USING AN ATTENTION MECHANISM IN DISCRIMINATIVE AUDIO SOURCE SEPARATION

Author(s) Name(s)

Author Affiliation(s)

ABSTRACT

In this paper, we explore utilizing an attention mechanism in a neural network model for discriminative audio source separation. The attention mechanism enables the model to infer a template vector which best helps to boost the source separation performance. We test our hypothesis on male-female speech mixtures from the TIMIT dataset, and observe that the proposed methodology helps to increase the source-to-distortion ratio on test mixtures.

Index Terms— Audio source separation, neural networks, attention mechanism

1. INTRODUCTION

In audio source separation the goal is to recover the original audio signals in an additive mixture signal. The discriminative source separation models are trained in such a way to directly estimate the underlying sources [**?**] conditioned on the mixture, as opposed to generative source separation models [**?**] which first train a generative model separately on the mixtures, which are later to be used for separation [**?**].

The attention mechanism was first introduced in the context of automatic machine translation [**?**], and then later on used in a variety of computer vision tasks [**?**] and audio tasks [**?**]. In this paper, we explore using an attention mechanism in a neural network model to infer *template* vectors in order to enhance the separation performance vectors. In each time instant, the attention mechanism computes a probability vector corresponding to each template. A weighted sum of the chosen templates are then fed into a separator network in addition to the mixture spectrum.

We perform an extensive search over possible neural network architectures as well as a search over possible hyperparameter configurations. We test on performing source separation on speech mixtures created using the TIMIT [**?**] dataset. We observe that the proposed attention mechanism enables a performance boost in terms of source-to-distortion ratio [**?**].

Thanks to XYZ agency for funding.

2. DISCRIMINATIVE SOURCE SEPARATION

Given a mixture signal x^{mix}, the discriminative source separation system $f_\theta = [f_{\theta^1}, f_{\theta^2}]$, outputs the estimates $\hat{x}^1 := f_{\theta^1}(x^{\text{mix}})$, $\hat{x}^2 := f_{\theta^2}(x^{\text{mix}})$ for the original sources x^1 and x^2. The system is trained with the following cost function:

$$\min_\theta \sum_n \left(\|f_{\theta^1}(x_n^{\text{mix}}) - x_n^1\| + \|f_{\theta^2}(x_n^{\text{mix}}) - x_n^2\| \right), \quad (1)$$

where for the sake of simplicity we have omitted the time index to reduce clutter. The neural networks f_{θ^1}, f_{θ^2} can be chosen as multilayer perceptrons, convolutional networks or recurrent neural networks [**?**]. The diagram of the system is given in Figure **??**. Note that in this paper we are restricting our system to predict two sources, the system described above can be generalized to more than two sources.

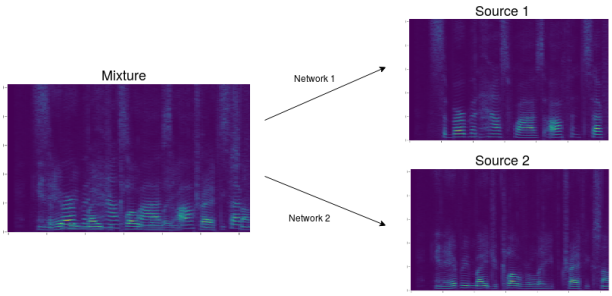

Fig. 1. Discriminative source separation diagram for a two source system

3. ATTENTION MECHANISM IN DISCRIMINATIVE SOURCE SEPARATION