Speech Communication 104 (2018) 1–11

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Single-channel multi-talker speech recognition with permutation invariant training

Yanmin Qian[a,*], Xuankai Chang[a], Dong Yu[b]

[a] Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[b] Tencent AI Lab, Bellevue, USA

ARTICLE INFO

Keywords:
Permutation invariant training
Multi-talker mixed speech recognition
Feature separation
Joint-optimization

ABSTRACT

Although great progress has been made in automatic speech recognition (ASR), significant performance degradation is still observed when recognizing multi-talker mixed speech. In this paper, we propose and evaluate several architectures to address this problem under the assumption that only a single channel of mixed signal is available. Our technique extends permutation invariant training (PIT) by introducing the front-end feature separation module with the minimum mean square error (MSE) criterion and the back-end recognition module with the minimum cross entropy (CE) criterion. More specifically, during training we compute the average MSE or CE over the whole utterance for each possible utterance-level output-target assignment, pick the one with the minimum MSE or CE, and optimize for that assignment. This strategy elegantly solves the label permutation problem observed in the deep learning based multi-talker mixed speech separation and recognition systems. The proposed architectures are evaluated and compared on an artificially mixed AMI dataset with both two- and three-talker mixed speech. The experimental results indicate that against the state-of-the-art single-talker speech recognition system our proposed architectures can cut the word error rate (WER) by relative 45.0% and 25.0% across all speakers when their energies are comparable, for two- and three-talker mixed speech, respectively. To our knowledge, this is the first work on the single-channel multi-talker mixed speech recognition on the challenging speaker-independent spontaneous large vocabulary continuous speech task.

1. Introduction

Thanks to the significant progresses made in recent years (Yu et al., 2010; Seide et al., 2011; Hinton et al., 2012; Dahl et al., 2012; Abdel-Hamid et al., 2012; 2014; Yu and Deng, 2014; Sainath et al., 2015; Bi et al., 2015; Qian et al., 2016; Qian and Woodland, 2016; Mitra and Franco, 2015; Peddinti et al., 2015; Sercu et al., 2016; Amodei et al., 2016; Zhang et al., 2016a; Yu et al., 2016; Xiong et al., 2017), ASR systems have now surpassed the threshold for adoption in many real-world scenarios and have enabled services such as Microsoft Cortana, Apple's Siri and Google Now, where close-talk microphones are commonly used.

However, current ASR systems still perform poorly when far-field microphones are used. This is because many difficulties hidden by close-talk microphones now surface under distant recognition scenarios. For example, the signal to noise ratio (SNR) between the target speaker and the interfering noises is much lower than that when close-talk microphones are used. As a result, the interfering signals, such as background noise, reverberation, and speech from other talkers, become so distinct that they can no longer be ignored.

In this paper, we aim at solving the speech recognition problem when multiple talkers speak at the same time and only a single channel of mixed speech is available. Although multi-channel speech processing is important for many applications now, it is still very necessary (and interesting) to conduct research on single-channel multi-talker speech recognition for four reasons. First, many recording devices, such as those used by reporters, only have one microphone. Second, even with microphone array, single-channel multi-talker ASR is still needed when the two speakers are in the same direction and thus cannot be separated by a beamformer. Third, single-channel results set a lower-bound on what is achievable when multi-channel information is available. Fourth, it sheds lights on new solutions other than beamformer when multi-channel information (esp. with ad hoc mic-array) is available.

Many attempts have been made to address the problem of single-channel multi-talker speech recognition. Before the deep learning era, the most famous and effective model is the factorial GMM-HMM, which outperformed humans in the 2006 monaural speech separation and recognition challenge (Kristjansson et al., 2006; Hershey et al., 2010; Ming et al., 2010; Cooke et al., 2010). The factorial GMM-HMM, however, requires the test speakers to be seen during training so that the

* Corresponding author.
E-mail address: yanminqian@sjtu.edu.cn (Y. Qian).

https://doi.org/10.1016/j.specom.2018.09.003
Received 13 November 2017; Received in revised form 13 July 2018; Accepted 3 September 2018
Available online 7 September 2018

interactions between them can be properly modeled. Recently, several deep learning based techniques have been proposed to solve this problem (Weng et al., 2015; Hershey et al., 2016a; Isik et al., 2016; Yu et al., 2017b; Kolbaek et al., 2017; Chen et al., 2017; Chang et al., 2018a; Tan et al., 2018; Chen and Droppo, 2018; Qian et al., 2018; Chen et al., 2018; Chang et al., 2018b). The core issue that these techniques try to address is the label ambiguity or permutation problem (refer to Section 3 for details).

In Weng et al. (2015) a deep learning model was developed to recognize mixed speech directly. To solve the label ambiguity problem, Weng et al. assigned the senone labels of the talker with higher instantaneous energy to output one and the other to output two. Although this addresses the label ambiguity problem, it causes frequent speaker switch across frames. To deal with the speaker switching problem, a two-speaker joint-decoder with a speaker switching penalty was used to trace speakers. This approach has two limitations. First, energy, which is manually selected, may not be the best information to assign labels under all conditions. Second, the frame switching problem introduces burden to the decoder.

In Hershey et al. (2016a); Isik et al. (2016) the multi-talker mixed speech is first separated into multiple streams. An ASR engine is then applied to these streams independently to recognize speech. To separate the speech streams, they proposed a technique called deep clustering (DPCL) (Hershey et al., 2016a). They assume that each time-frequency bin belongs to only one speaker and can be mapped into a shared embedding space. The model is optimized so that in the embedding space the time-frequency bins belonging to the same speaker are closer and those of different speakers are farther away. The work in Isik et al. (2016) further introduced an enhancement network to refine the DPCL output, in which the soft mask can be obtained to achieve a better reconstruction performance. During evaluation, a clustering algorithm is first used upon embeddings to generate a partition of the time-frequency bins, and then the separated audio streams are reconstructed based on the partition. In these works, the speech separation and recognition are usually two separate components.

Chen et al. (2017) proposed a similar technique called deep attractor network (DANet). Following DPCL, their approach also learns a high-dimensional embedding of the acoustic signals. Different from DPCL, however, it creates cluster centers, called attractor points, in the embedding space to pull together the time-frequency bins corresponding to the same source. The main limitation of DANet is the requirement to estimate attractor points during evaluation time and to form frequency-bin clusters based on these points.

In Yu et al. (2017b), Kolbaek et al. (2017), Chang et al. (2018a), Tan et al. (2018), Chen and Droppo (2018), Qian et al. (2018), Chen et al. (2018) and Chang et al. (2018b), a simpler yet equally effective technique named permutation invariant training (PIT) was proposed to address the speaker independent multi-talker speech separation problem. In PIT, the source targets are treated as a set (i.e., order is irrelevant). During training, PIT first determines the output-target assignment with the minimum error at the utterance level based on the forward-pass result. It then minimizes the error given the assignment. This strategy elegantly solved the label permutation problem. However, in these original works PIT was used to separate speech streams from mixed speech. For this reason, a frequency-bin mask was first estimated and then used to reconstruct each stream. The minimum mean square error (MMSE) between the true and reconstructed speech streams was used as the criterion to optimize model parameters. It is noted that a similar permutation free technique was also proposed in Hershey et al. (2016a) but with negative results and conclusions; and this idea was also used in Isik et al. (2016) but within the DPCL framework which is more complex.

Moreover, most previous works on single-channel multi-talker speech still focus on *speech separation* (Hershey et al., 2016a; Isik et al., 2016; Chen et al., 2017; Yu et al., 2017b; Kolbaek et al., 2017). In contrast, single-channel multi-talker *speech recognition* is much harder and there is less related work. There have been some attempts, but the related tasks are relatively simple. For example, the 2006 monaural speech separation and recognition challenge (Cooke et al., 2010; Hershey et al., 2010; Rennie et al., 2010; Weng et al., 2015) was defined on a speaker-dependent, small vocabulary, constrained language model setup, while in Isik et al. (2016) a medium vocabulary reading style corpus was used. We are not aware of any extensive research work on the more real, speaker-independent, spontaneous large vocabulary continuous speech recognition (LVCSR) on single-channel multi-talker mixed speech before our work.

In this paper, we attack the multi-talker mixed speech recognition problem with a focus on the speaker-independent setup given just a single-channel of the mixed speech. Different from Hershey et al. (2016a), Isik et al. (2016), Yu et al. (2017b) and Kolbaek et al. (2017), here we extend and redefine PIT over log filter bank features and/or senone posteriors. In some architectures PIT is defined upon the minimum mean square error (MSE) between the true and estimated individual speaker features to separate speech at the feature level (called PIT-MSE from now on). In some other architectures, PIT is defined upon the cross entropy (CE) between the true and estimated senone posterior probabilities to recognize multiple streams of speech directly (called PIT-CE from now on). Moreover, the PIT-MSE based front-end feature separation can be combined with the PIT-CE based back-end recognition in a joint optimization architecture. We evaluate our architectures on the artificially generated AMI data with both two- and three-talker mixed speech. The experimental results demonstrate that our proposed architectures are very promising and flexible. Note that compared to our previous preliminary attempt in Yu et al. (2017a), this paper gives more comprehensive and detailed exploration, and other architectures are also developed and compared in this work: PIT is performed not only on the front-end feature separation module to obtain better separated feature streams but also on the back-end recognition module to predict the separated senone posterior probabilities directly. Moreover, PIT can be implemented on both the front-end and back-end with a joint-optimization architecture. Then a comprehensive experiment is designed and compared to evaluate the performance of different PIT based architectures for multi-talker speech recognition, and the further improvement and analysis is performed for the proposed framework with the evaluation on both artificially generated multi-talker AMI and WSJ0 corpus.

The rest of the paper is organized as follows. In Section 2 we describe the speaker independent single-channel multi-talker mixed speech recognition problem. In Section 3 we propose several PIT-based architectures to recognize multi-streams of speech. We report experimental results in Section 4 and conclude the paper in Section 5.

2. Single-channel multi-talker speech recognition

In this paper, we assume that a single-microphone signal $y[n]$ is observed. $y[n]$ is a mixture signal and assumed to be a linear combination of multiple speech sources, i.e. $y[n] = \sum_{s=1}^{S} x_s[n]$, where $x_s[n], s = 1, \cdots, S$ are S streams of speech sources from different speakers. Our goal is to separate these streams and recognize every single one of them. In other words, the model needs to generate S output streams, one for each source, at every time step. However, given only the mixed speech $y[n]$, the problem of recognizing all streams is under-determined because there are an infinite number of possible $x_s[n]$ (and thus recognition results) combinations that lead to the same $y[n]$. Fortunately, speech is not a random signal. It has patterns that we may learn from a training set of pairs $\mathbf{y}$ and $\ell^s, s = 1, \cdots, S$, where ℓ^s is the senone label sequence for stream s, i.e. the senone alignment obtain on the original single-talker speech.

In the single speaker case, i.e., $S = 1$, the learning problem is significantly simplified because there is only one speaker's stream that needs to be recognized, thus it can be cast as a simple supervised optimization problem. Given the input to the model, which is some feature represen-