EUROCON 2003 Ljubljana, Slovenia

Improving speech recognition robustness using non-standard windows

Robert Rozman, Dušan M. Kodek
Faculty for Computer and Information Science, University of Ljubljana
Tržaška 25, 1001 Ljubljana
Slovenia
email: rozman@fri.uni-lj.si, kodek@fri.uni-lj.si

Abstract--**Windowing problem of the short-time frequency analysis in Speech recognition systems (SRS) is considered. Design possibilities for different non-standard window sequences are presented. Traditional "digital filtering" approach to the design of finite window sequences with linear and nonlinear phase response is examined. Since human hearing is relatively insensitive to phase distortions of speech signal, some other ideas of alternative windows with nonlinear phase response are also investigated. Two most promising design methods for nonlinear phase windows are discussed. Practical performance comparison of such windows with the Hamming window on two real SRS is presented. They show that the non-standard window sequences can contribute to greater SRS robustness. An additional research on non-standard windows and parametrization process as a whole is suggested.**

***Index Terms*--Robustness, speech processing, speech recognition, windowing.**

I. Introduction

In Speech Recognition Systems (SRS) Short Time Fourier Transform (STFT) is commonly used as a frequency analysis method. Long signals are divided into short frames of N samples. We get final values $x(n)$ in frame by multiplying signal $s(n)$ with nonzero samples of window sequence $w(n)$

$$x(n) = s(n)\, w(n), \quad n = 0, \ldots, N-1. \tag{1}$$

Typical frame durations in speech recognition are 10-30 ms.

Using STFT we get $X(e^{j\omega})$ as frequency response of $x(n)$. $X(e^{j\omega})$ is a convolution integral of Fourier Transform (FT) of the window sequence $W(e^{j\omega})$ and FT of the original (non-framed) signal $S(e^{j\omega})$

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

This equation gives information about the ideal frequency response $W(e^{j\omega})$ of a window that would minimize distortion introduced due to inevitable use of window sequence. It is typical for speech recognition that only the magnitude frequency response of signal samples in the frame is kept for further processing. Therefore we wish that the computed magnitude response $|X(e^{j\omega})|$ is as "near" as possible to the real magnitude response $|S(e^{j\omega})|$.

These statements are true for the common "digital filtering" point of view. But when we design window sequences for SRS some other aspects are also important. We have no theoretical reason to believe that best window sequences, which fulfill the common digital filtering criterion, will also perform adequately in these systems. We therefore studied carefully the properties of human auditory perception and tried to incorporate some selected features into our implementations of SRS. Several interesting ideas appeared when this approach was used.

One of them is the idea of windows with wider main-lobes in magnitude response. Wideband time-frequency signal representation is usually used in SRS as a base for further processing. In this case it is obvious that accurate frequency analysis or use of windows with narrow main-lobes is probably not needed. There is also reasonable doubt that partially constant desired magnitude response (common case in digital filtering) is also the best criterion when designing windows for speech recognition systems.

After the definition of the desired window's magnitude response we can solve the design problem with methods that are well known for the design of FIR digital filters. But we use an approach that is quite different from the standard FIR filter design that is concerned with the linear phase response filters. We are interested in the design of FIR filters with nonlinear or nearly linear phase. It is a well-known fact that human auditory perception is quite insensitive for phase distortion of speech signals. Another fact that follows from the digital filter theory is that we can get a better amplitude response of a FIR filter if the linear phase constraint is relaxed. This leads to a more accurate magnitude spectrum of signal samples in a frame.

As a second possibility we also investigate an approach that leads to windows with certain human auditory perception features. The last section presents a practical evaluation of different windows on two real SRS with system's robustness as major optimality criteria. Detailed comparison of practical performance under different conditions is given in two tables.

II. Windows design possibilities

Most SRS implementations use one of the standard windows (Hamming, Hann, Blackman, Kaiser). Since speech recognition systems are still in a pre mature stage, more attention is given to general problems rather than to details such as the design of windowing functions.

Our research showed that windows are more important than one might expect, especially when we evaluate system's robustness. In the following sections we will show different possibilities of designing window functions and their comparison.

A. Standard windows

Using one of the standard windows gives a choice between different main-lobe width and side-lobes height relations. All windows are defined with a closed form expressions and therefore easily computable; they are also symmetrical (linear phase) and have a particular shape of the magnitude response.

B. Digital filtering design methods for windows

From a digital filtering viewpoint, we can design different windows with the help of the well-known design methods for Finite and Infinite Impulse Response (FIR and IIR) filters. Since windows are finite sequences the design methods for FIR filters will be examined first. The design problem can be defined as:

Find the optimal impulse response of length N, $\boldsymbol{h}^* = [h^*(0),\ h^*(1)\ ,...,\ h^*(N\text{-}1)]$*, that has the minimal error according to the minimax (or Chebyshev) criterion*[1]

$$\delta(\boldsymbol{h}^*) = \min_{\boldsymbol{h}} \delta(\boldsymbol{h})\,, \tag{3}$$

$$\delta(\boldsymbol{h}) = \max_{\omega\in\boldsymbol{\Omega}} W(e^{j\omega})\left|E(e^{j\omega})\right|, \tag{4}$$

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h(n)\, e^{-j\omega n}\,, \tag{5}$$

$$E(e^{j\omega}) = D(e^{j\omega}) \text{ - } H(e^{j\omega}), \tag{6}$$

where $\delta(\boldsymbol{h})$ is the Chebyshev error of sequence $\boldsymbol{h}$, $D(e^{j\omega})$ is the desired and $H(e^{j\omega})$ the real frequency response. $W(e^{j\omega})$ is a positive weighting function and $\boldsymbol{\Omega}$ is a set of discrete frequencies[2], on which the error function $E(e^{j\omega})$ is evaluated; its absolute value can be computed as

$$\left|E(e^{j\omega})\right| = \sqrt{(\mathrm{Re}\{E(e^{j\omega})\})^2 + (\mathrm{Im}\{E(e^{j\omega})\})^2}\,. \tag{7}$$

The minimax approximation problem (3)-(6) is nonlinear and therefore difficult to solve. The problem becomes linear and much simpler if linear phase constraints on $D(e^{j\omega})$ and $H(e^{j\omega})$ are introduced. It can be solved with the efficient Parks-McClellan algorithm. But we are not interested in the linear phase case and must solve the difficult version of the problem.

Several methods for solving (3)-(6) were tried [1], [2]. In addition to the complex error function (6) we also tried the simpler "magnitude only" error function

$$E(e^{j\omega}) = |D(e^{j\omega})| \text{ - } |H(e^{j\omega})|\,. \tag{8}$$

When (6) is used the phase error is weighted with the same weight function as the magnitude error. This leads to a "nearly linear phase" filter. When (8) is used the phase error is completely ignored and this leads to a "nonlinear phase" filter. An example of the corresponding solutions is given in Figures 1 and 2. It shows how the gradual relaxation of phase linearity constraints leads to a better magnitude response.

We solved problem (3)-(5) and (6) as error function with the modified linear programming method. In case of (8) the general-purpose optimization procedure "SOLVOPT"[3] was used.

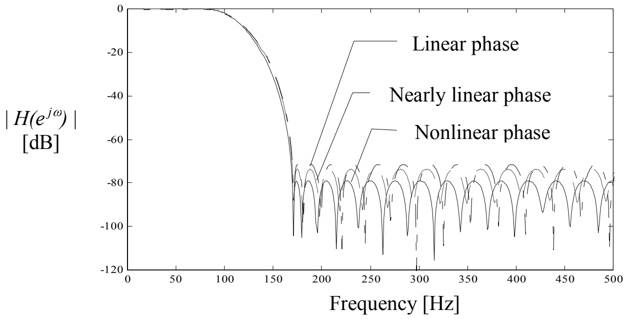

Fig. 1. Magnitude responses of windows designed with FIR design methods (*Fs*=1000Hz, *N*=272).

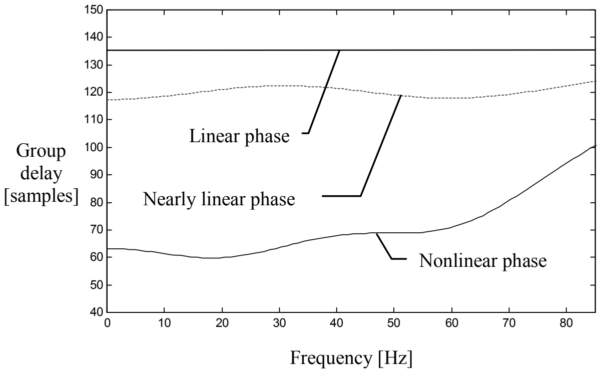

Fig. 2. Group delay in main-lobe of windows designed with FIR design methods (*Fs*=1000Hz, *N*=272).

Our previous research [3], [4] confirmed that linearity relaxation in windows, and hence a better magnitude response, leads to better robustness of SRS. A major drawback of this approach is the complexity of the design process that requires a solution of a difficult minimax approximation problem in (3)-(5) and (6) or (8).

C. IIR windows

We examined several simpler methods of getting similar windows with comparative robustness enhancements. Two approaches emerged with good practical evaluation results.

The first is a family of "IIR windows"[4], where the window sequence is equal to first N samples of impulse response of a simple, even order IIR system:

$$H(z) = \frac{1}{\left(1-\alpha\, z^{-1}\right)^M} \quad \alpha \in (0..1),\ M = 2,4,6,8,10. \tag{9}$$

[1] Most common criteria used in digital filtering.

[2] $\boldsymbol{\Omega}$ is union of compact, non overlapping subintervals of [0 .. π].

[3] URL address: *http://www.kfunigraz.ac.at/imawww/kuntsevich/solvopt/*.

[4] Named after IIR system.

The second is a family of "smoothed exponential windows" that follow from (9) when M=2. In this case the impulse response $h(n)$ *is* equal to

$$h(n) = n\,\alpha^{n}, \quad n = 0\,..\,N\text{-}1\,. \tag{10}$$

The window sequence is obtained by smoothing $h(n)$ with a well-known Hann window

$$w(n) = h(n)\ w_{Hann}(n)\,, \quad n = 0\,..\,N\text{-}1\,. \tag{11}$$

Varying values N and α in (10) and (11) gives different "smoothed exponential windows" that are evaluated[5] below.

D. Comparative analysis

Examples of "IIR windows" and "smoothed exponential windows" of length 256 are shown in Figure 3 where a Hamming window is added for the reference. Their magnitude responses are given in Figure 4; the corresponding parameters can be found in Tables 1 and 2.

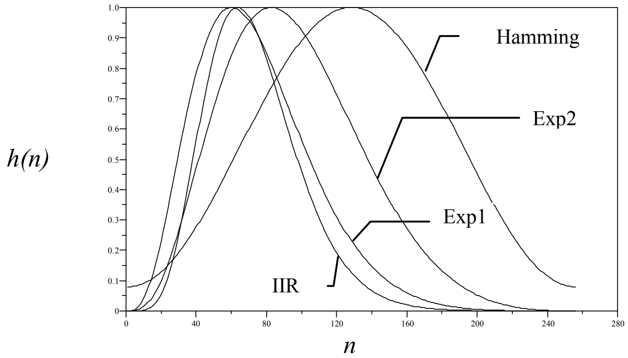

Fig. 3. Comparation of selected IIR sequences with reference Hamming window (N=256).

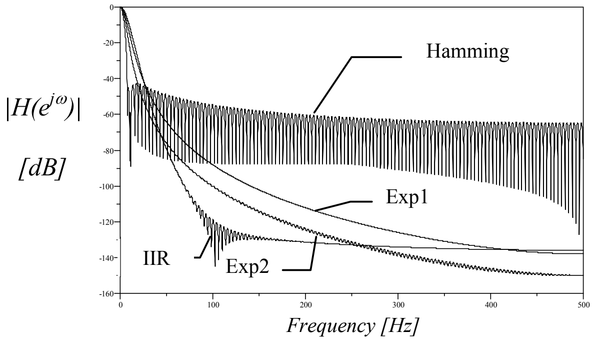

Fig. 4. Magnitude spectrum comparation of selected IIR windows with Hamming window (Fs=1000Hz).

Figure 3 shows that these windows are not symmetrical, which in turn gives nonlinear phase response. Figure 4 is more interesting; it shows that their main-lobes are wider and the side-lobes lower with a good asymptotical attenuation – properties that were desired at the first place.

III. Practical evaluation

IIR windows and smoothed exponential windows were tested on two real SRS that use different approaches and recognition task complexity:

- isolated word recognizer based on HM (Hidden Markov) Models,
- connected digit recognizer based on NN (Neural Networks).

In both systems STEVKE speech collection [4] was used. It consists of 780 Slovenian adult speakers' utterances recorded over public telephone lines with its inherent noise. Small portion of database was manually transcribed for further research.

In practical evaluations we randomly chose 200 speakers for train and 100 speakers for test set. Simple, 13-word vocabulary (digits from "0" to "9" and words "yes", "no" and "stop") was used. "Standard" set of MFCC[6] and corresponding Delta features served as speech signal representation. Connected digit recognizer was fed with whole utterances of 13 words from each speaker, while isolated SRS recognized one word at the time. In both cases Word Success Rate (WSR) was measured.

It should be stressed that both systems were trained and initially tested on "clean" train set. Their robustness was evaluated as their performance on noisier, simulated conditions that were not present in training phase. There was no additional adaptation performed prior such testing.

We simulated new testing conditions with addition of following additive noise recordings from NOISEX database [5] :

- speech in background ("Babble"),
- noise in pilot cockpit of F-16,
- factory noise,
- car noise,
- pink noise,
- white noise.

Three test groups were formed for practical evaluations:

- "*Clean*" test group is actually original test set of 100 speakers,
- "*Additive noise*" test group consisted of 6 test sets acquired with mixing noise recordings to "*Clean*" test set,
- "*Additive noise+lowpass filter*" test group was gained by additional lowpass filtering[7] of "*Additive noise*" test group.

In Tables 1 and 2 the recognition rates for both systems on three test groups are shown. Due to different noise adding methods, only comparison among different windows on each system can be made. Despite similar performances of all windows in original ("*Clean*") conditions, quite impressive robustness enhancement can be noticed on second and even more on third test group with additive noise and lowpass convolutional speech signal distortion.

[5] Named as "Exp1" and "Exp2".

[6] Most frequently used feature type in today SRS.

[7] Example of convolutional signal distortion.

TABLE I
PRACTICAL EVALUATION ON ISOLATED DIGIT TASK – HMM BASED
WORD SUCCESS RATE (WSR) IN PERCENTS

	Clean	Additive noise	Additive noise + lowpass filter
Hamming	97,69	75,52	48,87
IIR *α=0.9, M=8*	98,15	75,50	69,37
Exp1 *α=0.9564, M=2*	98,08	75,98	69,10
Exp2 *α=0.9725, M=2*	97,31	75,26	68,05

TABLE 2
PRACTICAL EVALUATION ON CONNECTED DIGIT TASK – NN BASED
WORD SUCCESS RATE (WSR) IN PERCENTS

	Clean	Additive noise	Additive noise + lowpass filter
Hamming	95,96	82,10	75,98
IIR *α=0.9, M=8*	96,30	82,06	81,05
Exp1 *a=0.9564, M=2*	96,70	86,14	82,82
Exp2 *a=0.9725, M=2*	96,79	86,00	83,75

IV. CONCLUSION

Practical evaluations of speech recognition system's robustness suggest that further research on using nonlinear phase and more general non-standard windows is worthwhile. Speech recognition robustness can greatly eliminate the need for developing the new systems from scratch for each type of the real conditions. This also dramatically reduces costs. An existing robust speech recognition system requires only a simple and fast adaptation, if needed at all, for any new circumstances.

REFERENCES

[1] R. Rozman and D. Kodek, "Design of optimal FIR filters according to the complex Chebyshev criteria," *Proceedings of the sixth International Electrotechnical and Computer Science Conference ERK 1997, Portorož, Slovenia*, vol. A, pp. 175-178, Sept. 1997.

[2] R. Rozman and D. Kodek, "Generalized Remez algorithm for design of optimal FIR filters according to the complex Chebyshev error criteria, " *Proceedings of the seventh International Electrotechnical and Computer Science Conference ERK 1998, Portorož, Slovenia*, vol. A, pp. 245-248, Sept. 1998.

[3] R. Rozman, A. Štrancar and D. Kodek, "Analysis of window function influence on robustness of speech recognition systems," *Proceedings of the ninth International Electrotechnical and Computer Science Conference ERK 2000, Portorož, Slovenia*, IEEE Region 8, vol. B, pp. 177-180, Sept. 2000.

[4] R. Rozman and D. Kodek, "Speech database ŠTEVKE and robustness of Speech Recognition Systems," *Language technologies: proceedings of the conference*, Ljubljana, Institut Jožef Stefan, pp. 75-78, 2000.

[5] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, 12(3), pp. 247-251, 1993.