Acoustic Features Combination For Robust Speech Recognition

Author : Andras Zolnay, Ralf Schluter, Hermann Ney

Source: Acoustics, Speech, and Signal Processing

Professor: 陳嘉平

Reporter: 吳國豪

Outline

- Introduction
- Signal analysis
- Combination
- Experiments

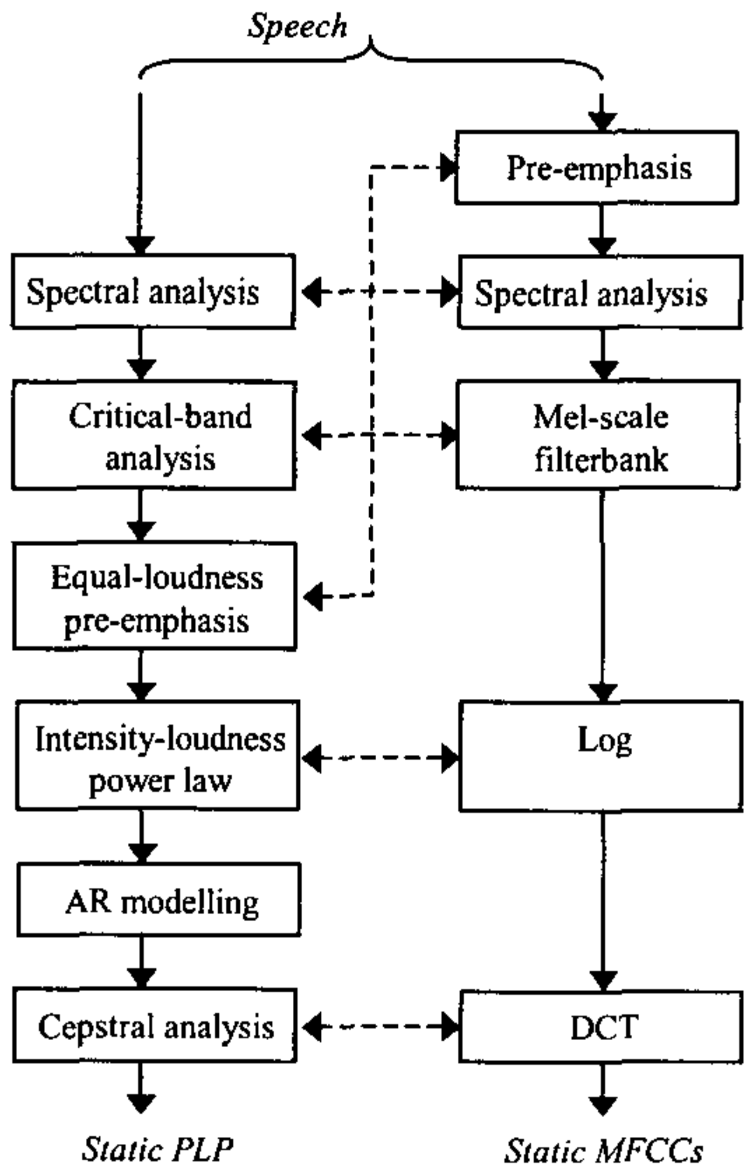
Introduction

- Most automatic speech recognition systems use auditory based representation of the speech signal, e.g. Mel Frequency Cepstrum Coefficients (MFCC), Perceptual Linear Prediction (PLP).

- In this paper, we consider the use of multiple acoustic features of the speech signal for robust speech recognition.

Signal analysis

- Mel Frequency Cepstrum Coefficients (MFCC)
- Perceptual Linear Prediction (PLP).


Speech
Pre-emphasis
Spectral analysis
Spectral analysis
Critical-band analysis
Mel-scale filterbank
Equal-loudness pre-emphasis
Intensity-loudness power law
Log
AR modelling
Cepstral analysis
DCT
Static PLP
Static MFCCs

MFCC-AllPoles

- In this method, MFCCs are derived from the **all-poles magnitude spectrum estimate** instead of the magnitude spectrum estimated by using Fast Fourier Transform.

- In the all-poles estimate, the magnitude spectrum $\left|X^t(w)\right|$ of a time frame t is assumed to have the form of

$$\left|X^t(w)\right| \approx \frac{g^t}{\left|1+\sum_{k=1}^{M} a_k^t e^{-jwk}\right|}$$

where g^t is called the gain, at *k* is a autoregressive coefficient, and *M* is number of autoregressive coefficients.

MF-PLP

- In this method, the MFCC and PLP techniques are merged into one algorithm.

- The first steps until generating the output of the Mel scale triangular filter bank are taken from the MFCC algorithm.

- The only difference here is that the filter bank is applied to the power spectrum instead of the magnitude spectrum.

- The last steps generating the cepstrum coefficients are taken from the PLP algorithm.

Voicedness Feature

- Voicedness feature is a measure representing the state of the vocal cords.

- The measure describes how periodic the speech signal is in a given time frame *t. We use the autocorrelation* function to measure periodicity.

Voicedness Feature

- Autocorrelation $\tilde{R}^t(\tau)$ expresses the similarity between the time frame $x^t(v)$ and its copy shifted by τ. We have used the unbiased estimate of autocorrelation $\tilde{R}^t(\tau)$ *:*

$$\tilde{R}^t(\tau) = \frac{1}{T-\tau} \sum_{v=0}^{T-\tau-1} x^t(v) x^t(v+\tau)$$

where *T* is the length of a time frame.

- Autocorrelation of periodic signals with frequency *f* attains its maximum $\tilde{R}^t(\tau)$ not only at $\tau = 0$ but also at $\tau = \frac{k}{f}$ $k = 0, \pm 1, \pm 2, \ldots$ integer multiples of the period.

Voicedness Feature

- In order to produce a bounded measure of voicedness, autocorrelation is divided by $\tilde{R}^t(0)$.

- The voicedness measure v^t is thus the maximum value of the normalized autocorrelation in the interval of natural pitch periods [2*.5ms..12.5ms]:*

$$v^t = \frac{\max\limits_{2.5ms \cdot f_s \leq \tau \leq 12.5ms \cdot f_s} \tilde{R}^t(\tau)}{\tilde{R}^t(0)}$$

where f_s denotes the sample rate. Values of v^t close to 1 indicate voicedness, values close to 0 indicate voiceless time frames.

LDA based feature combination

- The Linear Discriminant Analysis (LDA) based approach combines directly the different acoustic feature vectors.
- In the first step, feature vectors extracted by different algorithms $x_t^{f_i}$ are concatenated for all time frames *t.*
- In the second step, 2*L + 1* successive concatenated vectors are concatenated again for all time frames *t* which makes up the large input vector of LDA

LDA based feature combination

- The combined feature vector y_t is created by projecting the large input vector on a smaller subspace:

$$\boldsymbol{y}_t = \left[\boldsymbol{V}^T \right] \left[\begin{array}{c} \left[\begin{array}{c} x_{t-L}^{f_1} \\ \cdots \\ x_{t-L}^{f_F} \end{array} \right] \\ \left[\begin{array}{c} x_t^{f_1} \\ \cdots \\ x_t^{f_F} \end{array} \right] \\ \left[\begin{array}{c} x_{t+L}^{f_1} \\ \cdots \\ x_{t+L}^{f_F} \end{array} \right] \end{array} \right]$$

Log-linear model combination

- In this approach, different acoustic features are combined indirectly via the log-linear combination of acoustic probabilities $P_{f_i} = (X^{f_i} | W)$ *where W* denotes a sequence of words and X^{f_i} denotes a sequence of feature vectors extracted by the algorithm f_i *.*

- The basic idea is to modify the modeling of the posterior probability *P(W|X) in Bayes' decision rule:*

$$W_{opt} = \arg\max_{W} P(W|X)$$

Log-linear model combination

- In the standard case, posterior probability is decomposed into language model probability *P(W)* and acoustic model probability *P(X|W):*

$$P(W|X) = \frac{P(W)P(X|W)}{\sum_{W'} P(W')P(X|W')}$$

- In the case of log-linear model combination, the posterior probability has the following form:

$$P(W|X) = \frac{e^{\sum_i \lambda_i g_i(W,X)}}{\sum_{W'} e^{\sum_i \lambda_i g_i(W',X)}}$$

where g_i is called feature function

Log-linear model combination

- The basic feature function types are negative logarithm of probabilities:
  language model: $g_i^{LM}(W, X) = -\log P_i(W)$
  acoustic model: $g_i^{AM}(W, X) = -\log P_i(X|W)$

- Finally, in order to combine different acoustic features, we introduce a separate acoustic model $P_{f_i}(X^{f_i}|W)$ for each feature.

Log-linear model combination

- Using a single language model feature function and for each feature a separate acoustic model feature function, the Bayes' decision rule for log-linear feature combination can be written as:

$$W_{opt} = \arg\max_{W} P(W)^{\lambda_{LM}} \prod_{i} P_{f_i}(X^{f_i} | W)^{\lambda_{f_i}}$$

- Acoustic training of the combined system consists of two steps: independent training of each acoustic model $P_{f_i}(X^{f_i} | W)$ *and* training of the language model weight λ_{LM} and the acoustic model weights λ_{f_i}.

Experiments

- Recognition tests have been conducted on the large-vocabulary corpus ***VerbMobil II****.* The corpus consists of German conversational speech: 36k training-sentences (61.5h) from 857 speakers and 1k test-sentences (1.6h) from 16 speakers.

Acoustic Feature	Error Rates [%]		
	Del	Ins	WER
MFCC	6.3	2.4	23.1
MFCC-VTLN	5.0	2.7	21.3
MFCC-AllPoles	6.2	2.7	24.2
PLP	6.6	2.3	23.1
MF-PLP	6.2	2.7	23.2

Table 1. Baseline recognition results with different features.

Combined Features	Error Rates [%]		
	Del	Ins	WER
MFCC	6.3	2.4	23.1
MFCC + MFCC-AllPoles	5.8	2.5	22.6
MFCC + MF-PLP	6.0	2.6	22.9
MFCC + MF-PLP + PLP	5.6	2.6	22.1

Table 2. Recognition results of combining state-of-the-art features (MFCC, MFCC derived from all-poles magnitude spectrum, MF-PLP, and PLP) by using log-linear model combination.

LDA				Log-Linear			
Combined Features	Error Rates [%]			Combined Features	Error Rates [%]		
	Del	Ins	WER		Del	Ins	WER
MFCC + Voice	5.7	2.8	22.4	MFCC + Voice	6.1	2.7	23.0
				MFCC + LDA(MFCC + Voice)	5.9	2.7	22.2
MFCC-VTLN + Voice	5.1	2.6	20.8	MFCC-VTLN + LDA(MFCC + Voice)	5.3	2.3	20.3
MFCC + MFCC-VTLN + Voice	5.1	2.5	20.7	LDA(MFCC + Voice)+LDA(MFCC-VTLN + Voice)	5.3	2.2	19.9

Table 3. Recognition results of combining MFCC, vocal tract length normalized MFCC (MFCC-VTLN), and voicedness features (Voice). On the left, features are combined by LDA, on the right by log-linear model combination. LDA(MFCC + Voice) denote an acoustic model trained on the LDA based combination of MFCC and voicedness features.