homework\_1-3

1. In speech perception, speech signal is decomposed into a number of frequency bands. Please explain the deference between human perception（the auditory models） and Fourier transformation.

First of all, the human hearing range is commonly given as 20 to 20,000 Hz, while Fourier transformation does not put such restrict on the frequency domain.

One of the main drawbacks of Fourier transformation is that the frequency bins are linear. In contrast, the human ear responds to frequency logarithmically, not linearly, thus having different resolutions of different frequency ranges while Fourier transformation keeps a constant resolution. It is worth knowing that most of the frequencies that are of concern to us tend to be below 8 kHz. The Fourier transformation’s linearity can lead to the effect that much of the Fourier transformation data is wasted on recording high-frequency information very accurately, at the expense of the low-frequency information that is generally more useful in a speech context. In respect to that, the auditory models of humans are more sensitive to lower frequencies and are suitable to be applied in the scenarios related to human voice.

There are another several powerful functions of human perception system. Like, human’s auditory model can mask the sound that are not interested by human in the presence of multiple sounds, hence being able to do perception and separation simultaneously.

1. Please explain the relation of the linear spectral frequency (LSF) and the linear predictive coding (LPC).

Linear spectral frequency (LSF) uniquely represent the linear predictive coding (LPC) filter of a speech frame. Linear spectral frequencies have several properties (e.g. smaller sensitivity to quantize noise) that make them superior to direct quantization of linear predictive coding (LPC). For this reason, LSFs are very useful in speech coding. LSF’s encode speech spectral information more efficiently than other transmission parameters. This can be attributed to the intimate relationship between the LSF’s and the formant frequencies. Accordingly, LSF’s can be quantized taking into account spectral features known to be important in perceiving speech signals. In addition, LSF’s lend themselves to frame-to-frame interpolation with smooth spectral changes because of their frequency domain interpretation.

Linear predictive coding (LPC) is a method used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. LPC starts with the assumption that a speech signal is produced by a buzzer at the end of a tube (for voiced sounds), with occasional added hissing and popping sounds (for voiceless sounds such as sibilants and plosives). Although apparently crude, this model is actually a close approximation of the reality of speech production. LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the residue.

3. The LPCNet uses the traditional algorithm of LPC to increase the calculation effectiveness. Do you have any idea to combine other traditional algorithms into the neural network?

Mel-Scale Frequency Cepstral Coefficients (MFCC) are widely used in various speech processing techniques, especially commonly used as features in automatic speech recognition. One key point is that the front-end processing of extracting the MFCC is considerably simple. The basic procedure to develop MFCCs is the following:

1) Convert the frequency domain from Hertz to Mel Scale

2) Take logarithm of Mel representation of audio

3) Take logarithmic magnitude and use Discrete Cosine Transformation

4) This result creates a spectrum over Mel frequencies as opposed to time, thus creating MFCCs.

Leveraging MFCCs is a fantastic way to process audio such that various Deep Learning and Machine Learning problems can learn from the recorded sounds.

The generalized cross-correlation with phase transform (GCC-PHAT) is the most popular method for estimating the time difference of arrival (TDOA) between microphones, which is an important clue for sound source localization (SSL). GCC-PHAT is computed as the inverse Fourier transform of a weighted version of the cross-power spectrum (CPS) between the signals of two microphones. The TDOA estimate is then obtained by finding the time-delay between the microphone signals which maximizes the GCC-PHAT function. Since the GCC-PHAT is a good high-level feature that represents interaural time difference, there are so many neural network based SSL systems taking the full GCC-PHAT function as the input feature and exhibiting substantial performance on DOA estimation.

homework\_4-7

4. Please describe the relationship between GMM-HMM and DNN-HMM based speech recognition systems. How to calculate P(X|W) by DNN?

First of all, both the GMM-HMM and DNN-HMM are used as the acoustic model in the speech recognition system. They play the same role in determining P(X|W) (the probability of the feature vectors given the recognized words).

Plus, both of them use HMM to make a transition from its current state (phoneme) to one of its connected states every time step.

One of the differences is that the observation probabilities are generated from GMM or DNN. In GMM-HMM, every state has a probability distribution described by the GMM. In contrast, DNN-HMM estimates the posterior probability of each state from the sequence of acoustic feature.

Another main difference is that GMM-HMM has its limitations in modeling the continuous-time speech signal, while DNN-HMM can easily model correlated features. The reason is that unlike GMM-HMM uses the feature from one single frame as input, DNN-HMM is not built on the hypothesis that the acoustic features from different frames are independent of each other, thus can utilizing the context information from multiple continuous frames.

At last, both the GMM-HMM and DNN-HMM may share the same decoding algorithm, like Viterbi algorithm.

When it comes to how to calculate P(X|W) by DNN, first we need to generate labels using a trained GMM-HMM, then train the DNN to associate a phone label with a frame of acoustic feature using the cross-entropy loss function to optimize. Finally, the DNN outputs the posterior P(W|X), then Bayes Rule can be applied here to calculate the P(X|W): P(X|W)=.

1. Please summarize the advantages and disadvantages of DNN-HMM and end-to-end ASR system, and explain the main solutions to address these disadvantages of end-to-end ASR system.

Advantages of DNN-HMM:

1) The DNN-HMM can utilize the context information of frames by estimateing the posterior probability of each state from the sequence of acoustic feature and is stable processing to estimate phoneme states in a frame-by-frame manner.

2) The training can be performed using the Viterbi algorithm and the decoding is generally quite efficient.

Disadvantages of DNN-HMM:

1) The training process is complex and difficult to be globally optimized. HMM-based model often uses different training methods and data sets to train different modules. Each module is independently optimized with their own optimization objective functions which are generally different from the true LVCSR performance evaluation criteria. So the optimality of each module does not necessarily mean the global optimality. That being said, the cascading processing of DNN-HMM may lead to omitting global optimization and get local optimization instead.

2) Conditional independent assumptions. The HMM-based model uses conditional independence assumptions within HMM and between different modules. This does not match the actual situation of LVCSR.

Advantages of end-to-end ASR system:

1) Compared with the HMM-based model, the end-to-end model uses a single model to directly map audio to characters or words with no requirement of domain expertise, thus is simpler for constructing and training.

2) End-to-end ASR system can achieve the total optimization.

Disadvantages of end-to-end ASR system

1) End-to-end ASR system often suffers from the problem in which redundant generations repeat and importance symbols vanish.

2) Current performance of the end-to-end model is still worse than that of the HMM-DNN model, at best just comparable.

Looking ahead, the end-to-end ASR system needs to at least be improved in the following aspects:

1) Better trade-off the model delay and the recognition performance. Reducing latency while ensuring the recognition accuracy is an important but challenging research issue for the end-to-end model.

2) Better language knowledge learning. HMM-based model uses additional language models to provide a wealth of language knowledge, while the end-to-end model can only learn from limited training data’s transcriptions. This leads to great difficulties in dealing with scenes with large linguistic diversity.

1. Please explain the main problems and solutions of traditional signal processing based and deep learning based single channel speech enhancement approaches.

Both of the traditional signal processing based and deep learning based single channel speech enhancement approaches have a common problem: the estimation of the noise signal.

Traditional signal processing based single channel speech enhancement approaches

1) Spectral Subtraction

In Spectral Subtraction, noise signals are assumed to be additive, so the estimate of the underlying clean speech spectrum could be obtained by subtracting the estimate of noise spectrum from the noisy spectrum. The noise spectrum is estimated from the silent periods i.e., absence of the speech signals.

1. Wiener Filtering

Wiener filtering based speech enhancement minimizes the mean square error (MSE) between the estimated speech magnitude spectrum and the original signal magnitude spectrum.

The traditional speech enhancement methods mentioned above work well when the additional noise signal is stationary. However, the hypotheses for such algorithms do not well under the non-stationary noisy conditions. This is the time we need to utilize the non-linear modeling ability of deep learning based speech enhancement methods to get better enhanced results.

The deep learning based single channel speech enhancement approaches have another unique problem: its low generalization.

Most spectral domain based speech enhancement techniques exploit some higher-level feature. Recent advanced speech enhancement approaches mainly operate on the waveform of signal directly and further improved speech quality. However, generalization remains a major problem in deep learning based single channel speech enhancement. In [1], they propose *learnable loss mixup (LLM),* a simple and effort-less training diagram, to improve the generalization of deep learning-based speech enhancement models.

[1] Single-channel speech enhancement using learnable loss mixup

7. What is the difference between generative embedding with DNN i-vector and deep speaker embeddings? How to extract X-vector embeddings?

We use a low-dimensional “identity vector” (i-vector for short) to represent a speech segment. An i-vector contains the voice characteristic of a person (attributed to the speaker subspace) and channel factors (attributed to the channel subspace). The i-vector approach has become state-of-the-art in the speaker verification field. The approach provides an elegant way of reducing high-dimensional sequential input data to a low-dimensional fixed-length feature vector while retaining most of the relevant information. In general, we may extract the i-vector by the following steps:

1. Extract from waveform to get acoustic feature vectors
2. Use a universal background model (which is a GMM) to extract sufficient statistics
3. Obtain i-vector using a low-rank projection
4. Score with PLDA

The architecture for deep speaker embedding is an encode-decoder model. Encoder takes a sequence of acoustic features and derive intermediate representations. Temporal aggregation converts the sequence of intermediate representations into a single fixed-dimensional vector. In decoder, one of the layers is designed to be a bottleneck layer whose output (before the non-linearity) is taken as the speaker embedding. The encoder-decoder network is trained end-to-end to classified utterances from a large set of speakers.

The x-vector system is based on the DNN i-vector architecture and its the configuration is outlined in the table below. x-vectors are extracted at layer *segment6*, before the nonlinearity.

|  |  |  |  |
| --- | --- | --- | --- |
| Layer | Layer context | Total context | Input x output |
| frame1 | [t-2,t+2] | 5 | 120x512 |
| frame2 | {t-2,t,t+2} | 9 | 1536x512 |
| frame3 | {t-3,t,t+3} | 15 | 1536x512 |
| frame4 | {t} | 15 | 512x512 |
| frame5 | {t} | 15 | 512x1500 |
| stats pooling | [0,T) | *T* | 1500*T*x3000 |
| segment6 | {0} | *T* | 3000x512 |
| segment7 | {0} | *T* | 512x512 |
| softmax | {0} | *T* | 512x*N* |

Suppose an input segment has *T* frames. The first five layers operate on speech frames, with a small temporal context centered at the current frame *t*. The statistics pooling layer aggregates all *T* frame-level outputs from layer *frame5* and computes its mean and standard deviation. The statistics are 1500 dimensional vectors, computed once for each input segment. This process aggregates information across the time dimension so that subsequent layers operate on the entire segment. The mean and standard deviation are concatenated together and propagated through segment-level layers and finally the softmax output layer.

8. Read one or more the latest papers about a certain cutting-edge research topic related to acoustic processing; summarize and report the key challenges, novelty and contribution of the referred paper; give your own comments on the referred paper, not limited to its limitations, potential developments, future extensions, etc.

My recent research interest is sound source localization for unknown number of sources. Sound source localization (SSL) is the technology of estimating the Direction-of-Arrival (DOA) of one or several sound sources from the multichannel signals captured by the microphone array. Most methods have been focusing on localizing a single sound source, which do not extend to multiple sources rather than unknown number of sources. However, to estimate the DOA of unknown number of sources is the real key challenge, cause the number of sources is often unknown in real applications and SSL's performance decrease sharply in unknown number of speakers condition.

The first paper I read is DNN for Multiple Speaker Detection and Localization. They propose a likelihood-based encoding of the network output so that the spatial pseudo-spectra (SPS) can be generated by the network and then select form peaks of the SPS that are above a certain threshold when the number of sound sources is unknown. The idea of generating SPS by the network can work and it is no worse than that generated by conventional methods. It does achieve relatively good results of localization with a unknown number of sources, but the threshold is hard to choose and a pre-determined threshold for all domains are prone to errors due to the domain shift problem.

The second one is Robust Source Counting and DOA Estimation Using SPS and CNN. They propose to use a CNN to estimate the number of sources from short-time SPS. And the estimated number of sources is used to select the final DOAs without using a fixed threshold. The idea of firstly estimating number of sources is fancy, but the accuracy of the estimation is not high enough (around 70%), thus misleading to wrong DOA estimation. In future, I’d like to do more research on improving the accuracy of estimating the number of sources, proposing a more efficient network combined with the SPS estimating network mentioned above, thus can do more accurate DOA estimation for an unknown number of sources in an end-to-end fashion.

References:

The Application of Hidden Markov Models in Speech Recognition

Voice recognition algorithms using Mel-frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

Overview of end-to-end speech recognition

An Overview of End-to-End Automatic Speech Recognition

END-TO-END ASR: FROM SUPERVISED TO SEMI-SUPERVISED LEARNING WITH MODERN ARCHITECTURES

Speech Enhancement Based on Teacher–Student Deep Learning Using Improved Speech Presence Probability for Noise-Robust Speech Recognition

Single-channel speech enhancement using learnable loss mixup（可学习的损失混合）