Feature Extraction for ASR

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

The goal of feature extraction is to find a representation for speech sound. Ideally, speech features are

- **discriminative**: The representations for different linguistic targets are distinct.
- **robust**: The representations are not easily and severely corrupted by environmental noise.
- **parsimonious**: Achieve the same performance with the number of features as small as possible.

Common Feature Vectors

Currently, most ASR systems uses a cepstral vector derived from a filter bank or linear prediction. The basic steps include

- estimate power spectrum of an analysis window
- integrate power spectrum over filter banks
- spectral adjustment by equal-loudness (or pre-emphasis)
- compress spectrum by cubic root or logarithm
- apply inverse DFT to get cepstrum
- further processing such as spectral smoothing, orthogonalization and liftering.

MFCC, LPC, and PLP

- The previous processings are used in MFCC (mel-frequency cepstral coefficients) and PLP (perceptual linear prediction).
- Both provide a representation of smoothed power spectrum that has been compressed.
- It is also interesting to compare LPC and PLP, as shown in Figure 22.4. PLP combines modules in LP and MFCC: it can be viewed as MFCC with LPC-like spectral smoothing, or as LPC analysis with MFCC-like auditory filters.

Dynamic Features

- MFCC or PLP represent smoothed estimates of local spectrum. They are called static features.
- However, it can be argued that a key characteristic of speech is its dynamic behavior. So it is desired to have local time derivative estimates in addition to the static features.
- The delta (a.k.a. velocity, from physics) features are the difference between current static feature and neighboring features.

Robustness

- Robustness for convolutional noise
  - The convolution of speech and noise becomes multiplication in the spectral domain. They are additive in the log spectral domain.
  - The cepstral mean subtraction (CMS) is quite effective in this case.
- Robustness for additive noise
  - The spectra of speech and noise are additive.
  - In this case, one can estimate the noise spectrum form non-speech segments and subtract it from speech segments. This is called spectral subtraction.

Temporal Processing, RASTA

- CMS can be viewed as a special case of a more general idea of filtering the time trajectory of speech features. In other words, we go through the feature sequences and change the feature values.
- Another example of temporal processing is the RASTA filtering, which is given in Figure 22.5.
- It suffices to say that temporal filtering is quite effective for noise robustness.