Sinusoidal Approach for the Single-Channel Speech Separation and Recognition Challenge

P. Mowlaee[1], R. Saeidi[2], Z. -H. Tan[3], M. G. Christensen[4], T. Kinnunen[2],
P. Fränti[2], S. H. Jensen[3]

[1]Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum (RUB), Bochum, Germany
[2]School of Computing, University of Eastern Finland, Joensuu, Finland
[3]Dept. of Electronic Systems, Aalborg University, Aalborg, Denmark
[4]Dept. of Architecture, Design & Media Technology, Aalborg University, Aalborg, Denmark
Pejman.Mowlaee@rub.de, {zt,shj}@es.aau.dk, mgc@imi.aau.dk,
{rahim.saeidi,tomi.kinnunen,pasi.franti}@uef.fi

Abstract

Most of the single-channel speech separation (SCSS) systems use the short-time Fourier transform as their parametric features. Recent studies have shown that employing sinusoidal features for the SCSS application results in a high perceived speech quality. In this paper, we make a systematic study on automatic speech recognition results for a SCSS system that uses sinusoidal features composed of amplitude and frequency. We compare the speech recognition results with those already reported by other participants in the single-channel speech separation and recognition challenge. Our results show that a newly proposed system achieves an overall recognition accuracy of 52.3%, ranges at the median over all other participants in the challenge.

Index Terms: sinusoidal modeling, single-channel speech separation and recognition challenge.

1. Introduction

Robust speech recognition for single-channel recorded mixture is known as one of the most challenging topics in speech processing. This difficulty is because of the rapid degradation in the accuracy of typical speech recognition systems as the desired speaker signal (target) gets corrupted by other interfering speaker signals (masker). To mitigate this difficulty, a large group of methods have been proposed.

For instance, in the single-channel speech separation and recognition challenge provided in [1], several participants developed separation systems in the form of the combination of a single-channel speech separation engine and automatic speech recognition back-end. The challenge aimed at comparing the speech recognition accuracy of different single-channel speech separation systems suggested by the participants. A range of different methods were proposed as the separation engine. The methods are divided into source-driven [2, 3] and model-driven [4, 5, 6, 7, 8] approaches. The first class of methods relies on the observed mixture only while the methods in the second class are entirely based on pre-trained speaker models. According to the speech recognition results reported in the aforementioned challenge, the system proposed by IBM [7] achieved the highest recognition performance, and since its performance surpassed the average performance of the human listeners, it was called the super human speech recognition system.

The methods suggested in the challenge were either based on the short-time Fourier transform (STFT), Gamma-tone filter or pitch frequency features. The objective and subjective performance measures reported in [8] show that employing sinusoidal parameters of amplitude and frequency for separation purposes offers a viable choice by achieving higher scores in terms of the perceived speech quality compared to the conventionally used STFT-based separation methods.

While the signal quality results reported for the sinusoidal single-channel speech separation methods are promising [8, 9], investigating the impact of sinusoidal model on automatic speech recognition has been less exploited in the robust speech recognition literature. In this paper, we study single channel speech separation and recognition using sinusoidal parameters. In this way, we build a full separation and recognition system using speaker identification, signal-to-signal (SSR) estimation, speech separation and speech recognition modules as shown in Figure 1. The full system performance is reported in terms of speech recognition accuracy and is compared to that of other systems in the challenge.

2. Sinusoidal Single-Channel Speech Separation and Recognition System

Figure 1 shows the block diagram of the sinusoidal single-channel speech separation and recognition system for the speech separation and recognition challenge in [1]. The system is targeted to separate speech mixtures composed of two speakers and then recognize without any *a priori* knowledge of their identities or gain value under which the speaker signals have been mixed together. The separation and recognition system is composed of these major parts: (1) speaker identification (SID) whose goal is to identify the identities of the underlying speakers in the mixture followed by a gain estimation module which estimates the mixing gain under which they are mixed together, (2) a speech separation module which aims at separating the mixed speech into two separated output signals, and (3) speech recognition which recognizes the separated signals. In the following, we briefly explain each block.

The work of Pejman Mowlaee was supported in part by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr), contract no. MEST-CT-2005-021175. The work of Rahim Saeidi was supported in part by a scholarship from NOKIA foundation.

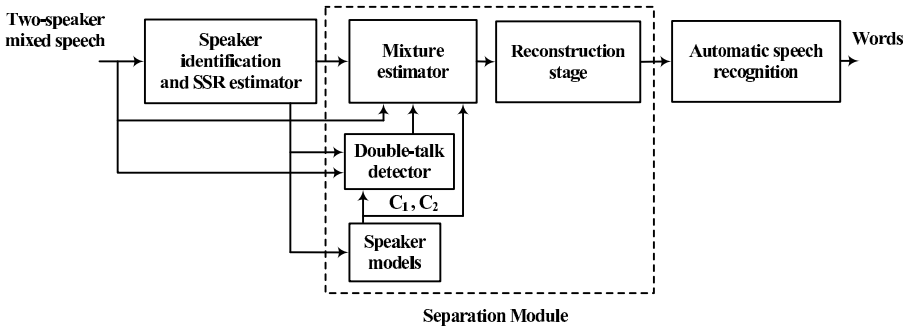

Figure 1: Block diagram of the proposed joint speaker identification, speech separation and recognition system. The codebooks for speaker one and two are indicated by C_1 and C_2, respectively.

2.1. Sinusoidal Feature Parameters

The features to be selected for a model-driven speech separation system need to meet at least two requirements: (1) low number of features for computational and statistical reasons (curse of dimensionality), and (2) high re-synthesized signal quality to ensure a good separation performance. Here we choose sinusoidal parameters which satisfy both of the aforementioned requirements [8]. Using the sinusoidal modeling as in [10], the underlying speaker signals are transformed into a parametric feature set composed of amplitude, frequency and phase vectors of sinusoidals. Furthermore, it was concluded in [8] that using the sinusoidal parameters results in improved speech separation performance compared to the STFT counterpart in terms of both objective and subjective measures.

At each frame, we represent the kth speaker signal as below

$$s_k(n) = \sum_{i=1}^{L} A_{k,i} \cos(n\omega_{k,i} + \phi_{k,i}) + e_k(n), \qquad (1)$$

where i is an index used to refer to the ith sinusoidal component characterized by the amplitude $A_{k,i}$, frequency $\omega_{k,i}$ and phase $\phi_{k,i}$, respectively, $0 \leq n \leq N-1$ with N as the window length, k denotes the kth speaker in the mixture with $k \in [1, 2]$, $e_k(n)$ is the additive noise and L is the sinusoidal model order. Among many sinusoidal modelings in the literature, here, we employ a modified version described in [10]. The spectral coefficients are translated according to the mel-frequency scale. At each frequency band, we select sinusoidal amplitude and frequency corresponding to the peak with the highest amplitude which is equivalent to choosing the maximum likelihood estimate for frequency of single sinusoid in white Gaussian noise per band.

2.2. Speaker Identification and SSR Estimator

The speaker identities and the SSR level, under which the underlying speakers are mixed together, are not known *a priori* in the observed mixture. These two parameters are required for the separation engine. The system in [7], named *Iroquois*, which in conjunction with a speech separation system, provided an average identification accuracy of 98% on the GRID corpus [1]. A modified version of the *Iroquois* system, by flooring the exponential argument in likelihood computation obtained slight improvement [4]. A text-independent stand-alone single-channel speaker identification system was proposed in [11] and is employed in current architecture, which is designed to be independent of the speech separation module. This helps in keeping the complexity of the algorithm low. It also provides the SSR estimate as a by-product.

2.3. Speaker Models

Following the selection of sinusoidal features, we need an appropriate approach to model the distribution of these features for each speaker. The separation system here works based on the pre-trained source models in the form of speaker codebooks in the sinusoidal domain [10]. Furthermore in [10], it was concluded that employing a sinusoidal coder as speaker models results in better quantization performance compared to when quantizers are trained for STFT features. A better quantization performance also means a higher upper-bound separation performance [8] because the upper-bound performance for a model-driven speech enhancement method is determined by the used quantizer [12].

2.4. Double-Talk Detector

In single-channel speech separation problem, it can be very helpful to classify the mixture speech segments into what is often referred to as single-talk, double-talk, and noise-only regions. The resulting detector is commonly referred to as a double-talk detector. Knowledge of such regions is useful since we are able to process the underlying signals differently depending on the type. In this regard, such a detector can be effectively used as a pre-processor for improving the performance [13].

The mixed signal together with the estimated identities are sent to a double-talk/single-talk detector which classifies the mixed speech signal into single-talk, double-talk, and noise-only regions. This information is used to simplify the computationally expensive separation task since one is required to process only the mixed frames with the separation system. To detect the double-talk regions with two speakers present, we employ a *maximum a posteriori* (MAP) detector proposed recently in [13]. The proposed method is based on the multiple hypothesis test and works in a speaker-dependent framework since the information for the speaker identities are already provided by the SID module.

2.5. Mixture Estimator

In a model-driven single-channel speech separation (SCSS) method, we need a mixture estimator aimed at searching the possible codevectors of the speaker models to find two optimal codevectors, one from each speaker model, such that when mixed, they satisfy a minimum estimation error criterion compared to the mixed signal. Previous separation systems use

either max-model or Wiener filter as their mixture estimator, which are the MMSE estimator for logarithm and power spectrum domain, respectively. In contrast, in the MMSE mixture estimator for the amplitude spectrum [14], the phase term is considered as a random variable which provides a more accurate mixture approximation compared to the log-max or Wiener filter estimators in terms of achieving lower mean square error for mixture estimation.

It should be noted that under specific conditions, the MMSE sinusoidal estimator for the magnitude mixture spectrum reduces to the log-max and Wiener filter estimators [14]. When one speaker dominates the other, the mixture estimate reduces to log-max mixture approximation. Another important case is when the speaker spectra are orthogonal. Then the mixture estimate reduces to the Wiener filter mixture estimate.

2.6. Reconstruction

The two codevectors provided by the mixture estimator in the previous stage are then passed to a reconstruction module, which produces the separated signals (see Figure 1). In terms of how to reconstruct the separated signals, separation methods can be divided into reconstruction-based [6, 7, 8] and mask methods [4, 5, 9]. In the former approach, the codevectors found in the mixture estimation stage are directly used for reconstructing the separated signals. The mask methods, as the name suggests, produce a mask based on the codevectors selected from the speaker models. These masks are then applied to mixture to provide separated speaker signals. Here we use the sinusoidal Wiener masks which balance the trade-off between the cross-talk suppression and minimizing the resulting speech distortion of the target signal. It also achieves a higher separation performance compared to the widely used STFT-based masks [9].

2.7. Automatic Speech Recognition

The last block as shown in Figure 1 is automatic speech recognition. The words are modeled as whole-word hidden Markov models with a left-to-right model topology, with no skips over states. We employed mean subtraction, variance normalization, and ARMA filtering (MVA) processing of speech features before modeling [15]. Following the challenge described in [1], here, we report the speech recognition results (percent correct) for the target separated output. For the recognition setup, 39-dimensional Mel-frequency cepstral coefficients (MFCC) including the logarithmic energy were used. Hamming window of 25 ms was used. The frame shift was set to 10 ms.

3. Experiments and Results

In this section, we present the speech recognition results obtained for the sinusoidal SCSS explained earlier. We compare the obtained speech recognition results with those reported by other participants in the challenge.

3.1. Database

The task in the challenge in [1] is to separate the speech mixtures of two speakers drawn from the test dataset composed of 34 speakers. The corpus consists of 34,000 distinct utterances from 34 speakers (18 males and 16 females). The sentences in the database follow a command-like structure with a unique grammatical structure as six word composed of verb, color, preposition, letter, digit and coda such as "*set blue at z five please*". The keywords emphasized for speech recognition task in the challenge are the items in position 4 and 5 referring to letter and digit, respectively.

Table 1: Percent correct for the sinusoidal separation system shown in Figure 1 at different mixing scenarios: same talker, same gender, and different gender.

SSR (dB)	-9	-6	-3	0	3	6
Same Talker	43.9	45.9	49.8	53.6	54.1	57.2
Same Gender	43.8	47.5	48.0	50.0	55.3	57.8
Different Gender	48.5	50.5	54.8	57.5	60.2	63.3
Average	45.4	47.9	50.9	53.8	56.5	59.4

3.2. System Setup

For each speaker, 500 clean utterances are provided for training purposes. The test data is a mixture of target and masker speakers mixed at six SSR levels ranging from -9 dB to 6 dB with a step of 3 dB. For each of the six test sets, 600 utterances are provided of which 200 are for same gender, 179 for different gender, and 221 for same talker. The sentences were originally sampled at 25 kHz. For practical reasons, we decrease the sampling rate to 16 kHz. The speech recognition results to be reported are averaged over all the utterances in the dataset. For speaker identification, we followed the fusion of the two-subsystems setup reported in [11]. The SSR is estimated as a by-product from the SSR-dependent speaker model providing the maximum likelihood.

For the separation setup, we extract features by employing a Hann window of length 32 ms and shift of 8 ms. We use split-VQ based on sinusoidal parameters composed of amplitude and frequency. The source models are divided into magnitude spectrum and frequency parts where each entry is composed of a sinusoidal amplitude vector and several sinusoidal frequency vectors as its candidates (for more details see [10]). According to previous experiments, we set the sinusoidal model order to 100 and for speaker modeling, we use 11 bits for amplitude and 3 bits for frequency part in the sinusoidal coder. The pre-trained speaker codebooks are then used in the test phase to use for the speech separation module. The codebooks are used for both the mixture estimator and double-talk detection blocks (see Figure 1).

3.3. Speech Recognition Results

Detailed results for the speech recognition accuracy are shown in Table 1 for different SSR levels and for different mixing scenarios. The speech recognition score for the sinusoidal separation approach gives an overall recognition accuracy of 52.3%. Comparing this result with those reported in [1], it is observed that the sinusoidal separation approach lies in the median among the others.

We also compare the speech recognition performance of the proposed method with those already reported by several other participants in the challenge [1]. Figure 2 shows the speech recognition accuracy for all participants in the challenge across different SSR levels. From the figure, we conclude that the proposed method achieves a relatively high performance at low SSR levels. Compared to other model-driven methods [2, 4, 6], the method handles low SSR scenarios rather well. This can be because at such low SSR levels, the MMSE estimator in sinusoidal domain performs better than log-max or Wiener filter.

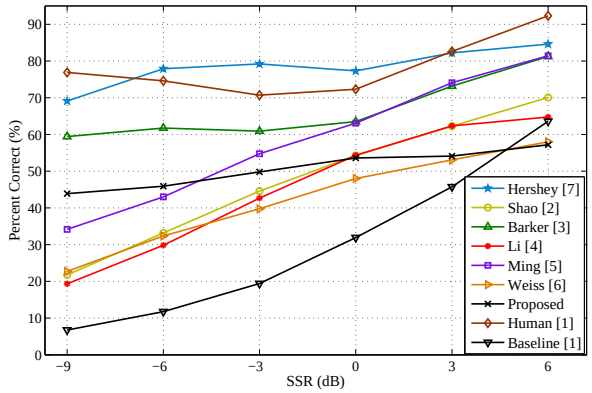

Figure 2: Recognition results (percent correct) of the automatic speech recognition systems entered in the the speech separation and recognition challenge [1].

The other reason for good ASR performance at low SSR levels might be because of strong interference rejection capability of sinusoidal method as reported in [8, 9]. On the other hand, as the SSR increases, the proposed method asymptotically gets close to its best possible performance (quantizer upper bound). This behavior is quite similar to other model-driven methods and to the conclusion in [12] that the performance of a model-driven speech enhancement method is upper-bounded by the quality of the used quantizer (equivalent to when the correct codebook indices are known *a priori*).

4. Conclusions and Future Work

In this paper, we presented the speech recognition performance for a sinusoidal single-channel speech separation (SCSS) system proposed for the speech separation and recognition challenge. The experimental results showed that the proposed system gives a comparable speech recognition performance to other model-driven SCSS methods and is located on the range of median over all other participants in the challenge.

In this work, we only reported the speech recognition results obtained for the task defined in the challenge. Future research should address some of the limitation existing in the current task in the challenge. To name a few, the training samples used to train the speaker models are noise-free and relatively large and the evaluation corpus consists of only digitally added mixtures with constant gains. The challenge also neglects the environmental or background noise effects, as well as the reverberation problem. In practice, each one of these issues and their effect on the overall performance should be carefully studied. Future work should systematically address how these simplifying yet restrictive and impractical pre-assumptions can be relaxed. As an example, recently in [16], a new corpus was provided for noise-robust speech processing research where the goal was to prepare realistic and natural reverberant environments using many simultaneous sound sources.

5. References

[1] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[2] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 77 – 93, 2010.

[3] J. Barker, M. Ning, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 94 – 111, 2010.

[4] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 30–44, 2010.

[5] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 67 – 76, 2010.

[6] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 16–29, 2010.

[7] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, 2010.

[8] P. Mowlaee, M. Christensen, and S. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 5, pp. 1265 – 1277, 2011.

[9] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "Sinusoidal masks for single channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2010, pp. 4262–4266.

[10] P. Mowlaee and A. Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *Proc. European Signal Processing Conf.*, Aug. 2008.

[11] R. Saeidi, P. Mowlaee, T. Kinnunen, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2010, pp. 4545–4548.

[12] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[13] P. Mowlaee, M. G. Christensen, Z. H. Tan, and S. H. Jensen, "A MAP criterion for detecting the number of speakers at frame level in model-based single-channel speech separation," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2010, pp. 538–541.

[14] P. Mowlaee, A. Sayadiyan, and M. Sheikhan, "Optimum mixture estimator for single-channel speech separation," *Proc. IEEE Int. Symposium on Telecommunications*, pp. 543–547, Aug. 2008.

[15] C. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 257–270, 2007.

[16] H. Christensen, J. Barker, N. Ma, and P. Green, "The chime corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. Interspeech*, 2010.