, such as radar, sonar or mobile communications. The wideband nature of speech requires a different approach. Furthermore, the information theoretic approach is normally reduced to the over-determined case.

Prior works in the field explore different approaches for speech source enumeration. In [7], a novel method for source enumeration and location in underdetermined multichannel mixtures is presented. The method identifies time-frequency (T-F) regions belonging to a predominant source to enumerate successfully speech sources in instantaneous mixtures and anechoic mixtures with small delays. However, it fails when the delay exceeds one sample as well as in the case of echoic mixtures. Another approach based on pitch estimation has been applied, for instance, in [8, 9]. Unfortunately, multi-pitch estimation is not very accurate due to the proximity and fluctuations of the pitches of different speech sources, and the performance degrades rapidly in the presence of noise or aperiodicity. Finally, information criteria has also been combined with multi-pitch estimation, for instance, in [10].

The algorithm proposed in this paper considers a two-microphone array and assumes a different Direction Of Arrival (DOA) for each of the speech sources. This implies that each impinging source incurs a different relative delay at the two microphones. Under the assumption of sparse representation of speech in the T-F domain [11], an instantaneous source delay estimator is computed. The Probability Density Function (PDF) of the estimates will include peaks associated to the different source delays. Consequently, the problem of speech source enumeration is equivalent to count the number of peaks in this PDF. We will use a linear model to determine the number of peaks, obtained from the PDF spectral series. The proposed linear model will reflect the modes of the PDF of the delays, and will subsequently be used in conjunction with the MDL criterion to minimize the coding length of the spectral series. In contrast to other algorithms that apply the MDL principle to observed data with an associated distribution, our proposed method seeks to capture the peaks of the PDF of the delays (much like one seeks spectral peaks of time series) by analyzing the *spectral series* of the PDF by considering its characteristic function [12].

2. DELAY-BASED ENUMERATION ALGORITHM

2.1. Instantaneous delay estimator

Let us consider a mixture of N speech sources recorded by a microphone array composed of two elements. Without loss of generality for the echoic case, the anechoic mixing model is defined by:

This work has been funded by the Spanish Ministry of Science and Innovation, under project TEC2009-14414-C03-03 and the scholarship AP2009-3932.

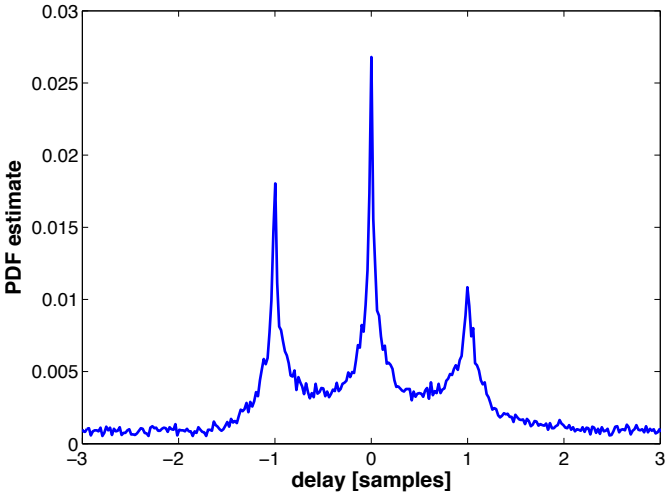

Fig. 1. Histogram-based PDF estimate of $\hat{\delta}$ in the case of an anechoic mixture of 3 speech sources with source delays of [-1, 0, 1] samples.

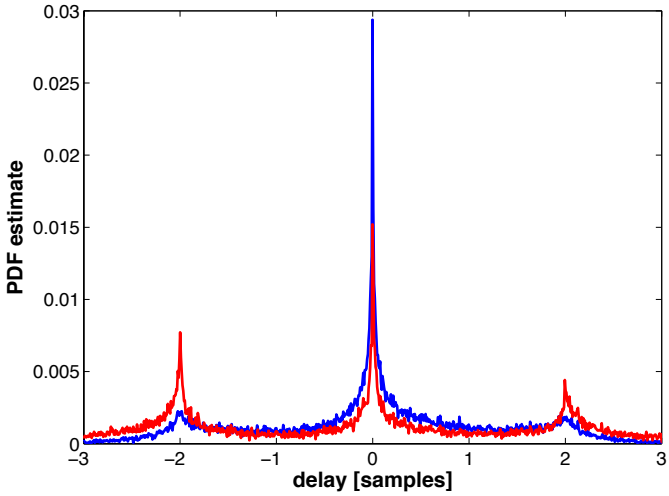

Fig. 2. Histogram-based PDF estimate of $\hat{\delta}$ with phase ambiguity (blue line) and with the phase unwrapped (red line) in the case of an anechoic mixture of 3 speech sources with source delays of [-2, 0, 2] samples.

$$y_m(t) = \sum_{j=1}^{N} \alpha_{mj} \cdot s_j(t - \Delta_{mj}),\ m = 1, 2 \qquad (1)$$

where y_m(t) are the microphone signals, $s_j(t)$ are the original speech sources, and α_{mj} and Δ_{mj} are the corresponding level and time differences respectively. The sampled signals in the discrete-time domain are: $y_m[n] = \sum_{j=1}^{N} \alpha_{mj} \cdot s_j(n - \delta_{mj})$, where δ_{mj} are the normalized delays with respect to the sampling period. Considering the first element of the array as a reference implies $\alpha_{1j} = 1$ and $\delta_{1j} = 0$, for $j = 1, ..., N$. In addition, we rename α_{2j} as α_j and δ_{2j} as δ_j. Taking the Short-Time Fourier Transform (STFT) of the discrete-time signals, the mixing model (1) in the T-F domain is

$$\begin{bmatrix} Y_1(k,l) \\ Y_2(k,l) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ \alpha_1 e^{-i\omega\delta_1} & ... & \alpha_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} S_1(k,l) \\ ... \\ S_N(k,l) \end{bmatrix}, \qquad (2)$$

where $k = 1, ..., K$ is the frequency index, and $l = 1, ..., L$ is the time index, K is the number of frequency bands and L is the number of time frames. Considering sources that fulfill the condition of approximate W-Disjoint Orthogonality (WDO) introduced in [13], i.e. a non-overlapping representation of the sources in the T-F domain, the instantaneous delays may be estimated according to:

$$\hat{\delta}(k,l) = -\frac{1}{\omega} arg\left\{ \frac{Y_2(k,l)}{Y_1(k,l)} \right\}. \qquad (3)$$

The fact that the delay changes with the position of the source together with the assumption of WDO sources, yield the probability density function (PDF) of the delay estimates whose peaks are associated with each of the sources. Figure 1 shows a histogram-based PDF estimate of the PDF of $\hat{\delta}$ in the case of an anechoic mixture of 3 speech sources with a delay of one sample between them. According to this, we propose to solve the problem of speech source enumeration counting the number of modes in the PDF of the delay estimator.

The estimator in (3) can be ambiguous due to the periodicity of the phase, and it is reliable only if $|\omega\delta_j| < \pi$, condition that is guaranteed when $\omega_{max}\delta_{jmax} < \pi$, which means that for relative delays between microphones larger than one sample, the estimated phase will be inaccurate. In order to overcome this limitation, we unwrap the radian phase of $\omega\delta_j$ by changing the absolute phase jumps greater to or equal to π to their 2π complement. This operation is performed along the frequencies of each frame. Once the frequency term is removed from the unwrapped phase, the delay estimate $\hat{\delta}_j$ reflects the true values even when they are larger than one sample. An illustrative example is shown in Figure 2, where the delays between the three speech sources of an anechoic mixture are set to two samples. In the PDF estimate represented with a blue line, the phase ambiguity has not been resolved, and the peaks corresponding to sources with delays of two samples are barely perceived. However, in the PDF estimate represented with a red line, where the phase has been unwrapped, the three peaks are clearly identifiable.

Finally, the delay estimates that have been estimated from T-F bins with low energy are not consistent, and they are consequently removed. The energy of a T-F bin is measured with the geometric mean of the energy of the signals at both microphones

$$E(k,l) = 10log_{10}(|Y_1(k,l)| \cdot |Y_2(k,l)|), \quad (4)$$

rejecting delay estimates from T-F points where $E(k,l) < Th$, where Th is the threshold. The number of delays rejected is R.

2.2. Parametric model-based PDF estimation

Let us consider the vector $\mathbf{d} = [\delta_1, ..., \delta_Q]$ containing all the delay estimates within every T-F bin, where $Q = K \cdot L - R$ is the length of $\mathbf{d}$. The information regarding the number of sources is contained in the PDF of the random variable δ, which is denoted by $f(\delta)$. As noted earlier, using the PDF of the delay estimates and its dual (the characteristic function), we construct a linear model which we exploit to explore the coding length of the dual of the PDF as we elaborate next. With the assumption that $f(\delta) = 0$ if $|\delta| > \pi$, we proceed to obtain the spectral sequence $\phi_\delta[m]$, defined as:

$$\phi_\delta[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\delta m} f(\delta) d\delta = \frac{1}{2\pi} E\{e^{i\delta m}\}. \quad (5)$$

Using the sample mean as an estimator of the probabilistic expectation, the sequence $\phi_\delta[m]$ can be estimated with the next expression:

$$\hat{\phi}_\delta[m] = \frac{1}{2\pi Q}\sum_{q=1}^{Q} e^{i\delta_q m}. \quad (6)$$

A linear predictive model of order P can be computed from $\hat{\phi}_\delta[m], m = 0, ...P$, by using the pseudo-inverse solution. This model is in some sense similar to an AR model. If the previous assumption ($f(\delta) = 0$ if $|\delta| > \pi$) is not fulfilled, the delays should be normalized before applying this method. Figure 3 shows an example of the aforementioned PDF estimation using the proposed linear model. The blue line represents the histogram-based PDF estimate of the source delays corresponding to a linear anechoic mixture of 4 speech sources, introducing delays of [-1, 0, 1, 2] samples between sources. The red line represents the linear-model estimation with $P = 4$, which clearly depicts the four peaks of the PDF.

2.3. Application of MDL for enumeration

The prediction error related to fitting the linear predictive model to the data is a monotonically decreasing function of the order model P. However, from a certain value of P, the linear model fits with sufficient accuracy the true PDF. In this context, the MDL method suggests choosing the model that provides the shortest description of our data, then considering that the order of that model is an estimation of the number of speech sources in the mixture. Additionally, encoding the prediction error is equivalent to encoding the best representation of the data [12].

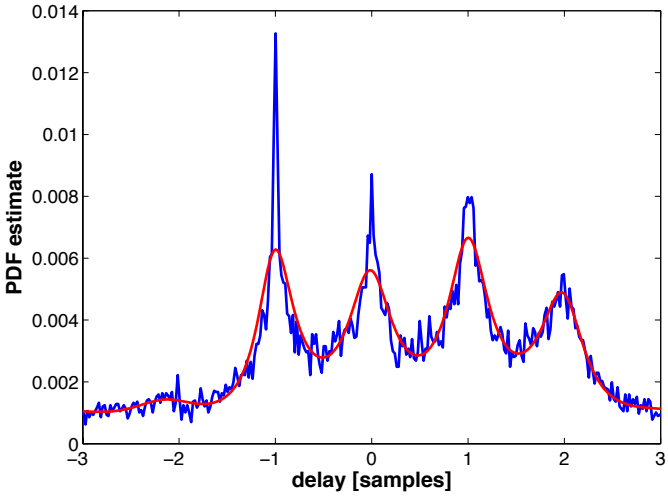

Fig. 3. Histogram-based PDF estimate (blue line) and 4^{th}-order linear-model estimate (red line) of $\hat{\delta}$ in the case of an anechoic mixture of 4 speech sources with source delays of [-1, 0, 1, 2] samples.

In the linear predictive model, the coefficients may be written as,

$$\hat{\phi}_\delta[m] = \sum_{p=1}^{P} a_p \hat{\phi}_\delta[m-p] + \sigma_\epsilon^2 \delta_K[m], \quad 0 \le m \le P, \quad (7)$$

where $\hat{\phi}_\delta[m]$ denotes the estimates with expression (6), σ_ϵ^2 is the variance of the model input random process ($\epsilon(n)$), and $\delta_K[m]$ is the Kronecker delta. For $m = 0$, the value of σ_ϵ^2 may be easily obtained. The model input random process is assumed to be a Gaussian random process with zero mean and variance σ_ϵ^2:

$$f(\epsilon) = \frac{1}{\sigma_\epsilon\sqrt{2\pi}} e^{-\frac{\epsilon^2}{2\sigma_\epsilon^2}}. \quad (8)$$

We are interested in determining the model order P used to estimate the PDF. Rissanen in [3] proposed the minimal code length required to describe the observed data and the free parameters (model parameters) as a general criterion for model order determination. The number of bits needed to encode the data determines the selected model. The choice of an estimator that achieves the shortest total code length via the MDL criterion is formulated as:

$$MDL(p) = -log(f(\mathbf{y}|\mathbf{a}) + \frac{p}{2} log(Q), \quad (9)$$

where $\mathbf{y}$ is the vector composed of the data of the spectral series (related to the delays), and $\mathbf{a}$ are the free parameters. Considering the fact that $f(\mathbf{y}|\mathbf{a}) = f(\epsilon(n))$, and introducing (8) into (9), we obtain after some manipulations:

Table 1. Source delays for linear mixtures

N	delays (samples)
2	[-1, 0]
3	[-1, 0, 1]
4	[-1, 0, 1, -2]
5	[-1, 0, 1, -2, 2]

$$MDL(p) = Qlog(\pi) + Qlog(\frac{1}{Q}||\epsilon_p||^2) + \frac{p}{2}log(Q), \quad (10)$$

where ϵ_p is the prediction error corresponding to the linear predictive model of order p. Finally, the number of speech sources is given by

$$\hat{p} = \min_{p \in \{1,\dots,P\}} MDL(p) \quad (11)$$

3. EXPERIMENTS

The algorithm proposed has been tested over 50 different anechoic speech mixtures of 2, 3, 4 and 5 sources. The speech sources are randomly selected from the TIMIT database [15]. The T-F decomposition is performed by a STFT with frames of 256 samples and 50% overlap, using a hamming window. The sampling rate is 16 kHz. All signals have been normalized with equal power, and the threshold value to remove low-energy T-F points is set to 0 dB. The source delays introduced in the mixtures are included in table 1.

Figure 4 represents the enumeration accuracy rate averaged over 50 mixtures with a varying number of sources. The enumeration in the case of 2 and 3 sources is almost perfectly performed, but when the number of sources increases, the error in the estimation also increases, as it was expected. Nevertheless, the accuracy rate in the 5 sources case is still 80%, which is a noticeable good value for speech enumeration.

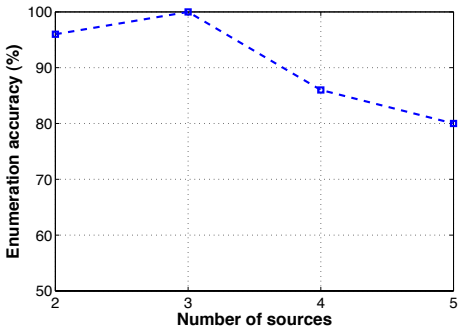

Fig. 4. Averaged accuracy rate (%) obtained in the estimation of the number of sources in anechoic speech mixtures of 2, 3, 4 and 5 sources.

4. DISCUSSION

In this paper, we propose a novel method to solve the problem of speech source enumeration, based on an information theoretic coding of the relative delays spectrum series. The number of sources is equivalent to the order of the optimal linear model whose selection is achieved by the MDL criterion. The performance of the proposed algorithm sets a standard in the enumeration of sources in anechoic speech mixtures, with a moderate number of sources not to violate the T-F disjointness of the delays. While not included in this paper for space reasons, the proposed linear model was successfully applied to echoic mixtures of sources. In such a case, the width of the peaks that appear in the PDF is larger, due to reverberation. Further investigations should be carried out in this direction. This technique is specially useful in source separation applications, when the number of sources is an input parameter for the separation algorithms. The presented results are promising, but further research is necessary in order to investigate the robustness in noisy environments, the dependence on the relative positions of speech sources as well as their energy.

5. REFERENCES

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach", IEEE Signal Processing Mag., Vol. 13, pp. 67-94, July 1996.

[2] H. Akaike,"A new look at the statistical model identification," IEEE Trans. Autom. Control, vol.19, pp. 716-723, Dec. 1974.

[3] J. Rissanen, "Modeling by shortest data description," Automatica, vol. 14, pp. 465-471, Sep. 1979.

[4] Z. Lu and A.M. Zoubir, 'Source enumeration using the pdf of sample eigenvalues via information theoretic criteria," Proc. 2012 IEEE Int. Conf. Acoust., Speech and Signal Process., pp. 3361-3364, March 2012.

[5] W. Cheng, S. Lee, Z. Zhang and Z. He, "Independent component analysis based source number estimation and its comparison for mechanical systems", Journal of Sound and Vibration, vol. 331-23, pp. 5153-5167, Nov. 2012.

[6] P. J. Green and D. P. Taylor, "Dynamic signal enumeration algorithm for smart antennas," IEEE Trans. Signal Process., vol. 50, pp. 1307-1314, Jan. 2002.

[7] S. Arberet, R. Gribonval, F. Bimbot, "A Robust Method to Count and Locate Audio Sources in a Multichannel Underdetermined Mixture," IEEE Trans. on Signal Process., vol. 58-1, pp.121-133, Jan. 2010.

[8] K.D. Gilbert and K.L. Payton, “Source Enumeration of speech mixtures using pitch harmonics,” IEEE Workshop Applicat. Audio and Acoust., pp. 89-92, Oct. 2009.

[9] A. de Cheveigné, A. Baskind, “F0 estimation of one or several voices,” Proc. Eurospeech 2003, pp. 833-836, 2003.

[10] H. Katmeoka, T. Nishimoto and S. Sagayama, “Separation of harmonic structures based on tied Gaussian mixture model and information criterion for concurrent sounds,” Proc. 2004 IEEE Int. Conf. Acoust., Speech and Signal Process., pp. 297-300, May 2004.

[11] A. Jourjine, S.Rickard, and O. Yilmaz, “Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures,” Proc. 2000 IEEE Int. Conf. Acoust., Speech and Signal Process., pp. 2985-2988, 2000.

[12] H. Krim, J.H. Cozzens, “A Data-Based Enumeration Technique for Fully Correlated Signals,” IEEE Trans. Signal Process., vol. 42-7, pp. 1662-1668, July 1994.

[13] O.Yilmaz,S.Rickard,“Blind separation of speech mixtures via time-frequency masking,” IEEE Trans. Signal Process., vol. 52 pp. 1830-1847, July 2004.

[14] S. Kay, “Model-Based probability Density Function Estimation,” IEEE Signal Process. Lett., vol. 5-12, pp. 318-320, Dec. 1998.

[15] W. Fisher, G. Doddington, K. Marshall, “The DARPA speech recognition research database: Specification and status,” Proc. DARPA Speech Recognition Workshop, pp. 93-100, 1986.