VAD, Wavelet based Data augmentation Submission to the Shared Task on Automatic Speech Recognition for Non-Native Children's Speech

Shreeharsha B S

Abstract

This paper describes the Voice Activity Detection (VAD), Wavelet based data augmentation techniques used in the submission to the Shared Task on Automatic Speech Recognition for Non-Native Children's Speech challenge. The augmentation techniques are performed with the help of a VAD module and wavelet filtering methods. These techniques were evaluated using the baseline system provided for the shared task (nnet3 TDNN i-vector Kaldi recipe) with minor changes to the decoding parameters. The result was an 8.48% absolute improvement in the Word Error Rate (WER) on the evaluation dataset and 10.1% absolute improvement in WER on the development dataset.

1 Introduction

One of the many challenges involved in speech recognition in the setting of children and second language learners is the wide variety of accents and the lack of large datasets. The motivation behind examining the techniques described here is to create background noise suppressed and background noise enhanced samples of an existing, well transcribed dataset of relatively smaller size. These new data points then constitute a larger dataset on which conventional ASR systems can be used. In this work, the background noise is identified using a VAD module that comprises of an adaptive linear energy detector (ALED) [1] and a zero frequency filter (ZFF)[2], both of which exploit different properties of the non-speech regions (energy and periodicity).

Using the power spectrum of these noisy regions, the existing recordings are then filtered by constructing wavelets from a non-speech power spectrum that is unique to every recording. The expectation, here, is that these new recordings help the system identify and learn a wider variety and severity of background noise conditions using the noise-enhanced recordings and also learn the actual phones uttered using the noise-suppressed portion of the recordings. The details of the experiments carried out are described next.

2 Details of the Augmentation Procedure

2.1 Voice Activity Detector (VAD)

The pre-processing VAD module produces speech, non-speech decisions for every 10 ms of a recording using a 20 ms window. The VAD is comprised of a combination of adaptive energy detector and a zero frequency filter both of which work in parallel to produce speech/non-speech (1/0) decisions. These decisions are then combined using certain temporal context rules and other temporal restrictions (non-speech regions less than 200 ms are re-cast as speech back into the surrounding speech regions) by ORing or ANDing the individual decisions (1/0) of every frame.

2.2 Wavelet Construction

Using the non-speech regions identified using the VAD module, an overall power spectrum of the background noise is determined by folding non-contiguous blocks of non-speech regions on top of each other and summing them. These power spectra, $(H_0(z))$, are unique to every recording and have been empirically determined to not vary highly with the sizes of the block used. From these power spectra, scaling and wavelet functions are constructed by approximating the following convergence conditions in L2($\mathbb{R}$) [3]:

- Scaling function

$H_0^{(i)}(z) = \prod_{p=0}^{i-1} \left(H_0\left(z^{2^p}\right)\right)$

- Wavelet function

$H_1^{(i)}(z) = H_1\left(z^{2^{i-1}}\right) \prod_{p=0}^{i-2} \left(H_0\left(z^{2^p}\right)\right), i = 2, 3, ..$

where $H_1^{(1)}(z) = H_1(z)$ which is the power complementary spectrum of the corresponding background power spectrum $H_0(z)$.

The scaling function, $\phi(t) = \lim_{i\to\infty} h_0^{(i)}(t)$, is guaranteed to exist provided the limit converges (here $h_0(t)$ is the time domain signal corresponding to $H_0(z)$). $\phi(t)$ is approximated by restricting i to be finite, which in turn is determined by the resolution of the DFT (Discrete fourier transform) used.

2.3 Augmentation Procedure

With the initial 9 hours of training data that were very well transcribed, the procedure described in Section 2.2 is used to construct approximate and detail versions of the recordings in the training set at 2 levels using the stationary wavelet transform. This results in 27 hrs of new data. The transcripts for these 27 hrs is kept the same and the system is re-trained on the baseline model using all 36 hrs. This will be called System I.

The additional 40 hrs of training data that was released later are not as well transcribed as the previous 9 hrs. Using some intermediate measures from the VAD module and the average spectral tilt, around 5500 recordings were identified as being highly noisy(presence of coughing noises, mic pops and bursts etc) in the 40 hrs dataset. These recordings' ground truth transcripts were updated

by running the recordings through System I, with higher acoustic weights and low language model weight while decoding.

Then the entire set of 40 hrs was again filtered using the procedure in Section 2.2 at 1 level, to generate 80 hrs of data in total. These recordings along with the 36 hrs from System I were used to retrain the baseline model again which was the final system used in the challenge. This will be called System II.

3 Results and Future work

The results of our work are as follows: Reduced baseline WER from 37.6% to 31.5% using System I on the development dataset. Redcued WER from 31.5% to 27.5% using System II on the development dataset. System II produced a WER of 26.61% on the final evaluation dataset used in the challenge.

Future work should involve formalizing certain aspects of the techniques used and examining its effectiveness in combination with other ASR models and scenarios.

References

[1] K. Sakhnov, E. Verteletskaya, and B. Simak, "Dynamical energy-based speech/silence detector for speech enhancement applications," in *Proceedings of the world congress on engineering*, vol. 1. Citeseer, 2009, p. 2.

[2] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[3] M. Sharma, A. V. Vanmali, and V. M. Gadre, "Construction of wavelets: Principles and practices," *Wavelets and Fractals in Earth System Sciences*, p. 29, 2013.