ASR: Past Paper May 2016

Completed on May 1, 2023

Patrick Tourniaire

Problem 1

Subproblem 1. *Spectral domain filter-bank features are now widely used in neural network acoustic models in preference to mel frequency cepstral coefficients (MFCCs). Why is this? – and why are filter bank features rarely used for acoustic models based on Gaussian mixture models (GMMs)?*

Answer
Spectral domain filter-bank features are now widely used in neural network acoustic mdoels in preference to mel frequency cepstral coefficients (MFCCs) for several reasons.

- Improved discriminability: filter-bank features have been shown to have better discriminability for speech recognition tasks compared to MFCCs. This is because filter-bank features are designed to capture the spectral characteristics of speech in a more fine-grained way, while MFCCs are based on a logarithmic compression of the frequency spectrum.
- Better integration with NN: NN are well-suited for learning complex representations from high-dimensional inputs. Filter-bank features can provide a richer input representationfor neural networks than MFCCs, as they capture more spectral information.
- Simplicity: computing filter-bank featrues is computationally simpler than computing MFCCs, as the former does not require the computation of the discrete cosine transform (DCT).

In contrast, GMMs are computationally less efficient than NN and require a larger number of features for accurate modelling. This is because GMMs require the estimation of a large number of paramaters, which can be computationally expensive. While computing filter-bank features may be computationally simpler than computing MFCCs, it can still be a bottleneck in GMM-based systems due to the large number of features required for accurate modeling. Additionally, filter bank features have a high degree of correlation with each other due to their spectral overlapping nature. this redundancy can lead to overfitting in GMM-based systems, where the number og Gaussian components is limited due ti computational constraints.

Subproblem 2. *What are the main characteristics of a log mel filterbank used for feature extraction with a neural network acoustic model? What features of human hearing are related to the computation of these features?*

Answer
A log mel filter-bank is a set of filters applied to the magnitude spectrum of a speech signal in order ot extract acoustic features that are relevant for speech recognition. The main characteristics of a log mel filter-bank used for feature extract with a NN acoustic model are.

- Logarithmic scale: the filter-bank is constructed on a logarithmic scale of frequency, as this is more in line with human perception of sound *(frequency resolution)*. The logarithmic scale allows for a better representation of low-frequency components and a more uniform distribution of filter across the frequency spectrum.
- Mel scale *(pitch perception)*: the filters are space according to the mel scale, which is a non-linear transformation of frequency that is based on the preception of pitch in human hearing. The mel scale has been found to provide a better representation of speech signals than a linear frequency scale.
- Overlapping filters *(critical bands)*:the filters in the log mel filter-ank overlap in frequency, with adjecent filters overlapping by 50% or more. This allows for a better representation of the spectral characteristics of speech signals.
- Triangular shape: the filters in the log mel filterbank are typically triangular in shape. This is because the mel scale is not linear and so the width of the filters needs to vary according to frequency in order to maintain a uniform coverage of the spectrum.

Problem 1 continued on next page...

Subproblem 3. *Outline two ways in which neural networks can be used to generate features for an acoustic model. What advantages do such features bring to a speech recognition system.*

Answer
There are many way one can use NN architectures to generate features for an acoustic model. However, two traditional examples include CNNs and LSTMs.

- Concolutional neural networks (CNNs): can be trained to learn feature representations directly from the raw waveform of speech signals. These features are typically learned in a hierarical manner, where lower layers of the network learn low-level features such as frequency filters, and higher layers learn more abstract representations. The output of the final layer of the CNN can be used as input features to an acoustic model.
- Long short-term memory (LSTMs): LSTMs can be used directly with raw waveforms for temporal modelling, but higher level modelling of the features helps to disentangle underlying factors of variation within the input. Further, the LSTM cells are able to capture long-term dependencies in the input signal. LSTMs are also robust to variations in the acoustic environment, as they can learn features that are invariant to changes in noise level, , and other acoustic factors.

Subproblem 4. *In what circumstances might the fundamental frequency $F0$ be a useful feature for a speech recognition system?*

Definition 0.1: What is the fundemental frequency $F0$?

The fundamental frequency of a speech signal, often denoted by F0 or F0, refers to the approximate frequency of the (quasi-)periodic structure of voiced speech signals. The oscillation originates from the vocal folds, which oscillate in the airflow when appropriately tensed. The fundamental frequency is defined as the average number of oscillations per second and expressed in Hertz. Since the oscillation originates from an organic structure, it is not exactly periodic but contains significant fluctuations. In particular, amount of variation in period length and amplitude are known respectively as jitter and shimmer. Moreover, the F0 is typically not stationary, but changes constantly within a sentence. In fact, the F0 can be used for expressive purposes to signify, for example, emphasis and questions.
Typically fundamental frequencies lie roughly in the range 80 to 450 Hz, where males have lower voices than females and children. The F0 of an individual speaker depends primarily on the length of the vocal folds, which is in turn correlated with overall body size. Cultural and stylistic aspects of speech naturally have also a large impact.
The fundamental frequency is closely related to pitch, which is defined as our perception of fundamental frequency. That is, the F0 describes the actual physical phenomenon, whereas pitch describes how our ears and brains interpret the signal, in terms of periodicity. For example, a voice signal could have an F0 of 100 Hz. If we then apply a high-pass filter to remove all signal components below 450 Hz, then that would remove the actual fundamental frequency. The lowest remaining periodic component would be 500 Hz, which correspond to the fifth harmonic of the original F0. However, a human listener would then typically still perceive a pitch of 100 Hz, even if it does not exist anymore. The brain somehow reconstructs the fundamental from the upper harmonics. This well-known phenomenon is however still not completely understood.

Answer
The fundemental frequency (F0) is the frequency of the lowest harmonic of the voice signal, and it can carry important information about the speaker, the language, and the prosody of speech. In some circumstances, F0 can be a useful feature for a speech recognition system, such as:

- Speaker recognition: $F0$ is a speaker-dependent feature that can be used for speaker identification or verification. Since $F0$ varies grately between speakers, it can be used as a feature for speaker modelling in speaker recognition systems.
- Emotion recognition: $F0$ can be used as a feature for emotion recognition, as the pitch of the voice is closely related to the emotional state of the speaker. For example, high $F0$ values may be associated with excitement or anxiety, while low $F0$ values may be associated with sadness or depression.
- Tone languages: in tone languages, such as Mandarin or Cantonese, the pitch contour of speech is used to distinguish between words with different meanings. In these languages $F0$ can be a useful feature for speech recognition, as it provides important information about the lexical feature for speech recognition, as it provides important information about the lexical and syntactic structure of the language.
- Noisy environments: $F0$ can be useful in noisy environments, as it is less affected by noise than other spectral features. This is because $F0$ is a periodic feature that is less affected by noise than spectral features, which are based on the energy distribution of the signal.

Subproblem 5. *Recently there has been a lot of interest in using raw waveforms as direct input to neural network acoustic models. What potential advantages could motivate this approach? What are the drawbacks?*

Answer
Using raw waveforms as direct input to NN acoustic models has several potential advantages.

- Improved feature representation: by using raw waveforms, the model can capture more detailed and subtle aspects for the speech signal that may not be captured by traditional feature extraction techniques.
- Simplified processing pipeline: the use of raw waveforms eliminates the need for complex signal processing pipelines.
- Reduced training time: since the feature extraction step is eliminated, training times may be reduced as the network can learn directly from the raw data.

However, there are also some potential drawbacks to using raw waveforms:

- Increased computational complexity: raw waveforms are generally larger in size than traditional feature representations, which can lead to increased computational complexity during training and inference.
- Increased memory requirements: takes up more memory due to the size, can be a concern for low memory resources.
- Increased training data requirements: training a NN on raw waveforms requires significantly more data than training on traditional feature representations, which can be a challenge if sufficient data is not available.
- Limited interpretability: raw waveforms are inherently less interpretable than traditional feature representations, which can make it difficult to diagnose and correct errors in the system.

Problem 2

Subproblem 6. *Explain concisely how Viterbi training of a hidden Markov model (HMM) differs from training using the forward-backward algorithm. What are the advantages of forward-backward training?*

Answer
Viterbi training and forward-backward training are two methods for training HMMs. The main difference between the two is the way they estimate the model parameters. In Viterbi training, the MLE of the HMM parmaeters is obtained by aligning the training data with the model's most likely state sequence using the Viterbi algorithm. The Viterbi algorithm calculates the most likely sequence of hidde states that generated the observed data. The model parameters are then updated based in the state sequence alignment.
In contrast, forward-backward training uses the forward and backward algorithms to compute the posterior probabilities of the hidden states given the observations. The model parameters are then updated based on these parameters. The advantage of forward-backward training is that it takes into account all possible state sequences and can provide more accurate estimates of the model parameters. Viterbi training, on the other hand, only considers the most likely state sequence and may produce biased estimates if the data is noisy or the model is complex. However, Viterbi training is computationally more efficient then forward-backward training and may be preferred in some applications.

Subproblem 7. *How is Viterbi training used to train a hybrid neural network – HMM system? Could forward-bakward training be used in this case? How would the neural networks be trained?*

Answer
Viterbi training can be used ti train the HMM part of a hybrid HMM/NN system. In this approach, the NN is used to estimate the emission probabilities for each state in the HMM. The Viterbi algorithm is then used to find the most likely state sequence given the observed data and the HMM parameters are updated based on this alignment. The NN is then retrained using the updated HMM parameters, and the process is repeated until convergence.
It is possible to use the forward-backward in this context, but it requires additional computation and is less commonly used. In forward-backward training, the NN is trained to predict the posterior probabilities of the HMM states given the input features, and these probabilities are used ti update the HMM model parameters. The NN is then fine-tuned using backpropagation to further improve its predictions.

Subproblem 8. *A phonetic decision tree is used to obtain state-clustered context-dependent models. How is the log likelihood used to determine state splitting? Assuming each state has a single Gaussian pdf explain how the log likelihood of a state can be computed efficiently.*

Answer
The log likelihood is used to determine the splitting of states in a phonetic decision tree by computing the log likelihood ratio between the original state and the two candidate split states. If the log likelihood ratio exceeds a predefined threshold, the state is split into two child states, and the process is recursively applied to each child state. Assuminb each state has a single Gaussian PDF, the likelihood of a state can be computed efficiently using the following states.

1. Compute the mean vector and the covariance matrix of the Gaussian PDF for the state using the training data.

2. For each training sequence, compute the log likelihood of the sequence under the Gaussian PDF for the state using the formula.

$$logP(\boldsymbol{O}|\mu,\Sigma) = -\frac{1}{2}log|\Sigma| - \frac{D}{2}log(2\pi) - \frac{1}{2\pi}(\boldsymbol{O}-\mu)^\top\Sigma^{-1}(\boldsymbol{O}-\mu) \tag{1}$$

   Where $\boldsymbol{O}$ is the observation sequence, μ is the mean vector, Σ is the covariance matrix, and D is the dimension of the observation vector.

3. Sum the log likelihoods of all training sequences to obtain the total log likelihood of the state.

Since each state has a single Gaussian PDF, the covariance matrix is diagonal, and its determinant can be computed efficiently. The matrix inverse in the log likelihood formula can also be computed efficiently using the diagonal elements of the inverse covariance matrix. The results in a computationally efficient method for computing the log likelihood of a state.

Subproblem 9. *Explain how scaled likelihoods are obtained from the neural network in a hybrid system and why the outputs of the neural network are not used directly. If the system was to be used for TIMIT phone recognition (rather than large vocabulary speech recognition) would it be necessary to compute the scaled likelihoods? Justify your answer.*

Answer