

ASR: Past Paper May 2014

Completed on April 26, 2023

Patrick Tourniaire

Problem 1

Subproblem 1. Use a block diagram to show the main steps involved in computing PLP coefficients from a sampled speech waveform.

Answer

The computation of the PLP coefficients can be summarised by the following block diagram visualised in Figure .

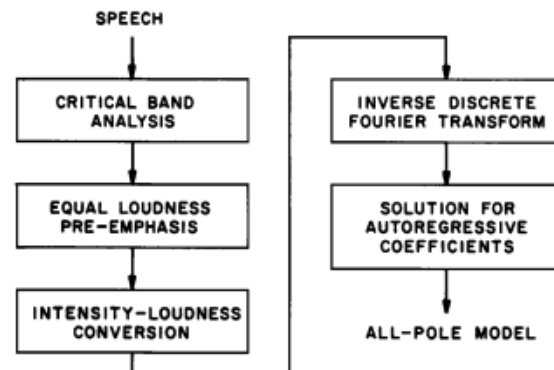


FIG. 1. Block diagram of perceptual linear predictive (PLP) speech analysis.

Subproblem 2. How do PLP coefficients differ from Mel frequency cepstral coefficients (MFCCs)?

Answer

Both PLP coefficients and MFCCs are commonly used feature representations for speech processing tasks such as speech recognition, speaker identification, and emotion detection. PLP and MFCCs are similar in that they both use a Mel filterbank to group spectral components of the speech signal into perceptually meaningful frequency bands. However, there are several key differences between the two feature representations.

1. Non-linear mapping: PLP coefficients use a non-linear mapping of the filterbank outputs that is designed to better model the human auditory system. In contrast, MFCCs use a linear transformation of the filterbank outputs.
2. Cepstral smoothing: PLP coefficients use cepstral smoothing to reduce the effect of spectral noise and emphasise the formant structure of the speech signal. In contrast MFCCs do not use cepstral smoothing.
3. Spectral warping: PLP coefficients use a linear prediction model to estimate the spectral envelope of the speech signal, which can be used to further reduce the effects of spectral noise and improve the representation of the formant structure. In contrast, MFCCs do not explicitly model the spectral envelope.

In conclusion, PLP coefficients are designed to provide a more robust and accurate representation of the speech signal than MFCCs, particularly in noisy or reverberant environments. However, the choice of feature representation ultimately depends on the specific application and the characteristics of the speech data being processed.

Definition 0.1: What is a spectral envelope?

The spectral envelope of a speech signal refers to the shape of the overall spectral energy distribution across the frequency range of the speech signal. It represents the underlying pattern of resonances in the vocal tract, which are responsible for shaping the sounds of speech.

The spectral envelope is called an "envelope" because it encloses the spectral shape of the speech signal, much like an envelope encloses a letter. It is a smooth function that describes the overall shape of the spectrum of the speech signal, without providing detailed information about individual spectral components.

The spectral envelope can be estimated using various techniques, such as linear predictive coding (LPC), which models the speech signal as a linear combination of past samples, or the homomorphic filtering approach, which uses a logarithmic transformation of the speech signal to separate the spectral envelope from the fine spectral structure.

Estimating the spectral envelope is important for many speech processing tasks, such as speech enhancement, speaker identification, and speech synthesis, because it provides a compact and informative representation of the underlying vocal tract characteristics that are responsible for the sound of speech. By modeling the spectral envelope, it is possible to separate the information that is most relevant to the perceptual quality of the speech signal from irrelevant or noisy spectral components.

Subproblem 3. *Which parts of the PLP computation are inspired by properties of human perception?*

Answer

The PLP feature extraction method is designed to mimic the properties of human auditory perception, particularly with respect to the non-linearities and frequency selectivity of the auditory system. Several parts of the PLP computation are directly inspired by properties of human perception, including.

1. Mel filtering bank: the Mel filtering bank used in PLP is inspired by the frequency selectivity of the human cochlea, which acts as a bank of bandpass filters that separate the incoming sound wave into narrow frequency bands. The mel scale is used to approximate the non-linear frequency spacing of the cochlea, which are more closely spaced at lower frequencies than at higher frequencies.
2. Non-linear mapping: the non-linear mapping used in PLP is designed to mimic the compressive non-linearity of the inner hair cells in the cochlea, which act to increase the dynamic range of high level sounds.
3. Cepstral smoothing: the cepstral smoothing used in PLP is designed to model the effect of the integration time constant of the auditory system, which causes rapid fluctuations in the spectral envelope to be smoothed out over time.
4. Linear prediction: the linear prediction model used in PLP to estimate the spectral envelope is based on the idea that the vocal tract can be modeled as a linear filter with time-varying coefficients, which can be estimated from the speech signal.

By incorporating these perceptually inspired processing steps, PLP is able to provide a more accurate and robust representation of the human auditory system.

Definition 0.2: What is a Mel frequency bank?

The Mel frequency bank is a collection of bandpass filters that are used to group the spectral components of a speech signal into perceptually meaningful frequency bands. The Mel frequency bank is based on the Mel scale, which is a non-linear scale of frequency that approximates the frequency spacing of the human auditory system.

The Mel scale is designed to mimic the frequency selectivity of the human cochlea, which is the organ in the inner ear responsible for transducing sound into neural signals. The cochlea contains a bank of bandpass filters that separate the incoming sound wave into narrow frequency bands, with the bandwidth of each filter increasing with frequency.

The Mel scale is based on the observation that the frequency spacing of these cochlear filters is non-uniform, with smaller frequency intervals at low frequencies and larger intervals at high frequencies. The Mel scale is defined such that the distance between two frequencies on the Mel scale corresponds to the difference in pitch that is perceived by the human ear.

To construct the Mel frequency bank, a set of triangular bandpass filters are placed evenly spaced on the Mel scale, with the center frequencies of each filter corresponding to the peak of each triangle. The bandwidth of each filter is determined by the width of the triangle, which is typically chosen such that adjacent filters overlap slightly.

The output of the Mel frequency bank is a set of filterbank coefficients, which represent the energy in each frequency band. These coefficients can be used as features in speech processing tasks such as speech recognition or speaker identification. The Mel frequency bank is a common pre-processing step for feature extraction methods such as Mel Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Prediction (PLP).

Subproblem 4. *Describe two ways in which information from multiple frames of the speech signal may be incorporated in the acoustic features used for an HMM / GMM (hidden Markov model / Gaussian mixture model) acoustic model, in order for the features to encode information about the speech dynamics*

Answer

in order to capture the temporal dynamics of speech, it is often necessary to incorporate information from multiple frames of the speech signal into the acoustic features used for an HMM/GMM acoustic model. Here are two ways in which this can be done.

1. **Dynamic features:** dynamic features are a class of acoustic features that incorporate information from multiple adjacent frames of the speech signal. One common type of dynamic feature is the delta feature, which represents the rate of change of the acoustic features over time. Delta features are computed by taking the first-order difference of the original acoustic features between adjacent frames. Another type of dynamic feature is the delta-delta feature, which represents the acceleration of the acoustic features over time. Delta-delta features are computed by taking the second-order difference of the original acoustic features between adjacent frames. By incorporating information from multiple frames of the speech signal in this way, dynamic features can capture the temporal dynamics of speech, such as changes in pitch, **formant transitions**, and coarticulation effects.
2. **Contextual features:** are a class of acoustic features that incorporate information from neighboring frames of the speech signal, both preceding and following the current frame. Contextual features can be computed using a sliding window approach, where the acoustic features for each frame are concatenated with the features from neighboring frames within a certain window size. By incorporating contextual information from multiple frames of the speech signal in this way, contextual features can capture the contextual dependencies between adjacent speech sounds, such as phoneme-to-phoneme transitions or word-to-word transitions. One common type of contextual feature

is the delta delayed feature, which incorporates information from both preceding and following frames of the speech signal to capture the coarticulation effects between adjacent speech sounds.

Definition 0.3: What are formant transitions?

Formant transitions are changes in the resonant frequencies of the vocal tract that occur during the production of speech sounds. Formants are the peaks in the frequency spectrum of a speech signal that correspond to the resonant frequencies of the vocal tract. The formants are determined by the shape and size of the vocal tract, which can be modified by the position and movement of the tongue, lips, and other articulators.

When producing a speech sound, the vocal tract undergoes a series of rapid changes in shape as the articulators move to produce different phonemes. These changes cause the formant frequencies to shift, creating formant transitions between adjacent speech sounds. Formant transitions are particularly important for distinguishing between similar-sounding phonemes, such as /b/ and /p/, which differ primarily in the timing and direction of their formant transitions.

Formant transitions can be captured in the acoustic features used for speech processing tasks such as speech recognition or speaker identification. Dynamic features such as delta and delta-delta features can be used to capture the rate of change of the formant frequencies over time, while contextual features such as pitch-synchronous features can be used to capture the pitch-dependent changes in the formant frequencies. Modeling formant transitions is an important aspect of acoustic modeling for speech processing, as it can help improve the accuracy and robustness of speech recognition and other speech processing tasks.

Definition 0.4: What are coarticulation effects?

Coarticulation effects are changes in the acoustic properties of speech sounds that are caused by the influence of neighboring sounds in a sequence of speech. When we speak, we don't produce each sound in isolation; instead, the production of one sound can influence the production of the sounds that come before and after it. Coarticulation effects can include changes in the timing, duration, amplitude, and spectral characteristics of speech sounds.

For example, when producing the word "bat," the articulation of the consonant /b/ requires the lips to be pressed together, which affects the shape and size of the vocal tract. As a result, the formant frequencies of the following vowel /a/ are shifted slightly compared to the formants of the same vowel produced in isolation. Similarly, when producing the word "cab," the articulation of the consonant /k/ requires the tongue to be pressed against the velum, which affects the shape and size of the vocal tract and causes a shift in the formant frequencies of the preceding vowel /æ/.

Coarticulation effects can make speech recognition and other speech processing tasks more challenging, as the acoustic properties of a given speech sound can vary depending on the sounds that come before and after it. However, coarticulation effects can also provide useful information for understanding the dynamics of speech production and for developing more accurate and robust models of speech processing. Dynamic and contextual features, as well as other acoustic modeling techniques, can be used to capture coarticulation effects and improve the accuracy of speech processing algorithms.

Subproblem 5. *In hybrid HMM / DNN (hidden Markov model / deep neural network) systems, log mel filterbank features have been found to be superior to PLP coefficients or MFCCs (mel frequency cepstral coefficients). Why might this be the case, and why is it not the case for HMM / GMM systems?*

Answer

Log mel filterbank features have been found to be superior to PLP coefficients or MFCCs in hybrid HMM/DNN systems for several reasons.

1. Better representation of spectral information: log mel filterbanks better represent the spectral information in speech signals compared to other features like MFCCs or PLPs. This is because the human auditory system is more sensitive to changes in frequency at lower frequencies than at higher frequencies, and log mel filterbanks capture this property by mapping the spectral frequencies into mel-scale, which is logarithmically spaced at lower frequencies and linearly spread at higher frequencies.
2. More robust to noise: log mel filterbanks are more robust to additive noise than MFCCs and PLPs. This is because they are less sensitive to changes in the amplitude of the signal and more sensitive to changes in the spectral shape.
3. Better regularisation: the use of log mel filterbanks as input to DNNs provides better regularisation than PLPs or MFCCs. This is because the number of parameters in a DNN is much larger than in an HMM, and the use of log mel filterbanks helps to reduce overfitting by providing a more compact representation of the spectral information.

However, the superiority of log mel filterbanks over other features is not always the case for HMM/GMM systems. This is because HMM/GMM systems are based in Gaussian mixture models (GMMs) that assume the statistical distribution of the features is Gaussian. In this case, the use of PLPs or MFCCs may provide a better fit to the Gaussian distribution assumption of the GMM. In addition, HMM/GMM systems are typically trained using maximum likelihood estimation (MLE), which is less sensitive to the choice of input features than DNN-based systems.

Problem 2

Subproblem 6. Consider a hidden Markov model (HMM) which has a state output distribution given by a single Gaussian. We would like to estimate the mean μ_j of state s_j , given a training sequence of acoustic observations $X = x_1, \dots, x_t, \dots, x_T$.

If the time-state alignment is known, write an expression for the maximum likelihood estimate of μ_j .

Answer

The expression for the estimated μ_j , given that we have a single Gaussian optimised using MLE is given by.

$$\mu_j = \frac{1}{T} \sum_{t=1}^T x_t \quad (1)$$

Subproblem 7. If the time-state alignment is not known, how is a maximum likelihood estimate of μ_j obtained? Your answer should include a definition of $\gamma_t(s_j)$, the state occupation probability of state s_j at time t , and a description of how $\gamma_t(s_j)$ is computed using the forward-backward algorithm.

Answer

The expression for the estimated μ_j when the time-state alignment is not known, is given by the following.

$$\hat{\mu}_j = \frac{\sum_{t=1}^T \gamma_t(s_j) x_t}{\sum_{t=1}^T \gamma_t(s_j)} \quad (2)$$

The state occupation probability $\gamma_t(s_j)$ is the probability of occupying state s_j at time t given the sequence of observations. Which can be expressed in terms of the forward and backward probabilities.

$$\gamma_t(s_j) = P(S(t) = s_j | X, \lambda) \quad (3)$$

$$= \frac{1}{\alpha_T(s_E)} \alpha_t(s_j) \beta_t(s_j) \quad (4)$$

$$\alpha_t(j) = p(\mathbf{x}_1, \dots, \mathbf{x}_t, S(t)=j | \lambda) = \sum_{i=1}^N \alpha_{t-1}(i) a_{ij} b_j(\mathbf{x}_t)$$

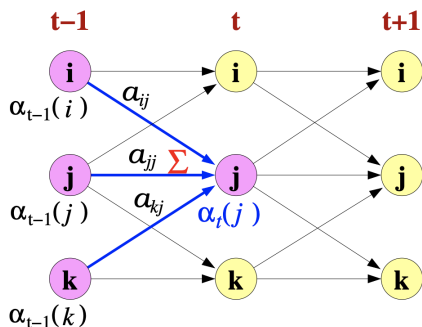


Figure 1: Forward probability computation

$$\beta_t(j) = p(\mathbf{x}_{t+1}, \dots, \mathbf{x}_T | S(t)=j, \lambda) = \sum_{j=1}^N a_{ij} b_j(\mathbf{x}_{t+1}) \beta_{t+1}(j)$$

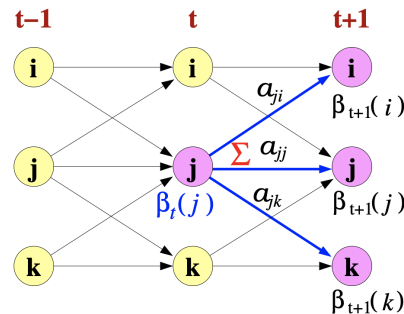


Figure 2: Backward probability computation

The forward and backward probabilities are given by the equations in Figure 1 and 2, as well as an explanatory visualisation for how these likelihoods are indeed computed.

Subproblem 8. *In a large vocabulary speech recognition system for English, the fundamental units of speech are typically context-dependent phone models. Why is some degree of parameter sharing necessary when constructing context-dependent phone models?*

Answer

Parameter sharing is necessary when constructing context-dependent phone models in a large vocabulary speech recognition system for English because there are a large number of possible contextual variations that a phone can have, and it is not feasible to collect sufficient data to model all of them independently. Parameter sharing helps solve this by having the same set of parameters which are shared for multiple contexts, which also helps prevent overfitting.

Subproblem 9. *Which approach to parameter sharing is predominantly used in current context-dependent acoustic models, and why is it preferred to alternative approaches?*

Answer

The approach to parameter sharing that is predominantly used in current context-dependent acoustic models is known as "tied-state" or "shared-state" modelling. In this approach, multiple context-dependent states are tied together to share the same set of acoustic parameters, such as Gaussian mixture means and variances. This approach is preferred to alternative approaches for several reasons.

- Parameter sharing reduces the number of parameters that need to be estimated, which leads to more efficient and faster training of the model.
- Tied-state modelling helps to regularise the model and prevent overfitting, as the sharing of parameters encourages the model to generalise better to new unseen contexts.
- Tied-state modelling also reduces the risk of over-segmentation, where the model segments the speech signal into too many short segments, which can lead to poor recognition performance.
- Tied-state modelling has been shown to produce better recognition accuracy than other approaches, such as state clustering or decision tree clustering.

Overall, tied-state modelling is a powerful and widely used technique for parameter sharing in context-dependent acoustic models, as it provides a good balance between modelling complexity and recognition accuracy.

Subproblem 10. *How is a phonetic decision tree used to enable parameter sharing? (Your answer does not need to include details of the splitting algorithm.)*

Answer

A phonetic decision tree is used to enable parameter sharing in context-dependent acoustic models by grouping similar context-dependent states together, based on their phonetic properties. An example of such a tree can be seen in Figure 3.

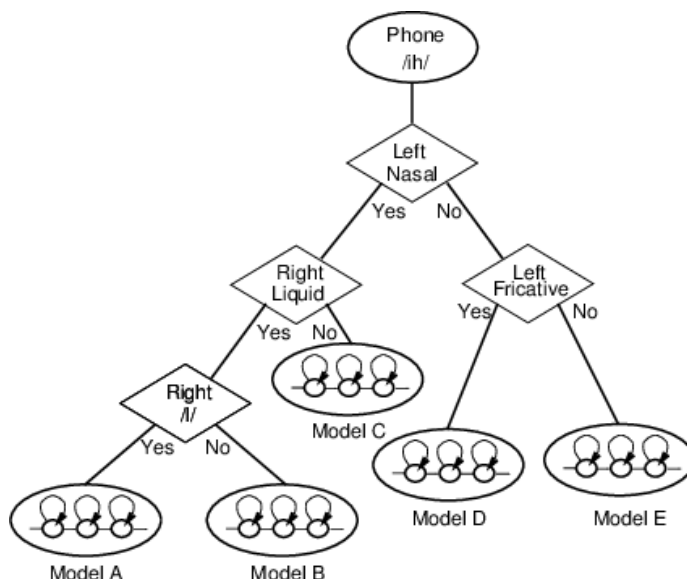


Figure 3: Example of a phonetic decision tree

The tree is constructed by recursively splitting the states into two groups based on a phonetic feature, such as the identity of the preceding phone, the position of the current phone within a word, or the identity of the following phone. The goal is to group states that have similar acoustic properties and share the same acoustic parameters, and to separate states that have different acoustic properties and require different properties. Once the decision tree is constructed, each leaf node represents a context-dependent state, and the path from the root to the leaf node defines the set of phonetic features that characterize that state. The nodes along the path to the leaf node determine the context in which the state occurs, and the acoustic parameters associated with the leaf node are used to model the acoustic properties of the state.

The decision tree is used to enable parameter sharing by tying together context-dependent states that share the same path from the root to a common ancestor node. These tied states share the same acoustic parameters, which reduces the number of parameters that need to be estimated and improves the model's ability to generalize to new contexts.

In summary, a phonetic decision tree is used to group similar context-dependent states together and enable parameter sharing by tying together states that share the same phonetic context and acoustic properties. This approach helps to reduce the number of parameters that need to be estimated and improves the model's ability to generalize to new contexts.

Problem 3

Subproblem 11. *Explain how a neural network trained as a classifier may be used to provide the output probabilities in a hybrid neural network / HMM system.*

Answer

In a hybrid NN/HMM speech recognition the NN is typically used to provide the output probabilities for each frame of acoustic features, and the HMM is used to model the temporal dynamics of speech. The NN is trained as a classifier to predict the posterior probabilities of the phone labels given the input acoustic features. During training, the network is trained to minimize the cross-entropy loss between the predicted posterior and the true phone labels.

At runtime, the NN is used to compute the posterior probabilities of the phone labels for each frame of acoustic features. These probabilities are then used as emission probabilities in the HMM, which models the temporal dynamics of speech. The HMM computes the joint probability of the sequence of phone labels and the sequence of observed acoustic features, using the forward-backward algorithm. This involves computing the likelihood of each frame of acoustic features given each possible phone label, using the emission probabilities provided by the NN. The HMM also models the transitions between phone labels using transition probabilities, which are estimated from the training data.

Subproblem 12. *Discuss two approaches to speaker adaptation which may be used in a hybrid neural network / HMM system.*

Answer

Here are two common approaches using Linear Input Networks (LIN) and i-vectors.

- **Linear Input Network (LIN):** LIN is a method for speaker adaptation that involves transforming the input features of the speech signal to better match the speaker's voice characteristics. During training, a set of speaker-independent models is trained using a large corpus of speech data from multiple speakers. Then, during adaptation, a linear transformation matrix is learned from a small amount of speaker-specific data, which is used to transform the acoustic features to better match the speaker's voice characteristics. The transformed features are then used as input to the neural network, which is trained using a technique such as backpropagation. LIN is computationally efficient and can be performed quickly in real-time applications.
- **i-vectors:** i-vectors are a method for speaker adaptation that models the distribution of speaker-dependent information in the speech signal using a low-dimensional vector. During training, a set of speaker-independent models is trained using a large corpus of speech data from multiple speakers. Then, during adaptation, an i-vector is extracted from a small amount of speaker-specific data, which captures the speaker-dependent information in the speech signal. The i-vector is then used to adapt the acoustic models to better match the speaker's voice characteristics. The adaptation is typically performed using a technique such as maximum likelihood linear regression (MLLR) or maximum a posteriori (MAP) adaptation. i-vectors are a powerful approach for speaker adaptation, as they capture speaker-dependent information in a low-dimensional space and can be used to adapt a wide range of acoustic models.

Subproblem 13. *Compare and contrast the following three approaches to language modelling:*

- *Smoothed 4-gram model*
- *Bengio's neural probabilistic language model with inputs corresponding to the three previous words*
- *Mikolov's recurrent neural network (RNN) language model*

Your answer should refer to: the structure of the models and how they represent words; how they are trained; and how they may be used for speech recognition.

Answer

Smoothed 4-gram model, Bengio's neural probabilistic model, and Mikolov's recurrent neural network (RNN) language model are all approached to language modelling, but they differ in their structures, training methods and how they are used for speech recognition.

Smoothed 4-Gram Model

Structure: a smoothed 4-gram model is a statistical language model that calculates the probability of a word given the three previous words in the sentence. The model represents words as discrete symbols and stores their frequencies in a large n-gram table.

Training: the model is trained on a large corpus of text, where the frequencies of word n-grams are counted and smoothed using techniques like Laplace or Kneser-Ney smoothing. The model is then used to predict the probability of a word given its preceding three words.

Speech recognition: the smoothed 4-gram model is commonly used in speech recognition systems as a language model to constrain the search space and improve recognition accuracy. The model is typically combined with an acoustic model, such as a HMM, to perform speech recognition.

Bengio's Neural Probabilistic Language Model

Structure: Bengio's neural probabilistic language model is a feedforward neural network that takes as input the embeddings of the three previous words and predicts the probability distribution over the next word in the sentence. The model represents words as continuous-valued vectors (embeddings) that are learned during training.

Training: the model is trained on a large corpus of text using backpropagation and stochastic gradient descent. During training, the model learns to map the embeddings of the three previous words to a probability distribution over the vocabulary.

Speech recognition: Bengio's neural probabilistic language model is a feedforward neural network can be used in speech recognition systems in a similar way as the smoothed 4-gram model, as a language model to constrain the search space and improve recognition accuracy.

Mikolov's RNN Language Model

Structure: Mikolov's RNN language model is a type of NN that uses recurrent connections to model the sequence of words in a sentence. The model takes as input the embeddings of the current word and the hidden state of the previous time-step, and predicts the probability distribution over the next word. The model represents words as continuous-valued vectors (embeddings) that are learned during training.

Training: the model is trained on a large corpus of text using back-propagation through time and stochastic gradient descent. During training, the model learns to map the embeddings of the current word and the previous hidden state to a probability distribution over the vocabulary.

Speech recognition: Mikolov's RNN language model can also be used in speech recognition systems as a language model to constrain the search space and improve recognition accuracy. However, RNN language models may be computationally more expensive than simpler models like 4-gram models or Bengio's neural probabilistic language model.