Speech Classification Notes

What are linguistic and paralinguistic traits?

Linguistic traits are the words and phrases that are used in speech. Paralinguistic traits are the non-verbal cues that are used in speech. For example:

- **Linguistic**: "I am happy"
- **Paralinguistic**: The tone of voice used to say "I am happy"
  Gender, Age, Language accent, Emotion, and Personality are all examples of paralinguistic traits.

What is the objective of speech classification?

Convert a speech input into a sequence of class labels. For example, given a speech input, classify it as "happy", "sad", "angry", etc.

Why speech classification is a dreamy solution?

1. Because the samples of the same class can ahve different forms due to properties like volume,s peaker, gender, accent, etc. This makes it difficult to classify speech.
2. Other problems are the data collection and annotation: It's hard to get data due to privacy problems and biases; even more annotation is a very slow process.
3. Machine Learning problems: the input data usually is very different in length and the output is a sequence of labels. The variable length have problems, like Segmentation/Alignment problem. About the output, the high number of classes lead to too many combinations.

Solution: SPlit into sub problems.

What is the problem Sequence2One?

Given a sequence of input data, the objective is to classify it into a single class. For example, given a sequence of speech, classify it as "happy, sad or angry".

How to deal with the input problem at different levels?

- Audio level: fix the length truncating the audio
- Feature level: Use functional features, like *mean*
- Model level: Use a model that can deal with variable length inputs, like RNNs
- Output level: Fuse output predictions (probabilities)

What are the main properties of the features used in speech classification?

- Informative: The feature should give information about the class. Similar sounds should have different features.
- Practical: The feature should be easy to compute and store.
- Robust: Should not change and be able to deal with noise.

How can be the features extracted from speech?

- Local: from the frame
- Global: mean, median, etc
- Segmental: phoneme, voiced/unvoiced, etc

Which types of features can be extracted from speech?

- Spectral: FFT, MFCC, etc
- Prosodic: pitch, energy, etc
- Voice quality: jitter, shimmer, etc
- Other: time-domain, model-based, etc

Usually the featurs are extracted from overlapping windows of 15-30ms and are inspired by the human auditory system. **LPC (Linear Predictive Coefficients)** and **Cepstral** coefficients are the most used.

What was the schema for MFCC (Mel Frequency Cepstral Coefficients)?

1. Hamming window
2. Fast Fourier Transform
3. Mel warp and log
4. Discrete Cosine Transform
5. Mean removal
6. Delta and Delta-Delta

What are delta and double-delta features?

The idea is to calculate the feature vectore at two different times. The delta is the difference between the two vectors and the double-delta is the difference between the two deltas. This si done to capture the dynamics of the speech.

For example: If the speech is "I am happy", the delta and double-delta will capture the change in the speech, like the change in the tone of voice.

What does Cepstral mean normalization?

Since the cepstral have important information we need to normalize this to make the features more robust. Moreover, convolutional effect in the domain are linear in cepstral domain, so this makes easier to tract the problem.
It's done comuting the mean feature vector using sliding window or complete segment.

Prosoic features

Prosodic features are related to the pitch, energy, and duration of the speech. They are used to capture the emotion of the speaker.

- **Fundamental frequency (F0)**: The lowest frequency of a periodic waveform. It's used to capture the pitch of the speaker. We use mean, median, standard deviation, etc.
- **Energy**: The energy of the speech. It's used to capture the intensity of the speech. We use mean, median, standard deviation, etc.
- **Duration** : The duration of the speech. We use speech ratem ratio of voiced/unvoiced, etc.
- **Formants**: The resonant frequencies of the vocal tract. We use first to fourth formants.

Voice quality features

These features are related to some pathologic issues of the speaker. They are used to capture the health of the speaker.

- **Jitter and Shimmer**: Measurement of F0 disturbace: Jitter frequency variation and Shimmer amplitude variation.
- **HNR** (Harmonic-to-Noise Ratio): Ratio of the energy of the harmonic components to the energy of the noise components.

What are other features?

- **Time-domain features**: Zero-crossing rate, Autocorrelation, Attack time, etc.
- **Model-based features**: Bottle neck features, posterior based features and embedding features
- **High level features**: Usually dependant on tet, phonetic, lexical and discourse marker.

What are the main feature pre processing steps?

- Optional: speech enhancement like wiener filter or spectral subtraction
- Mandatory: VAD (Voice Activity Detection) to remove silence. It's done because silence can corrupt the training and degrade the test. It's done using energy threshold.
- Others: Echo cancellation, bandwidth expansion, etc.

Common feature manipulation steps

1. Feature selection: make the feature space smaller. It's done to speed up training and reduce overfitting. It's done using PCA, LDA, etc.

1. PCA (Principal Component Analysis): not needed for labelled data
2. LDA (Linear Discriminant Analysis): needed for labelled data

2. Feature augmentation: Add more training data.

- Adding noise
- Adding reverberation
- Increase or decrease speech rate
- Random volume gain

Specaugment: It's a new technique that consist of time masking, frequency masking, and time warping. It's used to make the model more robust to noise.

Machine learning start

What are the 2 main elements in ML methods

- Type of *discriminant function* meaning the model
- Type of *loss function* meaing the training objective

Difference betweeen discriminative and generative models

- Generative models: are models that learn the joint probability distribution of the input and output data. They are used to generate new data.
- Discriminative models: are models that learn the conditional probability distribution of the output data given the input data. They are used to classify data.

What is a GMM (Gaussian Mixture Model)?

It's a model that assumes that the data is generated by a mixture of several Gaussian distributions. It's used to model the distribution of the data. It's in **1D** meaning that the data is 1-dimensional. Parameters are the mean μ, the variance σ.

Since a 1 dimension model cannot model the distribution with multiple modes and non linear correlation a weighted sum is used.

To use a GMM it's needed to:

1. estimate the parameters

2. compute the log likelihood of a sequence

Fast SVM explanation

SVM is a model that tries to find the hyperplane that separates the data into two classes. It's used to classify data. It's a discriminative model. Uses kernels that are used to transform the data into a higher dimension. The kernel trick is used to avoid the computation of the transformation.

Differences between supervised and end-to-end learning

The traditional supervised learning is based on learning from the pairs X,Y using handcrafted features and initializing randomly the classifiers.

End2End learning is based on learning from the pairs X,Y, but it jointly training a feature extractor. This requires a lot of labelled data.

Steps for E2E learning

1. Lean a feature extractor using suelf supervised learning. Define a pretext task and pseudo labels derived from data. Extractor and pseudo classifier are trained here. Some example are BERT, CPC, VAE.
2. Remove pseudo classifier, add random initialized classifier and then fine tune:
    i. Classifier only
    ii. Classifier + feature extractor (it's more heavy)

Some examples of models like this are:

- wav2vec2
- HuBERT

How are wav2vec2 and HuBERT trained?

They are trained using the following steps:

`Wav2vec2`：

1. Audio
2. Latent feature encoder (CNN)
3. Mask 50% of the latent features
4. Context network (transformer)
5. Contrastive loss

`HuBERT`：

1. Audio
2. Clustering feature extractor
3. K-means clustering
4. cross entropy loss
5. context network
6. latent feature encoder

What is S3PRL?

A tool for feature extraction.

Speaker recognition

What is the standard pipeline for speaker recognition?

1. Speech signal
2. Feature extraction
3. Representation
4. Variability compensation
5. Backend classification
6. Results

Difference between verification and identification

1. Verification: easy. check if the speaker is the same. It's a binary classification.
2. Identification: hard. check who is the speaker. It's a multi-class classification.

Enrollment and test

1. Enrollment is the process of registering the speaker in the system.
2. Test is the process of checking the speaker in the system.

Definition of trials

A trial is a task that consist of making the system decide if the speaker is speaking inside the audio track or not.

- Target trial: the speaker is speaking
- Impostor trial: the speaker is not speaking

The system can give result as:

- decision: true or false
- score: the probability of the speaker speaking

What are the speaker recognition measures for evaluation?

- Missed detections: percentage of the trials rejected incorrectly. They can be the False Negative.
- False alarms: percentage of the trials accepted incorrectly. They can be the False Positive.

What is the equal error rate (EER)?

ERR is the point where the false positive rate is equal to the false negative rate. It's used to evaluate the performance of the system.

What is the DET curve?

The DET curve is a curve that shows the relationship between the false positive rate and the false negative rate. It's used to evaluate the performance of the system.

How is the Representation task handled?

Speaker models are used to represent the speaker information. A lot of models have been used.

What are impostor models?

It's a model trained using the Universal Background Model UBM. Takes information about all the speakers and then is adapted to each of them using the MAP (Maximum A Posteriori) adaptation.

How to ahndle the Variability compensation task?

This refers to the change of the speaker's voice that can be due to the environment, the channel, the microphone, etc. Initially was done using the *supervector* concept

A sueprvector is a kind of feature extraction result for discriminative models. It's a concatenation of the mean and the covariance of the GMM.

The real MVP here is the i-vector. It's a low-dimensional representation of the supervector. It's used a lot for language dialect and native language identification.

The last model is the x-vector. It's a deep neural network.