ELSEVIER

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Signal Processing 86 (2006) 353–359

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

An analysis of the limitations of blind signal separation application with speech

Daniel Smith*, Jason Lukasiak, Ian S. Burnett

Whisper Laboratories, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, NSW, Australia

Received 15 August 2003; received in revised form 15 September 2004; accepted 25 May 2005
Available online 20 July 2005

Abstract

Blind Signal Separation (BSS) techniques are commonly employed in the separation of speech signals, using Independent Component Analysis (ICA) as the criterion for separation. This paper investigates the viability of employing ICA for real-time speech separation (where short frame sizes are the norm). The relationship between the statistics of speech and the assumption of statistical independence (at the core of ICA) is examined over a range of frame sizes. The investigation confirms that statistical independence is not a valid assumption for speech when divided into the short frames appropriate to real-time separation. This is primarily due to the quasi-stationary nature of speech over the temporal short term. We conclude that employing ICA for real-time speech separation will always result in limited performance due to a fundamental failure to meet the strict assumptions of ICA.

Keywords: Mutual information; Blind signal separation; Independent component analysis; Statistical independence

1. Introduction

Speech has proven to be a major area of interest within Blind Signal Separation (BSS) research. This is largely due to the potential of applying BSS for speech enhancement within an audio scene; examples of such applications include front-end enhancement for noisy speech recognition, hearing aids and mobile telephony [1].

BSS primarily employs Independent Component Analysis (ICA) as a criterion for separation [2], including applications in a speech/audio environment. Typically, in such applications, batch BSS techniques are employed and previous reported work such as JADE [3] and FastICA [4] (which assume instantaneous mixing environments) have indicated that these algorithms achieve reasonable performance operating on speech [5–7]. These BSS techniques presume that in batch mode, the large data lengths will meet a

*Corresponding author. Tel.: +61 2 42716957; fax: +61 2 42213236.
E-mail addresses: dsmith@titr.uow.edu.au (D. Smith), jasonl@elec.uow.edu.au (J. Lukasiak), i.burnett@elec.uow.edu.au (I.S. Burnett).

0165-1684/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2005.05.020

constraint of statistical independence. However, batch based algorithms fail to represent the dynamic nature of a realistic audio environment. For instance, it is stated in [8] that audio applications (involving live speakers) that "apply some means of inverse filtering would have to be adaptable on almost a frame-by-frame basis to be effective". The necessity of estimating the inverse mixing matrix on a frame-by-frame basis, combined with the 200 ms bound on the delay of interactive voice communication [9], indicates that audio based BSS must be applied in real time with a very limited dataset.

In addition to an algorithm's data efficiency, it is necessary to consider the more fundamental issue of whether signals comply with the ICA criteria (statistical independence between signals) for frame sizes suited to real-time application. Although this issue has not been addressed in the context of ICA in real time, other BSS work [10,11] has considered the statistical dependencies between speech signals. In [11], it was reported that strong cross-correlations may exist between speech signals for small (but non-negligible) time periods which can weaken the separation performance of Adaptive Decorrelation Filtering (ADF). In addition, [10] revealed that the performance of ICA in the frequency domain degrades as the number of samples in each frequency bin decreases. This is due to an increase in the statistical dependency between speech signals in each bin [10].

While [10,11] suggest that the statistical dependencies between speech signals increase across shorter periods, this paper considers the issue in more detail. In particular, it addresses the validity of the underlying ICA assumptions for a speech signal framed for real-time processing. We investigate ICA/real-time speech processing compatibility through a detailed analysis of the statistical dependence of speech signals with respect to frame size. This provides insights into the validity of ICA as a criterion for frame-based speech separation. The analysis conducted in this paper employs instantaneous mixtures of speech, and not the convolutively mixed speech of a more realistic acoustic environment. Instantaneous mixtures, however, present the best case scenario for signal separation via BSS. Thus, the conclusions of this analysis should be directly applicable to convolutively mixed speech observations.

2. Mutual Information

Mutual information (MI) is an information theoretic measure of the dependence of random variables. MI can be defined in the discrete two-dimensional domain for variables x and y as [12]

$$\mathrm{MI}(x,y)=\sum_{x\in N}\sum_{y\in N1}P(x,y)\,\log\frac{P(x,y)}{P(x)P(y)}, \tag{1}$$

where $P(x,y)$ is the joint pdf of x and y, and $P(x)$ and $P(y)$ are the marginal pdfs of x and y with observation spaces N and $N1$, respectively. MI is regarded as an ideal measure of statistical independence as it considers the whole dependency structure of the variables, unlike correlation measures that only consider linear dependencies [12]. In this analysis, the estimator detailed in [13] was used to compute (1).

3. Analysis of the relationship between statistical independence and speech

3.1. MI analysis data set

This section presents a detailed analysis of typical speech signal statistical dependence across a range of frame sizes that cover the entire spectrum of BSS applications; from batch (4–5 s) to real-time application (20–30 ms). To investigate the relationship between speech and statistical independence, a data set consisting of four classes of signals including speech, natural vowels, artificial vowels and Gaussian noise was applied to the Mutual Information estimator in [13]. The MI was estimated, across the entire range of frame sizes, between all possible combinations of frames taken from a pair of signals of the same class.

The corpus used in the MI analysis consisted of 30 speech sentences from the ANDOSL database [14] and 22 natural vowels from [15]. The speech sentences were all 5 s in length, sampled at 20 kHz and consisted of speakers of various age and both genders. The natural vowels were 0.5 s long, sampled at 16 kHz and consisted of 11 vowels spoken by a male and female. We employed the database from [15] as it allowed analysis to be conducted over a set of vowels that were sustained for a relatively long duration (0.5 s). This allowed MI analysis to be conducted across a broad range of frame sizes to generate statistically meaningful results. Artificial vowels were generated from the natural vowel corpus via replication of a single pitch cycle extracted from each natural vowel. The Gaussian noise test set was artificially generated to be uncorrelated and hence statistically independent [2]. Despite the fact that pairs of Gaussian signals violate the ICA framework [2], they were included in the MI test set to provide a benchmark comparison.

3.2. MI-frame size relationship for signal classes

A summary of the results of the MI versus frame size analysis for the speech and Gaussian signals defined in Section 3.1 is shown in Fig. 1. The MI values for all signals in Fig. 1 continue to asymptotically approach zero as the frame size increases from 0.5 to 5 s. However, for brevity, only the results for frame sizes spanning 20 ms–0.5 s are shown in Fig. 1.

The results shown in Fig. 1 indicate increasing dependency for the speech and Gaussian classes as frame size is reduced. However, the increase in dependency is more than an order of magnitude greater for the speech signals than for the baseline Gaussian signals. This characteristic is particularly evident for frame sizes less than 100 ms, where the speech exhibits a significant (33%) increase in dependency as frame size is reduced. The marginal increase in dependency exhibited by the Gaussian signals can be attributed to poorer estimates of the underlying statistics as the sample size is reduced. This same effect could account for a small increase in the MI value for speech as the frame size decreases. However, the dramatic increase exhibited by the speech signals must be attributed to some additional factor. This additional factor involves the physical characteristics imposed upon the speech signal by the auto-regressive structure of the speech production mechanism [16]. This is further examined in Section 3.4.

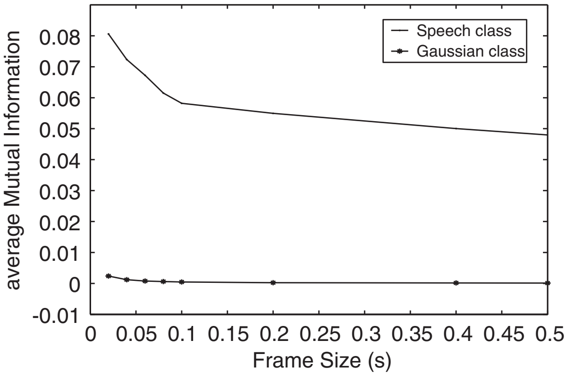

Fig. 1. Average mutual information estimated for speech and Gaussian classes for frame sizes 20 ms–0.5 s.

3.3. Deterministic and harmonic speech signal effects on MI

To examine the effects of a deterministic 'speech' signal on the relationship between MI and frame size, the set of artificial vowels (detailed in Section 3.1) were employed in a comparative analysis with the set of natural vowels. The artificial vowels were used as they represent the extreme of predictability (maximum autocorrelation) that a natural speech signal may exhibit in the temporal short term. In addition, a subset of the artificial and natural vowels employed in the analysis were chosen to possess pitch periods that were integer multiples of a common fundamental frequency. In the remainder of this paper these signals are referred to as being harmonically related. The harmonic artificial vowels differ from the harmonic natural vowels, however, as the artificial vowel's harmonic pitch relationships are constant across their duration. Whereas the natural vowel's pitch period may vary slightly due to the dynamic nature of the speech production mechanism; thus they are not consistently harmonically related across their duration. Fig. 2 compares the average MI measured between all 22 pairs of the natural vowels data set, pairs of harmonic artificial vowels and pairs of harmonic natural vowels. The MI was estimated across all

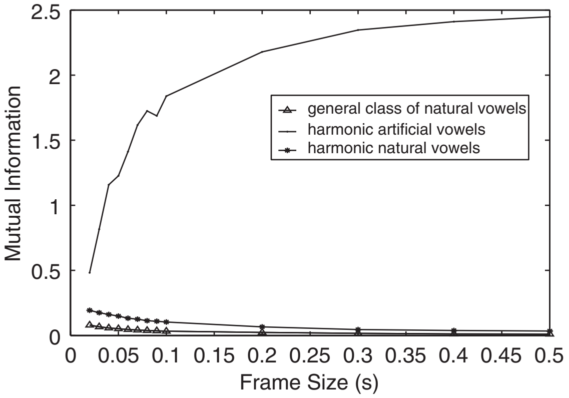

Fig. 2. Average mutual information estimated for harmonic artificial vowels, harmonic natural vowels and the entire class of natural vowels for frame sizes 20 ms–0.5 s.

frame permutations and for frame sizes 20 ms–0.5 s in all cases. The results shown in Fig. 2 indicate that the MI between pairs of harmonic artificial vowels is significantly higher than pairs of natural vowels (both harmonic and non-harmonic). In fact, the artificial vowels exhibit an increase in MI as frame size is increased. This relationship clearly violates the overall trend for all other signals examined (see Figs. 1 and 2) and can be attributed to both the harmonic relationship and deterministic nature of the artificial vowels. These characteristics result in highly predictable temporal relationships that actually increase as frame size extends to encompass multiple pitch cycles (higher autocorrelation). Fig. 2 also indicates higher MI results for harmonically related natural vowels when compared to the entire class of natural vowels. This result clearly indicates that harmonic relationships result in increased dependence. However, while harmonicity results in an approximate doubling of the MI for frame sizes below 200 ms, this is significantly smaller than the order of magnitude increase evident for artificial vowels. This distinction clearly indicates that the majority of the MI increase for the artificial vowels is due to their deterministic nature. The results presented in this section indicate that harmonically related vowels clearly defy the statistical independence criteria imposed by BSS algorithms and this characteristic is further accentuated by the predictability of the speech signals.

3.4. Influence of the speech production model on MI

The results shown in Sections 3.2 and 3.3 indicate that the mutual information between speech signals increase dramatically as the frame size of the speech signals decrease. It is proposed that the characteristics of the speech production model, in particular the quasi-stationary nature and inherent correlation of speech over the temporal short term (for up to 30 ms) [16], are responsible for this relationship. To objectively analyse the influence of the speech production model on MI, the MI between all possible frame combinations of two speech signals Speaker 1 and Speaker 2 (obtained from 1 s segments of the speech set described in Section 3.1) was calculated for frame sizes of 20 ms and 80 ms. These frame sizes were a reasonable choice to allow comparison of the statistical dependencies between speech signals considered quasi-stationary (20 ms) and speech signals less stationary in nature (80 ms).

It is evident comparing Fig. 3(a) and Fig. 3(b) that the MI for 80 ms frames is consistently low (<0.1), while the MI for the 20 ms frames vary dramatically (from 0 to 0.47). The large variation in MI for 20 ms frames is due to these smaller frames having sufficient temporal resolution to represent a single phoneme or at least a single phonetic classification. The mutual information troughs of Fig. 3(a) (labelled i) refer to the frames of a speaker that present minimum MI across all frames of the other speaker. These troughs correspond to unvoiced frames of speech. Unvoiced frames have previously been demonstrated to be substantially noise-like, and thus these frames approach statistical independence from all other speech frames. Portions of Fig. 3(a) that display more significant MI (labelled ii) correspond to frames of voiced speech for both Speakers 1 and 2. Voiced speech, examined in the form of natural vowels in Section 3.3, is quasi-periodic and possesses temporal correlation, resulting in some predictability between voiced frames. The maximum MI peaks of Fig. 3(a) (labelled iii) correspond to voiced sections of Speakers 1 and 2 that have the greatest temporal predictability, due to the formation of harmonic pitch relationships. The underlying reason that Fig. 3(b) fails to

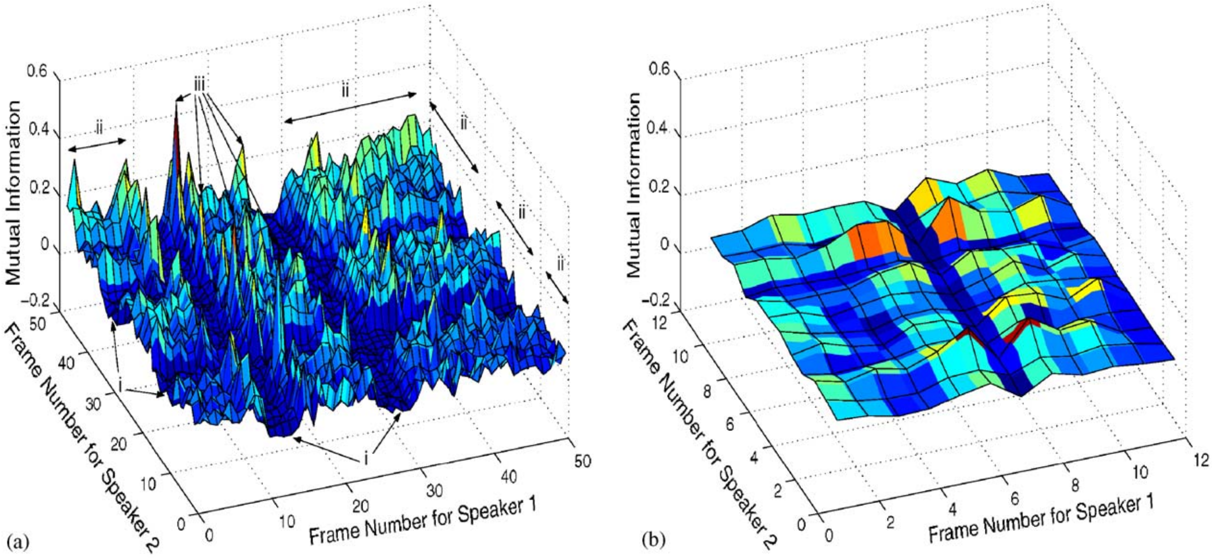

Fig. 3. Mutual Information estimated between all combinations of frames belonging to two 1 s sections of speech signals, Speakers 1 and 2, for frame sizes of 20 ms (Fig. 3(a)) and 80 ms (Fig. 3(b)). Label i corresponds to the unvoiced frames of Speakers 1 and 2. Label ii refers to frames of voiced speech between Speaker 1 and Speaker 2, while label iii corresponds to voiced frames that have formed harmonic pitch relationships.

exhibit the same peaks and troughs as those evident in Fig. 3(a) is due to longer frames capturing the time-varying nature of the vocal tract (the evolution of speech) and variation between voiced and unvoiced speech.

4. ICA application with speech in relation to frame size

In this section, the effect of increasing the statistical dependencies between pairs of speech is analysed for common ICA algorithms. The Joint Approximate Diagonalization of Eigenmatrices (JADE)[1][3] and FastICA[2][4] algorithms were applied to the 30 speech signals from the data set of Section 3.1 and a baseline class of iid Laplacian distributed data. The algorithms were applied to all possible combinations of frames for each signal pair and over a range of frame sizes from 20 ms to 5 s. The performance criterion employed for JADE and FastICA was an interference measure (IM) defined as

$$\mathrm{IM} = \frac{1}{2}\sum_{j=1}^{2} \frac{\left(p_j p_j^T - max(p_j)^2\right)^{1/2}}{max(p_j)}, \tag{2}$$

where p is the product of the mixing and separation matrix, and p_j is a row of p. Eq. (2) essentially measures p's distance from a matrix corresponding to the product of a permutation matrix and diagonal matrix. It is the inverse of the measure used in [6]. Informal listening tests conducted upon the speech mixtures of this analysis, indicate that an IM of 0.03 or less corresponds to a level of separation where interference is inaudible. The results of this analysis, the average IM versus frame size for both ICA algorithms (as shown in Fig. 4), indicate that both the speech and the Laplacian classes record higher levels of interference when frame size is reduced. The IM increase in the Laplacian data can be attributed to the sub-optimal estimation performance of JADE and FastICA for smaller frame

[1] Download real version 1.5, http://sig.enst.fr/cardoso/stuff.html, July 2003.

[2] Download version 2.1, http://www.cis.hut.fi/projects/ica/fastica, July 2003.

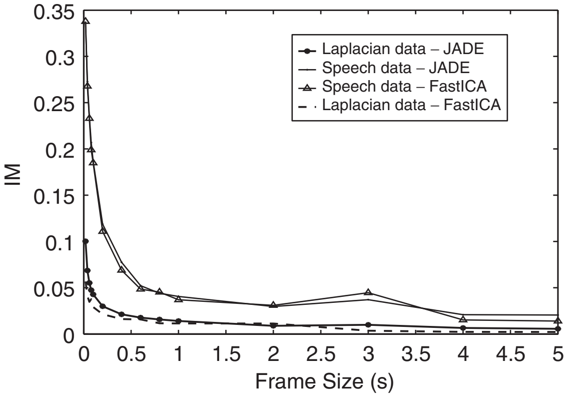

Fig. 4. The average IM obtained by applying JADE and FastICA to the set of speech signals and Laplacian data for frame sizes 20 ms–5 s.

sizes. However, it fails to account for the additional interference the speech class possesses when compared with the Laplacian class across all frame sizes. In particular, the speech signal's IM was up to five times greater than the Laplacian signals across shorter frame lengths (<100 ms). The speech signal's unacceptably high IM levels (0.18–0.3) correlate with the increased statistical dependencies for smaller frame sizes; reported in Fig. 1. This result clearly indicates that the increasing dependency exhibited by speech signals, as the frame size is reduced, is a significant factor contributing to the poor performance of the ICA algorithms for real-time framed speech.

The feasibility of applying ICA to speech in real time is also effected by another issue. It was detailed in the MI analysis of Section 3.4 that as frame size is reduced, a frame is more likely to consist entirely of unvoiced speech. Unvoiced speech has been reported to be Gaussian distributed in [17]. In order to retain signals, the ICA model requires that all independent components but one are non-Gaussian [2]. Thus, when more than one speech signal in the mixture frame is composed of unvoiced speech, the non-Gaussian assumption of ICA is violated. Given an approximation that unvoiced speech comprises 20% of a speech signal [18], it can be concluded that two real-time framed speech signals will violate the non-Gaussian assumption of ICA, on average, for 4% of the frames.

5. Conclusion

The mutual information analysis of this paper reveals a general trend in the MI—frame size relationship of speech. As frames of speech are decreased in size, the statistical dependencies between them increase. This relationship has particular relevance to BSS with speech in a time-varying mixing environment, which requires a real-time approach to separation. Significant statistical dependencies exist for the smaller frames of this MI analysis, due to the quasi-stationary nature and inherent correlation of speech over the temporal short term. As a consequence, the underlying ICA model is incompatible with frames of speech that are considered small enough for real-time application. However, as the size of the analysis speech frames increases, frames capture the long-term behavior of speech, exhibit less correlation and approach statistical independence.

Thus, it is concluded that although ICA is suitable for application with speech in batch techniques possessing substantial data, it is inevitably less reliable for realistic audio environments that require a real-time approach to separation.

References

[1] K. Torkola, Blind separation for audio signals—are we there yet?, Proceedings of the first International Workshop on ICA and BSS, 1999, Aussois, France, pp. 239–244.

[2] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[3] J.F. Cardoso, A. Souloumiac, Blind beamforming for nongaussian signals, IEE Proceedings-F 140 (6) (1993) 362–370.

[4] A. Hyvarinen, Fast, Fixed-Point Algorithms for Independent Component Analysis, Trans. Neural Network 10 (3) (1999) 626–634.

[5] B. Millar, J. Vonwiller, J. Harrington, P. Dermody, Australian National Database of Spoken Language (ANDOSL), CD ROM.

[6] K. Hild, D. Erdogmus, J. Principe, Blind source separation using Renyi's mutual information, IEEE Signal Process. Letters 8 (6) (2001) 174–176.

[7] A.J.W. van der Kouwe, D. Wang, G. Brown, A comparison of auditory and blind separation techniques

for speech segregation, IEEE Trans. Speech Audio Processing 9 (3) (2001) 189–195.

[8] M. Brandstein, On the use of explicit speech modeling in microphone array applications, in: Proc. of ICASSP98, Seattle, May 1998, pp. 3613–3616.

[9] D. Ferrari, Client requirements for real time communication systems, rfc 1193, 1990.

[10] S. Araki, R. Mukai, S. Makino, T. Nishikawa, H. Saruwatari, The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech, IEEE Trans. Speech Audio Processing 11 (2) (2003) 109–116.

[11] Y. Zhao, R. Hu, S. Nakamura, Whitening processing for blind signal separation of speech signals, in: Proceedings of ICA2003, 2003, Nara, Japan, pp. 331–336.

[12] A. Fraser, H. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (2) (1986) 1134–1140.

[13] G. Darbelly, I. Vadja, Estimation of the information by an adaptive partitioning of the observation space, IEEE Trans. Inform. Theory 45 (4) (1999) 1315–1321.

[14] T. Blaschke, L. Wiskott, CuBICA: independent component analysis by simultaneous third- and fourth-order, IEEE Trans. Speech Audio Processing 52 (5) (2004) 1250–1256.

[15] M. Hasegawa-Johnson, A. Alwan, J. Cha et al., Vowels MRI databse, http://www.ifp.uiuc.edu/speech/mri/vowels.html, last accessed 21/5/03.

[16] L. Rabiner, R. Schafer, Digital processing of speech signals, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[17] G. Kubin, Nonlinear Processing of Speech, Speech Coding and Synthesis, Elsevier Science, Elsevier, Amsterdam, 1995, pp. 557–611.

[18] R. Hagen, E. Paksoy, A. Gersho, Voicing-specific LPC quantization for variable-rate speech coding, IEEE Trans. Speech Audio Processing 7 (5) (1999) 485–494.