Using asymmetric windows in automatic speech recognition

Author: Robert Rozman *, Dusan M. Kodek

Professor:陳嘉平
Reporter:葉佳璋

Outline

- Introduction
- Window In Speech Recognition
- Asymmetric windows designed with FIR filter methods
- Practical Evaluation

Introduction

- Symmetric windows are widely used in the field of digital signal processing.

  -ease of design

  -linear phase property

- Potential drawbacks

  -longer time delay

  -frequency response limitations

Introduction(cont.)

- Human hearing is relatively insensitive to short-time phase distortion there is no apparent reason for the use of symmetric windows which give a linear phase response.

- Removal of the symmetry constraint can therefore give asymmetric windows having some better properties.

Window In Speech Recognition

- The short time Fourier Transform (STFT) is a common frequency analysis method in speech recognition.

x(n) =s(n)w(n), n=0, …., N-1

The frame length N must be short because of the rapidly changing spectrum of s(n). A longer N gives better spectral resolution but worse temporal resolution.

$$X(e^{jw}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{j\theta})W(e^{j(\omega-\theta)})d\theta$$

The windowed spectrum $X(e^{jw})$ is calculated as the frequency response of x(n). $X(e^{jw})$ is also equal to the convolution integral of Fourier Transform(FT) of the window sequence of $W(e^{j\omega})$ and the FT of the origin signal $S(e^{j\omega})$.

Window In Speech Recognition(cont.)

- Wish to select a window function w(n) in such a way that the computed magnitude response $|X(e^{jw})|$ is as " near possible" to the real magnitude response $|S(e^{jw})|$.

- For a given N the asymmetric window idea arises naturally here:

  removing symmetry constraint can increase the spectral resolution giving a better $|X(e^{jw})|$.

Window In Speech Recognition(cont.)

- It is possible to make the following fundamental assumptions about the window magnitude response for use in speech recognition:

  - Human speech perception is almost insensitive to short time phase distortions in the speech signal. The ear performs frequency analysis with lower frequecy resolution and heavily overlapped fitters with rapidly decaying side-lobes.

  - Speech recognition system usually discard the phase information and perform wideband frequency analysis in the parameterization process. This means that the linearity of phase constraint can be removed without any adverse effects.

Window In Speech Recognition(cont.)

- Basically there are two major signal representation distortions introduced by the inevitable windowing process:

  - spectral smearing

  - spectral leakage

- Since both smearing and leakage can not be minimize at the same time their importance should be established.

- Based on experiments it seems that the distant spectral leakage, or more general side-lobe, is important for speech robustness.

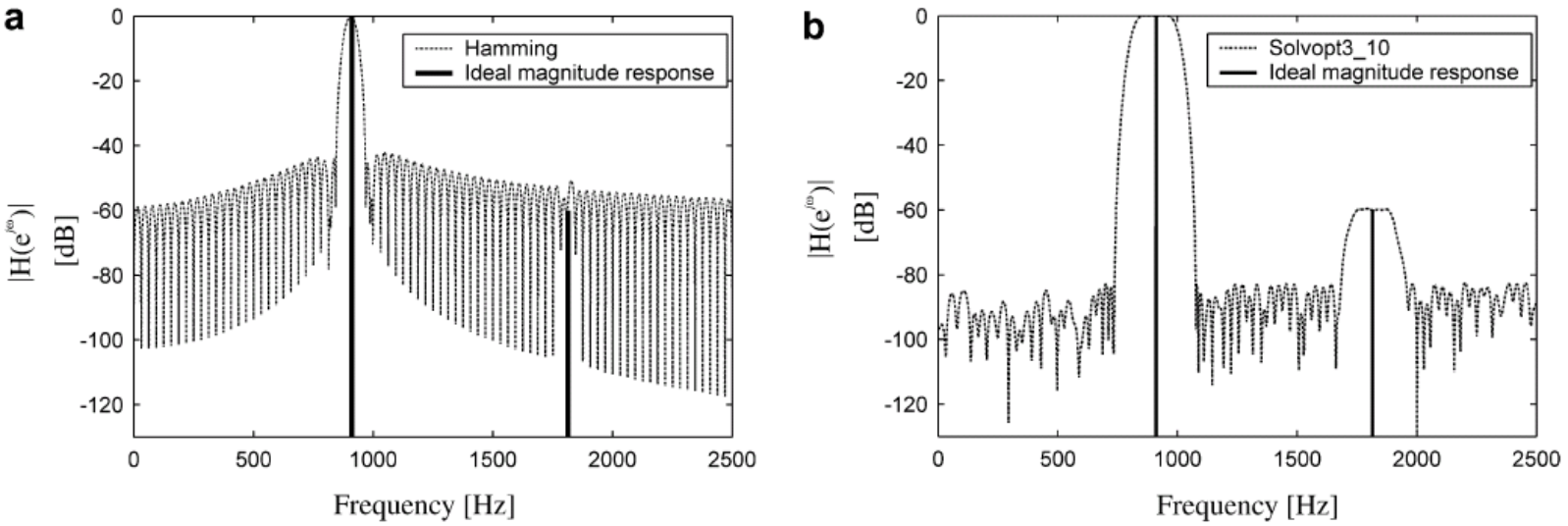


Fig. 2. Comparison of window influence on the computed magnitude response of a simple two tone signal using: (a) Hamming window, (b) Solvopt3_10 – asymmetric window with lower side-lobes.

Window In Speech Recognition(cont.)

- Most real speech recognition systems (SRSs) that are based on HMMs approach use diagonal covariance matrices as a computational simplification of the time consuming processing of full covariance matrices.

- The error introduced with this simplification is smaller if components of feature vectors are uncorrelated.

- In this context it is interesting to observe that the lowering the spectral leakage helps decorrelate FBANK features.

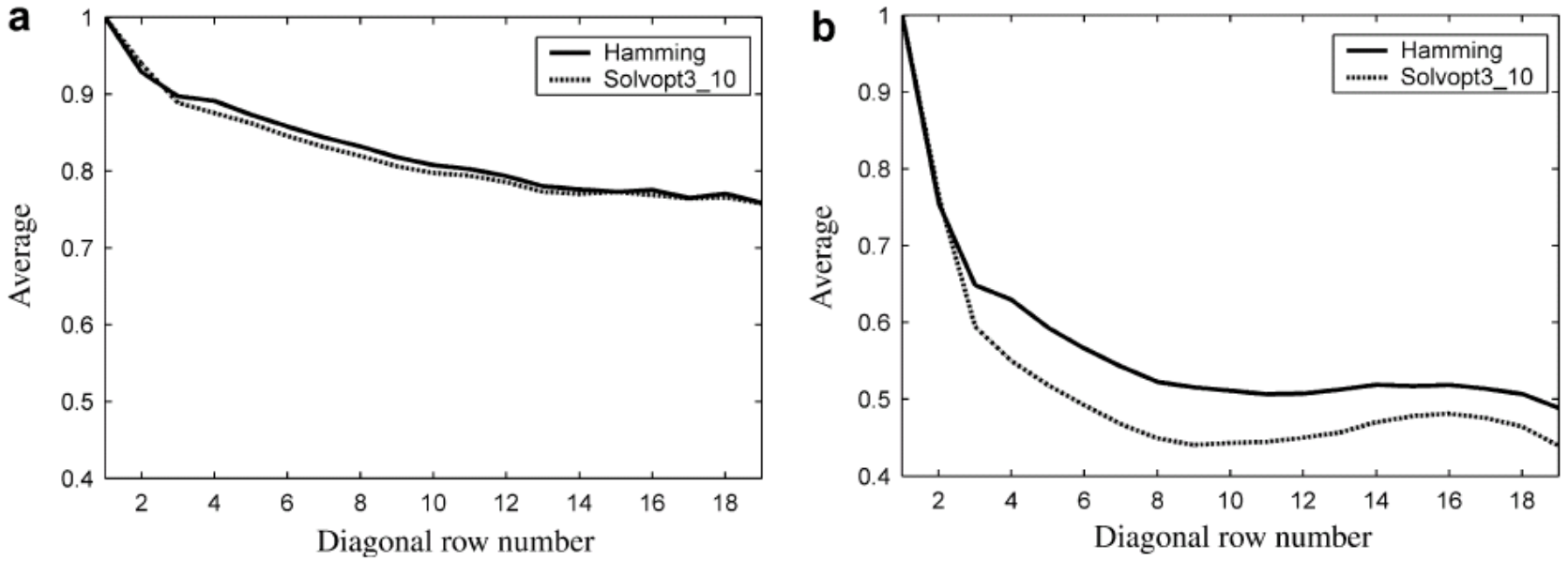
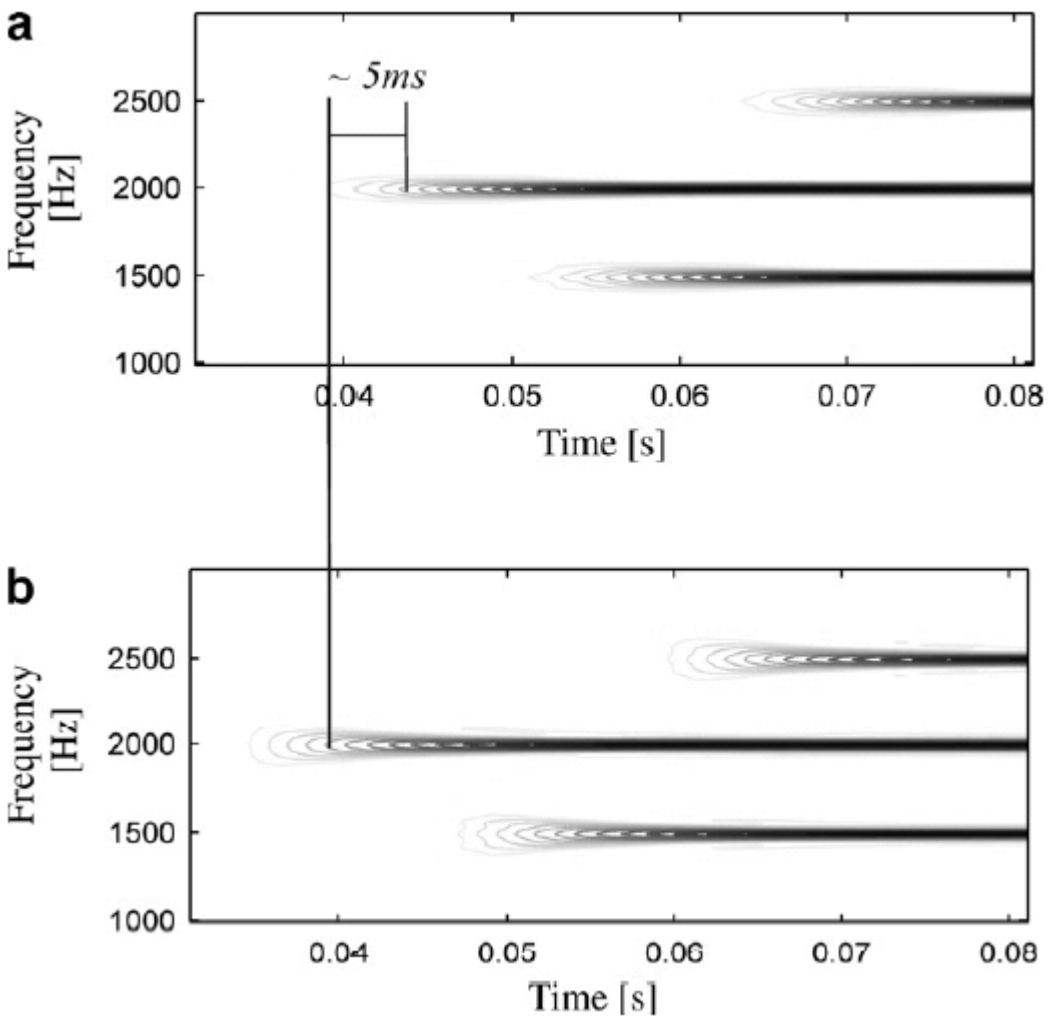


Fig. 3. Window influence on the average correlation of FBANK features in: (a) clean conditions and (b) added white noise at 6 dB.

Window In Speech Recognition(cont.)

- Asymmetric windows also bring shorter time delay which at first does not seem to be of major importance.
- But recently unified distributed platforms for speech communication, recognition, and synthesis have appeared.
- Therefore a fast reconstruction of the time-domain signal is required for “live” spoken communication.


a
~ 5ms
Frequency [Hz]
2500
2000
1500
1000
0.04
0.05
0.06
0.07
0.08
Time [s]
b
Frequency [Hz]
2500
2000
1500
1000
0.04
0.05
0.06
0.07
0.08
Time [s]

Window In Speech Recognition(cont.)

- These observation lead to a reasonable doubt that linear phase windows are optimal for speech recognition.

- The following properties are expected to be more important for speech recognition performance:

  - lower side-lobes

  - monotone, rapidly decaying height of side-lobes

  - shorter time delay(less important for recognition alone)

Asymmetric windows designed with FIR filter methods

- All window functions are of finite length N which makes it possible to treat them as if they are the impulse response h(n).

- Two types of asymmetric window design problems were investigated. The first one is denoted “nearly linear phase” window and is defined as:

- Find the optimal impulse response of length N, h*=[h*(0),h*(1),…,h*(N-1)], that has the minimal error according to the minmax (or Chebyshev) criterion

Asymmetric windows designed with FIR filter methods(cont.)

$$\delta(h^*) = \min_{h} \delta(h)$$

$$\delta(h) = \max_{\omega \in \Omega} W(e^{j\omega}) \mid E(e^{j\omega}) \mid$$

$$\mid E(e^{j\omega}) \mid = \sqrt{(\mathrm{Re}\{E(e^{j\omega})\})^2 + (\mathrm{Im}\{E(e^{j\omega})\})^2}$$

$$E(e^{j\omega}) = D(e^{j\omega}) - H(e^{j\omega})$$

Where $\delta(h)$ is the Chebyshev error of sequence h, $D(e^{j\omega})$ is the desired and $H(e^{j\omega})$ the real frequency response. $W(e^{j\omega})$ is a positive weighting function and Ω is a set of discrete frequencies

Asymmetric windows designed with FIR filter methods(cont.)

- The phase and magnitude errors contribute equally to the final error value. This lead to a window that is typically not too different from the symmetric linear phase window.

- The second type of asymmetric window is denoted “arbitrary phase” window. The complex error function is replaced by the magnitude-only error function

$$E(e^{j\omega}) = |D(e^{j\omega})| - |H(e^{j\omega})|$$

Practical Evaluation

- The asymmetric arbitrary phase Solvopt3 windows were design as described above.

- The stopband weighting function $W(e^{j\omega})$ was set to 10, 100, and 1000 the three windows Solvopt3_10, Solv3_100, Solvopt3_1000

Speech recognition system

- The practical evaluations were performed on two real SRS based on different approaches and recognition task complexities:

  - isolated word recognizer based on Hidden Markov Models – ”HTK”

  - connected digit recognizer based on Neural Networks(NN) – “CSLU”

Speech recognition system(cont.)

- The first one is base on the statistical approach. It recognizes one word at a time.

-Whole word continuous models are used with 8 internal states, mixtures of eight Gaussian densities and diagonal covariance matrices

- It is implemented using the HTK software

-12MFCC features together with logarithm of the frame energy and the corresponding 13 delta features were used giving a total of 26 features.

Speech recognition system(cont.)

- The second system uses a neural network in the form of a multi-layer perceptron with 200internal neurons. It is capable of recognizing whole utterances of concatenated word.

  - context dependent speech units are used.

  - a simplified form of Viterbi search procedure is use on the results from the perceptron classification stage.

  - the advantage of this approach is lower time and space complexity.

  - it is implemented in the CSLU Speech Toolkit.

  - the CSLU recognizer uses only MFCC feature without delta features.

  - although the vector at time t is actually a concatenation of 5 vectors at t-60ms, t-30ms, t+30ms, t+60ms. This sums up to a final 13*5=65 features

Speech databases

- All experiments were carried out on different speech database: one in English and Slovenian.
- SLO-DIGITS database we used in both SRS.

- it consists of 780 Slovenian adult speaker uterances recorded over public telephone lines with their inherent noise.
- simple 13-word vocabulary(0 to 9 digit and word yes ,no, stop)
- each utterance consisted of all 13 words in random order.
- in practical evaluations 234 speakers were used for training the recognizer.
- the test and validation sets consist of 156 speaker each.

Speech databases(cont.)

- To enhance the variability of evaluation condition the English “Number 95 ” database was also used with the CLSU recognizer.

  - it consists of connected numbers utterances recorded over telephone lines

  - for our task only utterances with digit strings were used(having an average of 5-6 digits per utterance)

  - there are 1368speakers in the training set, 555 speakers in the validation set and 1168 speakers in the test set.

Inherent robustness evaluation

- It should be stressed that recognizers ere trained on the “clean” training set that did not contain any added noise.

- The following seven additive noise recordings derived from the NOISEX database were used:
  - speech in background (‘‘Babble’’),
  - noise in pilot cockpit of F-16 (‘‘F-16’’),
  - factory noise (‘‘Factory1’’),
  - car noise (‘‘Volvo’’),
  - pink noise (‘‘Pink’’),
  - white noise (‘‘White’’),
  - filtered white noise centered around 900 Hz (‘‘Pass 900’’).

- Testing was performed on the following three major test groups:
  – ‘‘Clean’’ test group is the original test set.
  – ‘‘Additive’’ test group consists of 7 test sets that were obtained by adding the 7 noise recordings to ‘‘Clean’’ test set for a specific signal to noise ratio (SNR). Three different SNR values were used giving a total of 21 test sets.
  – ‘‘Additive+LP’’ convolutional test group was formed by additional lowpass filtering of all ‘‘Additive’’ test sets.

Table 1 WER on connected digit task

WER (%)	Clean	Additive (SNR)			Additive + LP (SNR)			Mean
		12db	6db	0db	12db	6db	0db	
Hamming	2.6	16.6	31.2	56.6	35.9	52.2	73.0	38.3
Solvopt3_10	3.0	14.3	30.2	58.1	26.2	40.6	64.2	33.8
Solvopt3_100	2.9	13.2	28.9	58.2	25.7	40.0	65.0	33.4
Solvopt3_1000	2.9	13.3	28.9	55.9	25.5	39.5	62.1	32.6

CSLU SRS on Number 95 database was used

Table 2 WER on connected digit task

WER (%)	Clean	Additive (SNR)			Additive + LP (SNR)			Mean
		12db	6db	0db	12db	6db	0db	
Hamming	5.9	13.7	26.7	46.7	45.7	58.0	69.4	38.0
Solvopt 3_100	5.0	12.3	24.2	44.6	20.9	31.3	49.8	26.9

CSLU SRS on SLO-DIGITS (isolated) database was used

Table 3 WER on isolated digit task

WER (%)	Clean	Additive (SNR)			Additive + LP (SNR)			Mean
		12db	6db	0db	12db	6db	0db	
Hamming	4.5	12.9	22.0	37.3	38.2	48.7	61.5	32.2
Solvopt 3_100	5.1	14.9	24.2	39.5	23.0	30.8	43.3	25.8

HTK SRS on SLO-DIGITS (isolated) database was used

Table 4 WER on isolated digit task

WER (%)	Clean	Additive (SNR)			Additive + LP (SNR)			Mean
		12db	6db	0db	12db	6db	0db	
Hamming	4.3	13.8	24.6	44.1	34.3	46.4	64.1	33.1
Solvopt 3_100	4.2	12.6	22.5	40.6	21.1	29.9	47.6	25.5

HTK SRS with RASTA + delta features on SLO-DIGITS (isolated) database was used

Table5 WER on connected digit task and different noise types(SNR=12)

WER (%)	White	Pink	Babble	Volvo	Factory1	F-16	Pass 900	Mean
Hamming	11.3	12.3	13.6	13.5	14.0	10.3	20.6	13.7
Solvopt 3_100	9.6	9.5	12.0	12.6	11.6	9.7	21.2	12.3

CSLUSRS with on SLO-DIGITS (isolated) database was used