Machine Learning Engineer Nanodegree

Capstone Proposal

Sen Li
March 16th, 2018

Proposal

Domain Background

Until a few years ago, the quality of voice telecommunications has been limited by design choices made over 100 years ago, which resulted in an 8 kHz sampling rate being used and in a practical frequency range of 300 – 3400 Hz. This so called narrowband (NB) frequency range severely limited speech quality. Recently, the industry has started to move to “HD voice” and “Ultra HD voice”, for example, the use of wideband (WB) or super-wideband (SWB) speech coders respectively, which use a sampling rate of 16 kHz or 32 kHz and correspond to a frequency range of 50-7000 Hz or 50 –14000 Hz respectively [1][2].

However, these deployments are not ubiquitous. A whole new infrastructure is needed to support these WB and SWB coders, at a substantial cost. It will likely take years before complete coverage is achieved. Until then, a significant proportion of calls will still use legacy narrowband. Further, it is likely that landline upgrades to WB or SWB will take even longer, meaning that even when the mobile networks have fully migrated to higher bandwidths, calls from landlines will still be narrowband.

Blind bandwidth extension (BBE) technology aims at solving this problem by transforming NB speech into WB or SWB speech. Typically using some form of either spectral folding or statistical modelling, the 4-8 kHz part of a speech signal is predicted from the 0-4 kHz part, to generate a signal having the general characteristics of wideband speech [3][4]. While perfect prediction cannot be expected, reasonably high quality speech can be obtained. In this project, we will focus on predicting the 4-8 kHz portion of speech, usually referred to as the high-band (HB), from the 0-4 kHz portion, known as the low-band (LB).

Problem Statement

Various approaches to BBE have been proposed and studied. Vector Quantization (VQ) codebook mapping is one of the classical method, which creates discreet mapping of speech parameters from LB to HB [5][6]. Gaussian Mixture Models(GMM) based method are used to preserve a more accurate transformation between LB and HB by modeling the speech envelope parameter continuously [7]. Hidden Markov Model (HMM) was the extension of GMM to improve the quality during speech transition by exploiting speech temporal information [8].

Recent advancement in neural networks learning, especially deep learning, suggested that such framework may have the potential to model more complex non-linear relationship between speech LB and HB, which leads to our proposal of this project. We will study the neural networks approach for blind bandwidth extension - from speech LB features to accurately predict speech HB features, such as spectral shape. Mean squared error (MSE), which is widely used for many regression modeling problems, can be used as the error metric between predicted HB features and target HB features in our project.

Datasets and Inputs

Since the BBE should be a generic speech enhancement algorithm, it should perform equally well for both male and female, and across various talkers and different languages. The ideal dataset for this project would be a multi-lingual speech database that contains multiple talkers and covers many languages. We decided to use the NTT 1994 multi-lingual corpus, containing 21 languages, 4 female and 4 male talks for each of the language [9]. Unfortunately, this speech corpus is not publicly available for free and the size of dataset is very large, we will extract the speech features from the raw speech signal for the project purpose. For our test inputs, we will evaluate the BBE performance on ITU P.501 British English test signal [10].

Solution Statement

In this project, we will build and implement a BBE system based on deep neural networks. The training data would be speech features calculated from NB speech and WB speech respectively. The neural network model will be trained based on NB spectral features as input from NTT 1994 corpus, and predict the corresponding HB spectral features. We will evaluate the model on unseen speech data from P.501 British English test speech. We will adopt the MSE in feature domain as the metric to quantify model performance. We hope the deep neural networks based model would be able to capture more complex non-linear relationship between speech LB and HB and could yield better prediction accuracy.

Benchmark Model

We plan to train a couple of benchmark models given the same input LB spectral features and the output HB spectral features data with classical machine learning approach - ranging from linear regression, decision trees and ensemble methods.

We also plan to implement the widely used VQ codebook approach as one of the benchmark model for this particular BBE problem. For the VQ codebook benchmark model, we will train a codebook for low-band spectral features concatenated with high-band spectral features using K-Means algorithm. In the prediction phase, we will take the spectral features from NB speech signal and calculated the 3-nearest neighbor codebook entries based on the spectral feature distance and use such distance-weighted average to synthesize the corresponding high-band spectral features. The model performance results will also be measured by the MSE in the spectral feature domain.

Evaluation Metrics

We will adopt the mean squared error as our project evaluation metric. We compare the predicted high-band spectral features with reference high-band spectral features extracted from the true wideband speech. The lower the score, the better the prediction model perform. the mathematic expression is,

$error = (y_p - y)^2$, where y_p is the predicted output, and y is the reference output, or the ground truth.

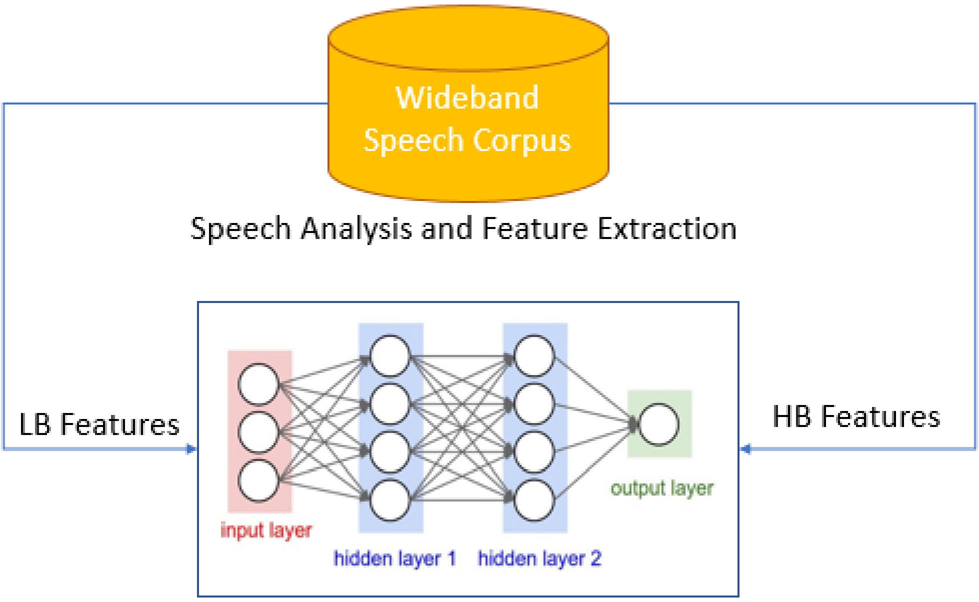
Project Design

Given that we are using the multi-lingual NTT 1994 corpus and it is not publicly available, we will need to perform pre-processing of the dataset and extract the speech features that we can directly use as the input and output training data for our project.

The wideband speech data from the corpus is sampled at 16 kHz sampling rate and digitized into 16-bit resolution.The ITU P.341 Tx filter is applied to the wideband speech to simulate the typical Tx response in the telecommunication system before speech parameter extraction for both low-band and high-band. Given the parameterized speech data, we prepares the training and validation data with classical 10-fold cross validation scheme for training.

The same pre-process and feature extraction procedure will also be applied to the test input speech, which is from ITU P.501 British English test signal. Since the original P.501 English test signal is sampled at 48 kHz, a 3:1 down sampling is required to convert it to wideband speech, sampled at 16 kHz and a further 2:1 down sampling is required to convert it to narrowband speech, sampled at 8 kHz. All the sampling rate conversion can be achieved using standardized ITU G.191 STL speech tools [11].

One problem still remains though, is for general speech related problems, especially for clean speech recordings, the background silence is not the point of interest. Voice activity detection (VAD) algorithms, which is implemented in many standardized speech coders, can effectively remove most of the silence within the recording, so that the feature that we extracted is truly representing the active speech.



Given that we have all the parameterized speech feature data after pruning out the silence, we can train a neural network model to predict from LB spectral features to LB spectral features, as shown in the Figure above. We plan the use the classical spectral representation of speech - line spectral frequency (LSF), which is widely used in speech coding and speech enhancement. Another widely used spectral feature is Mel frequency cepstrum coefficient (MFCC), which is widely used in speech recognition and synthesis for acoustic modeling. For the low-band or narrowband speech, a 10th order LPC analysis is sufficient enough to capture the speech spectral shape between 100Hz to 4kHz. For the high-band speech, a 6th order LPC analysis is more than enough to characterize spectral shape above 4kHz. As a results, the input low-band LSFs will have dimension of 10, and the output high-band LSFs will have the dimension of 6 respectively.

For the benchmark models, we plan to use the model utility from scikit-learn python library and train several classical machine learning models, including linear regression, regression decision trees an ensemble repressor. For the VQ codebook benchmark, we plan to use K-Means function from the scikit-learn python library as well to perform data clustering.

For the neural network framework, we plan to use Keras as our development tool and start with shallow neural network with only 1 hidden layer, 256 neurons will be used for the hidden layer. We will evaluate the performance based on this configuration and if such configuration achieves reasonable results, we will follow up by implementing wider and deeper neural networks to see if the prediction error gets even better with more complex models.

Since speech is a typical time series signal by its nature, we suspect that with the help of temporal and context information, the prediction accuracy might be further improved. We will further investigate the impact of transitioning from feed forward to recurrent type of framework, we will evaluate both LSTM and Bi-directional LSTM in particular in out project.

For this project, we always use the mean squared error as the metric to represent the performance of the prediction model, the lower the score, the better the model.

Reference

[1] 3GPP TS 26.190, “Adaptive multi-rate wideband (AMR-WB) speech codec; Transcoding functions,” 3rd Generation Partnership Project, Sept. 2012, version 11.0.0.

[2] 3GPP TS 26.441, “Codec for Enhanced Voice Services (EVS); General overview,” 3rd Generation Partnership Project, Dec. 2015, version 13.0.0.

[3] H. Carl and U. Heute,“Bandwidth enhancement of narrow-band speech signals,” in Proc. EUSIPCO, vol.2, Edinburgh, UK, Sept. 1994, pp. 1178–1181

[4] H. Pulakka and P. Alku,“Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum,”IEEE Trans. Audio, Speech, Language Process., vol. 19,no. 7, pp. 2170–2183, Sept. 2011

[5] Qian, Y. & Kabal, P.—Wideband speech recovery from narrowband speech using Classified codebook mapping, Proceedings of the 9th Australian International Conference on Speech Science & Technology Melbourne, December 2 to 5, 2002.

[6] J. Epps and W. H. Holmes,“A new technique for wideband enhancement of coded narrowband speech,” in Proc. IEEE Workshop Speech Coding, 1999, pp. 174–176.

[7] K.-Y. Park and H. S. Kim,“Narrowband to wideband conversion of speech using gmm based transformation,”in Acoustics, Speech, and Signal Processing, 2000. ICASSP’00. Proceedings. 2000 IEEE International Conference on, vol. 3. IEEE, 2000, pp.1843–1846.

[8] P. Jax and P. Vary,“Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model,” in Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on, vol. 1. IEEE, 2003, pp. I–680.

[9]N. A. T. Corporation, “Multi-lingual speech database for telephonometry,” http://www.ntt-at.com/product/multilingual/

[10] ITU-T P.501, “Test signalsfor use in telephonometry,” Int. Telecommunication. Union, Jan. 2012 https://www.itu.int/rec/T-REC-P.501/en

[11] ITU-T G.191, “Software tools for speech and audio coding standardization,” https://www.itu.int/rec/T-REC-G.191/en.