Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 95 – 100

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

A Temporal Coherence Loss Function for Learning Unsupervised Acoustic Embeddings

Gabriel Synnaeve[a,b,*], Emmanuel Dupoux[b]

[a] *Facebook A.I. Research, Paris, France*
[b] *École Normale Supérieure / PSL Research University / EHESS / CNRS, France*

Abstract

We train neural networks of varying depth with a loss function which imposes the output representations to have a temporal profile which looks like that of phonemes. We show that a simple loss function which maximizes the dissimilarity between near frames and long distance frames helps to construct a speech embedding that improves phoneme discriminability, both within and across speakers, even though the loss function only uses within speaker information. However, with too deep an architecture, this loss function yields overfitting, suggesting the need for more data and/or regularization.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: unsupervised learning; speech embeddings; speech recognition; temporal coherence; zero resource speech challenge; feature extraction

1. Introduction

Deep Neural Networks (DNNs) are becoming the dominant paradigm for speech technologies, regularly breaking the state of the art obtained previously with Hidden Markov Models and signal processing systems [1,2]. However, being more powerful, DNNs are also more hungry in human annotations: commercially deployed systems require supervised training on thousands of hours of human annotated data. Yet, there are situations where human annotations are not available or two expensive to gather. Half of the human languages, for instance, have no writing system. In addition, the fact that human infants can spontaneously learn their native language through mere immersion in a linguistic environment, shows that it is theoretically possible to learn acoustic and language models with little or no human labels. It is therefore of both practical and theoretical interest to explore the so-called "zero-resource" setting [3,4,5] where linguistic structures are been learned from large amounts of unannotated data.

Here, we examine the idea that useful representations can be learned in a DNN architecture using only generic knowledge about the temporal distribution of phonetic structure. Typically, in any human language, the building blocks of words (phones) have a duration of approximatively 60-150ms [6,7]. Therefore representations with a temporal

* Corresponding author.
E-mail address: gabriel.synnaeve@gmail.com

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.035

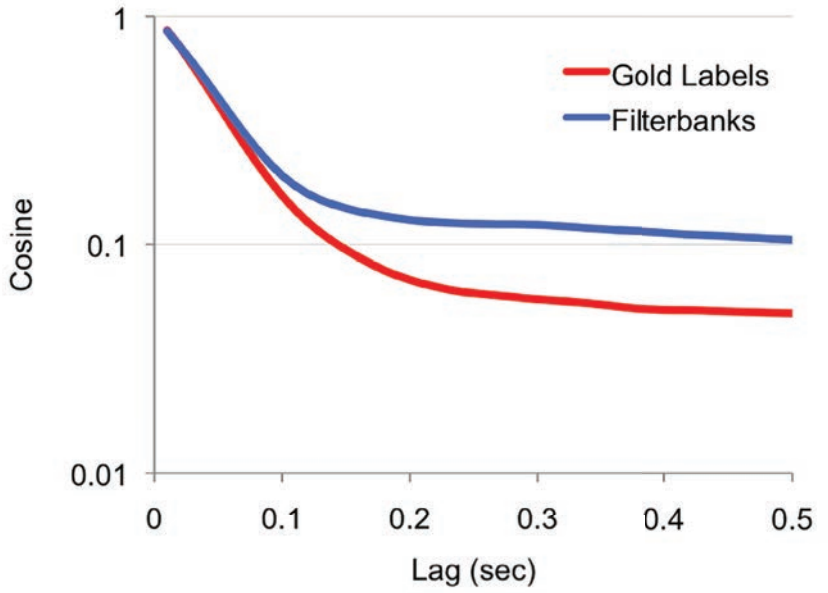

Fig. 1. Cosine similarity between different frames separated by a given time lag. In red is shown the gold labels (cosine of 1 when the labels are the same, and zero when not), in blue is shown the cosine between frames of filterbank values. The scale on the y axis is logarithmic.

profile that is either much smaller or much larger are likely to be not very useful for the purpose of word recognition. The typical duration of phonemes is illustrated in Figure 1, where we have plotted the average cosine similarity between short stretches of speech (frames) separated by different lags. As one can see, the similarity is a decreasing function of lag. On a "gold" phoneme representation (each 10ms frame is represented by a binary N-dimensional vector, where each dimension codes for one of the phoneme classes), the average cosine can be interpreted as the propensity for two frames separated by a given lag to belong to the same phoneme class. When the same plot is done on filterbank representations (each frame is composed of 40 Mel-frequency spectral coefficients over an Hamming window of 25ms, with a step size of 10ms), the same general curve is obtained, but the drop is less steep than for 'gold' labels. This is due to the fact that filterbank representations encode information that change less quickly than phoneme (e.g. information relevant to talker identity), and therefore display more long distance similarity than abstract, talker invariant labels do.

In this paper, we therefore propose a loss function for training DNNs that minimizes the difference between embeddings similarities at short lags (where frames are likely to belong to the same phoneme), and maximizes (shatters them) at long lags (where frames are likely to belong to different phonemes). We implement this idea within a siamese network architecture following our previous work on ABnets[8].

2. Related Work

The idea of using the temporal structure to evaluate speech representations has been proposed in[9]. In this study, the M-measure, defined as a between-frame correlation between posteriogramms of a speech recognizer can be used to predict the word error rate. In a more recent paper[10], the M-measure is modified to become the M-delta, which amounts to the difference in M-measure between a long and short lag correlogram. Models of the temporal structure of phonemes have also be used in systems that try to learn acoustic models from raw speech. As shown in[11], unsupervised clustering applied to continuous speech results in units smaller to phonemes (more akin to phone states). Other studies[12,13] improved on this idea by imposing three-state HMM structure to phonemes, the clustering being done using a Chinese restaurant process. Similarly,[14] model clusterized (unsupervised) phoneme states by splitting word chunks into phoneme-size units using the average duration of the phonemes.