INTERSPEECH 2009 BRIGHTON

# Model-based Speech Separation: Identifying Transcription using Orthogonality

*S. W. Lee* [1], *Frank K. Soong* [1,2], *and Tan Lee* [1]

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

10.21437/Interspeech.2009-417

## Abstract

Spectral envelopes and harmonics are the building elements of a speech signal. By estimating these elements, individual speech sources in a mixture observation can be reconstructed and hence separated. Transcription gives the spoken content. More important, it describes the expected sequence of spectral envelopes, if modeling of different speech sounds is acquired. Our recently proposed single-microphone speech separation algorithm exploits this to derive the spectral envelope trajectories of individual sources and remove interference accordingly. The correctness of such transcription becomes critical to the separation performance. This paper investigates the relationship between the correctness of transcription hypotheses and the orthogonality of associated source estimates. An orthogonality measure is introduced to quantify the correlation between spectrograms. Experiments verify that underlying true transcriptions lead to a salient orthogonality distribution, which is distinguishable from the counterfeit transcription one. Accordingly a transcription identification technique is developed, which succeeds in identifying true transcriptions in 99.74% of the experimental trials.[1]

**Index Terms**: speech separation, orthogonality, transcription, speech enhancement

## 1. Introduction

Speech separation is a fundamental problem in speech processing. In typical situations, multiple sound sources, in the form of signals from target speakers, competing speech and background noise, are present, constituting the resultant input mixtures. These sound sources overlap in both time and frequency domains, corrupting each other. Separation of individual speech sources from mixture signals becomes essential. One popular approach is independent component analysis (ICA), which relies on the statistical property between sources and the availability of multiple input mixtures [1]. This paper focuses on single-microphone speech source separation.

Human being is capable of segregating interested sound sources from interference and background noise, even with a single ear [2], [3]. Modeling of how human separates concurrent sources may be one viable way to extract target speech sources [4]-[6]. This approach is referred to as computational auditory scene analysis (CASA). Our perceptual system performs an auditory scene analysis for the input mixture, by examining primitive, acoustic regularities (for example, harmonicity and common fate etc.) and applying high-level knowledge like familiar speech patterns. Most of the CASA based separation methods utilize primitive regularities, in particular, the strong harmonicity cue. However, tracking multiple pitch frequencies is difficult per se [7] and perceptual experiments show that the knowledge of familiar speech patterns is indispensible to proper segregation [3], [8].

We recently proposed a speech separation algorithm based on speech production and modeling of familiar speech patterns [9]. Individual speech sources are estimated in terms of their spectral envelope trajectories and harmonic structures. Each spectral envelope trajectory is found by matching the input mixture with a transcription and speech models. The interference source is then removed accordingly. This transcription information provides linguistic knowledge of the source, but not the actual acoustic signal.

True transcriptions lead to estimates close to ideal sources, whereas counterfeit transcriptions do not. In the following, a transcription identification technique is proposed. The underlying transcriptions of sources are identified from a set of hypothesized word-level transcription candidates. We introduce an orthogonality measure and analyze the orthogonality generated from true and counterfeit transcriptions. Experimental results show that the orthogonality measure between source estimates and the input mixture provides reliable transcription identification.

## 2. Transcription-driven speech separation and orthogonality

The separation algorithm [9] is briefly reviewed to illustrate the role of a transcription in source estimation and the consequence of true and counterfeit transcriptions.

### 2.1. Model-based speech separation algorithm

An input mixture signal $x(n)$ is related to its two constituent speech sources $x_1(n)$ and $x_2(n)$ as

$$x(n) = x_1(n) + x_2(n) \quad (1)$$

Figure 1 gives the block diagram of the model-based speech separation algorithm. Let $x_i(n)$ be the current target source ($i \in \{1, 2\}$).

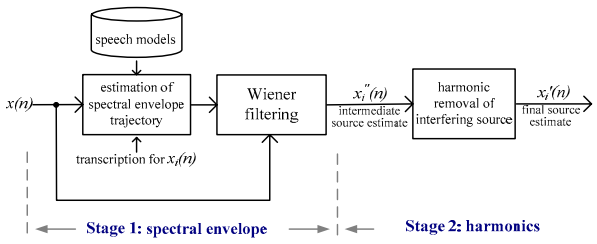


Figure 1: *Block diagram of the model-based speech separation algorithm.*

[1] This research is partially supported by an Earmarked Research Grant (Ref: CUHK 414108) from the Hong Kong Research Grants Council. The work is also affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-Centric Computing and Interface Technologies.

$x_i(n)$ is estimated in a 'synthetic' manner, by working out the associated spectral envelope trajectory and harmonics. In Stage 1, speech models are used to represent the phonetic-acoustic mapping (i.e. the normalized spectral envelopes for individual speech sounds). Consequently, the transcription dictates the expected sequence of spectral envelopes. Let $P_{x_i}(\omega)$ be the power spectral density of $x_i(n)$. We estimate $P_{x_i}(\omega)$ by forced alignment of $x(n)$ with the transcription. By concatenating and replicating the model parameters according to the resultant state-level time-alignment, $P_{x_i}(\omega)$ is revealed. $P_{x_i}(\omega)$ is further revised by adjusting the gain at different time instants. Accordingly, a Wiener filter [10] is derived and applied to $x(n)$. A source estimate $x_i''(n)$ is output. Wiener filtering is used here to remove the interference source. Recall that the input transcription explicitly determines the models and their order for the generation of spectral envelope trajectory, hence, correct transcription is necessary for proper separation. Stage 2 is aimed to retain the pitch harmonics of $x_i(n)$ and remove any harmonics that belong to the interference source, as the harmonic structures of both $x_1(n)$ and $x_2(n)$ remain after Stage 1.

The filter output $x_i''(n)$ varies with the input transcription, depending on $P_{x_i}(\omega)$ estimated. In Figure 2, two distinct Wiener filters are shown. They are derived for different sources. The input mixtures are identical. Comparing the power spectra between the output and the associated Wiener filter, the output estimate closely follows the filter response and reflects the spectral characteristics of the filter. Hence, substantial difference between the two output estimates is observed.

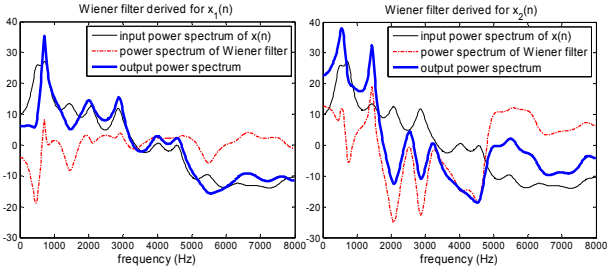


Figure 2: *Two output power spectra generated by distinct Wiener filters. (left) Filter derived for $x_1(n)$; (right) Filter derived for $x_2(n)$. Identical input mixture signal is used.*

## 2.2. True and counterfeit transcriptions

Suppose two Wiener filters are derived respectively by an underlying true transcription (that corresponds to the target source $x_i(n)$) and a counterfeit transcription (belongs to neither of the sources). Considering the correlation between each filter output with the input mixture $x(n) = x_1(n) + x_2(n)$, it is expected that the one from the true transcription has a higher degree of correlation with $x(n)$ than the one from counterfeit transcription. Note that the estimation error of a Wiener filter is orthogonal to the input mixture $x(n)$. As the output generated from a counterfeit transcription bears a component that represents the counterfeit transcription and an error component which is orthogonal to $x(n)$, consequently, this output will have negligible correlation to $x(n)$.

Moreover, for independent speech sources $x_1(n)$ and $x_2(n)$, if both true transcriptions are used to construct the Wiener filters, one of the filter outputs will be close to $P_{x_1}(\omega)$ with residue coming from $x_2(n)$. The other filter output will be close to $P_{x_2}(\omega)$ with residue coming from $x_1(n)$. Thus these two filter outputs will be correlated. If one or more counterfeit transcriptions are used instead, the two filter outputs will be orthogonal.

## 2.3. Orthogonality measure

The degree of correlation between random variables is often measured by the correlation coefficient [11]. For two waveforms or spectrograms, we use the angle of the inner product as a measure of correlation. The inner product of two real $n$-dimensional vectors ($\Re^n$), $y_1$ and $y_2$, is defined as

$$\langle y_1, y_2 \rangle = y_1^T y_2 = \sum_{i=1}^{n} y_{1i} y_{2i} \tag{2}$$

We calculate the orthogonality $\theta$ between the two non-zero vectors $y_1$, $y_2$ (as a measure of correlation in an angle sense) by

$$\theta = \angle(y_1, y_2) = \cos^{-1}\left( \frac{y_1^T y_2}{\|y_1\|_2 \|y_2\|_2} \right) \tag{3}$$

where $\|\cdot\|_2$ is the Euclidean norm. The two norms in the denominator normalize the correlation and make the measure of orthogonality bounded and independent of the lengths of $y_1$ and $y_2$. If $\langle y_1, y_2 \rangle = 0$, the $y_1$ and $y_2$ are orthogonal ($\theta = \pi/2$ rad). The closer $\theta$ to $\pi/2$ rad (90°), the more orthogonal are the two vectors. This cosine angle can be related to correlation coefficient that the correlation coefficient of two zero-mean random variables represents the expected cosine angle between the two sampled vectors.

To measure orthogonality for our transcription identification task, the above inner product is operated on $\Re^{m \times n}$ (the set of $m \times n$ real matrices). It is given by

$$\langle Y_1, Y_2 \rangle = tr\left(Y_1^T Y_2\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} Y_{1ij} Y_{2ij} \tag{4}$$

for $Y_1, Y_2 \in \Re^{m \times n}$, where $tr$ denotes the trace of a matrix. $\Re^{m \times n}$ here represents the magnitude spectrogram space. Hence, $m$ and $n$ are the numbers of frequency bins and frames respectively. This inner product $\langle Y_1, Y_2 \rangle$ is the same as the inner product of the corresponding vectors in $\Re^{mn}$ by taking the matrix elements column-wise or row-wise. As a result, the angle $\theta$, as our measure of the orthogonality between two magnitude spectrograms is given by

$$\theta = \angle(Y_1, Y_2) = \cos^{-1}\left( \frac{tr\left(Y_1^T Y_2\right)}{\|Y_1\|_F \|Y_2\|_F} \right) \tag{5}$$

where $\|\cdot\|_F$ represents the Frobenius norm. Given that the inner product is operated on the magnitude spectrograms, the elements of $Y_i$ are non-negative. This implies $0 \le \theta \le \pi/2$.

The experiments in the next section illustrate that the above angle-based orthogonality measure provides discrimination between true transcriptions and counterfeit transcriptions. Assume that $N$ possible transcription hypotheses are available and the underlying true transcriptions are included. By examining the distribution of $\theta$ generated from these hypotheses, a transcription identification technique is proposed to distinguish the true transcriptions of $x_1(n)$ and $x_2(n)$ from the others.

## 3. Experiments and discussions

The above inner-product based orthogonality measure is first applied to a set of speech sources. Orthogonality between ideal speech sources is examined. Then, statistics of the angle $\theta$ collected from Wiener-filtered source estimates with true transcriptions is compared with those obtained from counterfeit transcriptions. 100 continuous speech utterances of American English from TIMIT corpus [12] are used. They are in distinct spoken contents. The average duration is about three seconds. There are 62 male and 38 female speakers. 4950 mixture signals are generated by mixing all possible combinations of two utterances at equal power.

Figure 3 depicts the experimental setup, specifically, the two spectrograms which orthogonality is measured from. The measurement is taken either: between both ideal speech sources $x_1(n)$ and $x_2(n)$ ($\theta_A$); between one of the source estimates and the input mixture ($\theta_B$); or between both source estimates $x_1''(n)$ and $x_2''(n)$ ($\theta_C$). To focus on the orthogonality of Wiener-filtered source estimates, the magnitude spectrogram of the source bearing the considered transcription is directly adopted to compute $P_{x_i}(\omega)$, acting as the aligned model sequence.

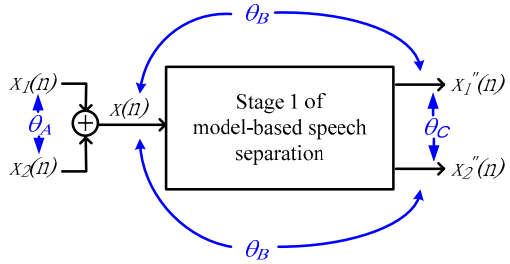


Figure 3: *Orthogonality measure is taken either between: both ideal speech sources ($\theta_A$); one of the source estimates and the input mixture ($\theta_B$); or both source estimates ($\theta_C$).*

### 3.1. Orthogonality between ideal speech sources

The orthogonality between ideal speech sources $x_1(n)$ and $x_2(n)$ is studied. Figure 4 shows the distribution of $\theta_A$ measured from different source combinations. Most speech sources are found to be approximately orthogonal. The maximum and minimum $\theta_A$ measured are 85.39° and 55.81° respectively. The mean and standard deviation are 74.03° and 4.21° respectively.

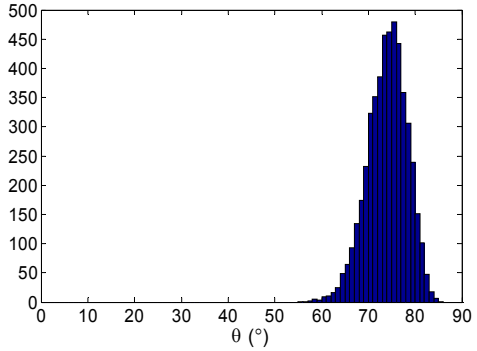


Figure 4: *Histogram of $\theta_A$ measured between ideal speech sources.*

### 3.2. Orthogonality between one source estimate and input mixture

The orthogonality between a source estimate and the input mixture is examined here. The transcription hypothesis set consists of the 100 transcriptions from all the source utterances. For an input mixture, the constituent speech sources give the two true transcriptions and 98 remaining transcriptions are counterfeit. $\theta_B$ from true transcriptions and counterfeit transcriptions are measured respectively.

Figure 5 depicts the histograms of $\theta_B$. Observing the distributions of $\theta_B$ under cases of true transcriptions and counterfeit transcriptions, they are highly different. With counterfeit transcriptions, the source estimate and the input mixture are orthogonal, with $\theta_B$ close to 90°; whereas with true transcriptions, the source estimate and the input mixture are correlated. Note that the distribution for true transcriptions is approximately symmetric with a mean value of 52.5°. This confirms that true transcriptions lead to source estimates correlated to input mixtures, but not for counterfeit transcriptions.

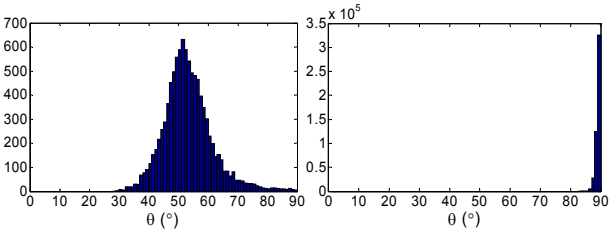


Figure 5: *Histograms of $\theta_B$ measured between input mixture and source estimate. (left) True transcription; (right) Counterfeit transcription.*

### 3.3. Orthogonality between source estimates

For an input mixture, there is a corresponding set of two underlying true transcriptions and $(C_2^{100} - 1) = 4949$ counterfeit transcription sets, where $C_2^N$ is the choose function. Over 4950 mixture signals, $\theta_C$ measured from true transcription sets and from counterfeit transcription sets are analyzed. The results are given in Figure 6.

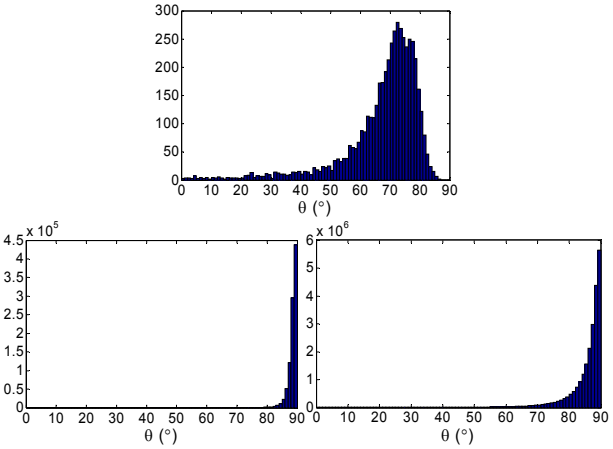


Figure 6: *Histograms of $\theta_C$ measured between source estimates. (top) Both transcriptions are true; (bottom left) One true and one counterfeit transcription; and (bottom right) Both transcriptions are counterfeit.*

Comparing these histograms, if there is at least one transcription is counterfeit, the filtered source estimates are highly orthogonal (as shown in bottom left and right histograms). On the other hand, if the true transcriptions for both sources are used (top histogram), the $\theta_C$ measured are smaller, showing that the estimates are relatively correlated. This confirms our conjecture discussed in Section 2 and the measure of $\theta$ between source estimates discriminates true transcriptions from counterfeit transcriptions.

### 3.4. Transcription identification

With the above statistical study of $\theta$, here we propose a transcription identification technique to choose the true underlying transcriptions from a set of hypotheses. Comparing the distributions in Figure 5 and 6 ($\theta_B$ and $\theta_C$ respectively), $\theta_B$ obtained from true transcriptions are much far away from those obtained from counterfeit transcriptions in Figure 5, located at smaller values. This implies that the orthogonality measure $\theta_B$ (between the mixture and a source estimate) better differentiates between true and counterfeit transcriptions. Moreover, considered the computation involved in measuring $\theta_B$ from all possible $N$ transcriptions and measuring $\theta_C$ from all $C_2^N$ transcription hypothesis pairs, the identification decision is made to: *choose the two transcriptions having minimum θs measured between the input mixture and the corresponding source estimate (i.e. $\theta_B$)*.

This transcription identification technique is applied to the 4950 mixture signals above with the 100 transcription hypotheses. Table 1 shows the identification results. 99.74% of the trials have correctly identified both underlying true transcriptions.

Table 1. *Identification rate of underlying true transcriptions.*

| | No. of successful identification trials (identification rate) |
|---|---|
| Both true transcriptions are identified | 4937 (99.74 %) |
| At least one true transcription is identified | 4945 (99.90 %) |

Traditional ICA or blind source separation (BSS) approaches tackle the separation problem by manipulating multiple microphone inputs and a demixing matrix [1]. Certain statistical properties between source estimates are achieved during iterations. Throughout the proposed transcription identification technique and the model-based separation algorithm, the underdetermined problem — separation of speech sources recorded from a single microphone input, becomes much feasible. This improvement is due to the use of prior knowledge about familiar speech sounds in the form of speech models in deriving the hidden transcriptions and source estimates in our formulation. The set of allowed transcriptions from the identification technique and source estimates from the separation algorithm are therefore constrained. Take an example: One trivial solution to achieve minimum $\theta_B$ is to use an all-pass filter as the Wiener filter, however, this is ruled out by generally low-pass speech models.

The set of transcription hypotheses can be constructed by enumerating all possible transcriptions with grammar and language model, which are typically available in speech recognition [13].

Compared with speech recognition, this transcription identification serves the similar purpose; however, the orthogonality measure together with the model-based separation algorithm provide another cue, besides the acoustic likelihood.

## 4. Conclusions

Estimation of spectral envelopes and harmonics of individual sources enables separation of concurrent speech signals from a mixture observation. One possible way to estimate the spectral envelope trajectory is to look up the sequence of associated speech models according to a presumed transcription. A transcription identification technique has been proposed in this paper to select the underlying transcriptions from a set of hypotheses. A concise orthogonality measure is introduced to compute the correlation between two spectrograms. Our study discovers that source estimate from counterfeit transcriptions are orthogonal to mixture observation, whereas this is not the case for estimates from true transcriptions. Furthermore, the proposed transcription identification technique has successfully identified the underlying true transcriptions of both sources in 99.74% of experimental trials.

## 5. References


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

[2] 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.

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

[4] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297-336, Oct. 1994.

[5] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135-1150, Sep. 2004.

[6] J. Barker, A. Coy, N. Ma and M. Cooke, "Recent advances in speech fragment decoding techniques," in *Proc. ICSLP*, Pittsburgh, Sep. 2006, pp. 85-88.

[7] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 2, pp. 255-266, Feb. 2008.

[8] R. Meddis and M. J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Amer.*, vol. 91, pp. 233-245, Jan. 1992.

[9] S. W. Lee, F. K. Soong, and P. C. Ching, "Model-based speech separation with single-microphone input," in *Proc. Interspeech*, Antwerp, Aug. 2007, pp. 850-853.

[10] S. Haykin, *Adaptive Filter Theory*, 4th ed. New Jersey: Prentice Hall, 2002.

[11] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw Hill, 1991.

[12] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status", in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93-99.

[13] J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, Cambridge University, Mar. 1995.