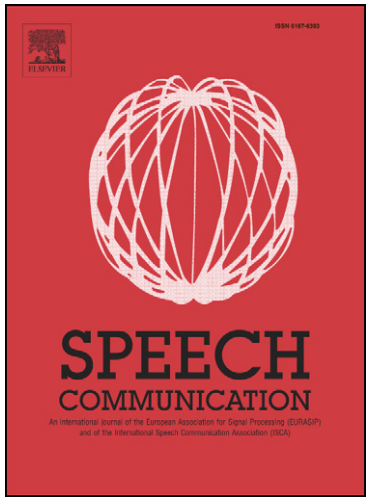
Using asymmetric windows in automatic speech recognition

Robert Rozman, Dušan M. Kodek



Using asymmetric windows in automatic speech recognition

Robert Rozman, Dušan M. Kodek

PII: S0167-6393(07)00025-8
DOI: 10.1016/j.specom.2007.01.012
Reference: SPECOM 1614

To appear in: *Speech Communication*

Received Date: 1 May 2006
Revised Date: 31 December 2006
Accepted Date: 29 January 2007

Please cite this article as: Rozman, R., Kodek, D.M., Using asymmetric windows in automatic speech recognition, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.01.012

Using asymmetric windows in automatic speech recognition

Robert Rozman*, Dušan M. Kodek
University of Ljubljana, Faculty of Computer and Information Science, Laboratory for Architecture and Signal Processing, Tržaška 25, 1001 Ljubljana, Slovenia

E-mail addresses: rozman@fri.uni-lj.si (R.Rozman), duke@fri.uni-lj.si (D.M. Kodek)

** Corresponding author:*
E-mail address: rozman@fri.uni-lj.si (R. Rozman), Tel.: +386 1 476 8374, Fax: +386 1 426 4647

Keywords: Asymmetric windows; Windowing; Robustness; Automatic speech recognition; Short Time Fourier Transform

1. Introduction

The fundamental problem of automatic speech recognition (ASR) is the variability of speech signals. Each written word has several possible spoken variants. In addition, the speech signal is often distorted which results in a reduced success rate of speech recognition systems (SRS). Undesired influence of distortions is addressed in different ways. The common procedure is inclusion of expected conditions in a training phase. In practice this is difficult to do because of the diversity of audio devices, channels and acoustical environments present in real spoken communications. It is therefore inevitable for SRS to meet unforeseen conditions. In such situations it is very important that they maintain their success rate as much as possible.

Symmetric windows are widely used in the field of digital signal processing due to their ease of design and linear phase property. But the latter also implies potential drawbacks like longer time delay and frequency response limitations. Removal of the symmetry constraint can therefore give asymmetric windows having some better properties. In speech recognition this can lead to a more robust signal representation and hence better recognition performance. A shorter time delay on signal processing can also be achieved. This property is gaining importance in contemporary spoken communications – particularly in Voice Over Internet Protocol (VOIP) related applications.

Human listeners perform substantially better than SRS in the presence of distortions. This suggests that properties of human hearing should be taken into consideration when SRS are designed. Although it can be argued that a "blind" replication of human properties cannot consistently enhance automatic recognition performance (Hermansky, 1997), some of them may be worth trying. For example, human speech perception is quite insensitive to short-time phase distortions of the speech signal. This may be attributed to the fact that the ear hair cells have an asymmetric impulse response and this fact is disregarded when symmetric windows are used. The same conclusion can also be reached from the purely signal processing point of view. We believe that the use of asymmetric windows could help in bridging the gap between human and ASR performance. Since the window function can be easily substituted in a SRS without any additional space or time complexity, the positive influence of such an act on recognition performance is of great interest.

It is perhaps surprising that so little attention has been given in the literature to the problem of asymmetric window design. Even more, little is known about the influence of window properties on the performance of SRS in general. However, the popularity of asymmetric windows in speech coding (ITU, 1996) suggests that their advantages could be applied to practical systems. Our initial research on the application of non-standard windows to speech recognition (Rozman and Kodek, 2003) confirmed a noticeable increase in overall recognition robustness. These facts contributed to motivation for further research with purpose of enhancing the knowledge about window influence on the performance of SRS.

The paper is organized as follows. Section 2 gives an overview of important human audio perception aspects in relation to windowing in a typical frequency analysis procedure used in SRS. Potential advantages of asymmetric windows for speech recognition are stressed. Also, different criteria and desired properties that could enhance SRS robustness are discussed. Several methods with practical

design examples are presented. In Section 3, practical evaluations are described. A reference testing environment with two real speech recognition systems, two speech databases, and a variety of simulated distortions is introduced. Results of practical recognition tasks are shown in several tables, focusing on speech recognition's performance in adverse conditions not present in the training phase; we denote this property as "inherent robustness". In Section 4 important conclusions are drawn and directions for further research are given.

2. Windows in speech recognition

The short time Fourier Transform (STFT) is a common frequency analysis method in speech recognition. The signal is divided into short frames of N samples as shown in Fig. 1. Final windowed values $x(n)$ in each frame are obtained by multiplying signal $s(n)$ with a nonzero window sequence $w(n)$

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

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

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

The frequency response $X(e^{j\omega})$ is obviously influenced by $W(e^{j\omega})$. In addition, it is typical for speech recognition that only the magnitude frequency response of signal samples in the frame is kept for further processing. We therefore wish to select a window function $w(n)$ in such a way that the computed magnitude response $|X(e^{j\omega})|$ is as "near as possible" to the real magnitude response $|S(e^{j\omega})|$. For a given N the asymmetric window idea arises naturally here: removing the symmetry constraint can increase the spectral resolution giving a better $|X(e^{j\omega})|$. This, however, does not necessarily translate into better speech recognition performance.

Window functions that satisfy some signal processing optimality criteria are well known in the literature. But when we design the window sequences for speech recognition other aspects are also important. There is no theoretical reason to believe that the best window sequences, which satisfy the signal processing optimality criteria, will also perform optimally in speech recognition. What is needed here is a careful study of the properties of human auditory perception and, based on this study, incorporation of selected window properties into SRS. We tried this approach in our implementations of SRS and several interesting ideas appeared.

One of them is the idea of windows with wider main-lobes in magnitude response. Wideband time-frequency signal representation is usually used in speech recognition as a basis for further computation of FBANK[1] and cepstral (MFCC[2]) feature vectors as shown in Fig.1. In this typical case it is obvious that accurate frequency analysis or use of windows with narrow main-lobes is not needed, at least not for higher frequencies. A similar conclusion can be drawn from knowledge of human speech perception (Fletcher, 1953). The main advantage of a wider main-lobe is that it can lead to lower side-lobes.

2.1 Finding optimal window for speech recognition

As mentioned above, it is not clear what window properties and design criteria are optimal for use in speech recognition. It is nevertheless 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 frequency resolution and heavily overlapped filters with rapidly decaying side-lobes.

[1] *FBANK features represent logarithm of energy in each frequency band.*

[2] *Stands for Mel Frequency Cepstral Coefficients. Computed as DCT transform of FBANK features.*

- Speech recognition systems 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.

Basically there are two major signal representation distortions introduced by the inevitable windowing process: spectral smearing and leakage. Both can be seen in Fig. 2 where a signal consisting of two unequal pure tones is shown. Spectral smearing is important for the discrimination of closely spaced spectral components, while spectral leakage influences the detection of the distant components. It is clearly shown in Fig. 2a that in the case of a window with high side-lobes (Hamming) the first tone will not be sufficiently suppressed and will add to the much smaller second tone. This gives an incorrect spectral estimate of the second tone. Things are different if a window with lower side-lobes[1] (Fig. 2b) is used. The first tone will be suppressed almost completely and will not influence the estimate of the second tone. Since both smearing and leakage cannot be minimized at the same time their importance should be established. Based on our experiments it seems that the distant spectral leakage, or more general side-lobe height, is important for speech robustness; hence it will be given more attention.

From the speech recognition point of view the distant spectral leakage is important because of another practical reason. Most real SRSs that are based on Hidden Markov Models (HMM) 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. As shown in Fig. 3 on the example of 50 randomly selected speakers from the SLO-DIGITS database (Sec. 3.1.2), asymmetric window with low side-lobes[1] decreases the average correlation in feature vectors[2]. Therefore HMM SRSs using FBANK features and diagonal covariance matrices are expected to perform better with asymmetric windows.

Asymmetric windows also bring shorter time delay, which at first does not seem to be of major importance. In general, the speech recognition process is time consuming and the time delay of spectral analysis represents only a small fraction. But recently unified distributed platforms for speech communication, recognition, and synthesis have appeared (Milner, 2006). They merge speech coding, recognition, and synthesis systems and introduce a major novelty – a uniform signal representation. Therefore a fast reconstruction of the time-domain signal is required for "live" spoken communication. The shorter time delay property alone is already successfully used in speech coding in the form of well known asymmetric "ITU Hamming-Cosine window" (ITU, 1996). An example of the time delay effect in frequency analysis can be seen in Fig. 4. A sliding STFT was computed to better show the time difference in both spectrograms. It is clearly shown that asymmetric windows give shorter time delay in spectral analysis and therefore faster signal reconstruction.

These observations 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).

For a given window function the lower side-lobes can only be obtained by widening the main-lobe which, based on the reasons presented, seems a small price to pay. Lower side-lobes that are also rapidly decaying are important because of the spectral leakage distortion. They prevent distant spectral components from affecting the output of a given band. The majority of additive noises in practice are band limited and hence preventing the spreading of noise energy into other spectral bands is important for robustness of recognition. As already stated, shorter time delay is of lesser importance for speech recognition alone.

---

[1] *"Solvopt3_10" - introduced in next section.*

[2] *Main diagonals in the matrix of correlation coefficients were averaged. Row '1' stands for main diagonal with elements (i,i), '2' stands for second diagonal with elements (i+1,i), etc…*

2.2 Window design methods

Most SRS implementations use one of the standard symmetric windows (Hamming, Hann, Blackman). It is a known fact from the field of FIR filter design that symmetry constraint relaxation can lead to better magnitude response. For the initial research in this field we have designed and evaluated asymmetric windows with lower side-lobes but without the rapidly decaying height of side-lobes property. It is important to remember that the inherent robustness of SRS will represent the final criterion. Two groups of windows are described in following subsections.

2.2.1 Standard symmetric windows

Using one of the standard windows gives a fixed relationship between different main-lobe widths and side-lobes heights. Windows in this group are usually defined with a closed form expression and are therefore easily computable. They are also symmetrical (linear phase) and have a particular shape of the magnitude response. But their magnitude responses are generally not in a full conformance with those of "speech recognition friendly" windows (Sec. 2.1). Also, symmetric windows imply constant, but generally longer time delay. In this group the Hamming window is most popular and will be used in further comparisons.

2.2.2 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 $h(n)$ of a FIR digital filter. Given the definition of the desired magnitude response a window function can be computed using methods similar to those used for the design of optimal linear phase FIR digital filters. The difference is that the symmetry constraint is removed which leads to an optimization problem that is significantly more difficult. 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, $\boldsymbol{h}^* = [h^*(0), h^*(1), \ldots, h^*(N\text{-}1)]$*, that has the minimal error according to the minimax (or Chebyshev) criterion*

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

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

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

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

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*[1]*, 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 described by Eqs. (3)-(7) is nonlinear and therefore considerably more difficult to solve than the one with the linear phase constraints on $D(e^{j\omega})$ and $H(e^{j\omega})$ (Parks and McClellan, 1972). In Eq. (6) the phase and magnitude errors contribute equally to the final error value. This leads 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 Eq. (6) is replaced by the magnitude-only error function

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

[1] $\boldsymbol{\Omega}$ *is a union of compact, non overlapping subintervals of* $[0 .. \pi]$.

in which the phase error is completely ignored. Examples of the corresponding solutions for linear phase, nearly linear phase, and arbitrary phase windows are given in Figs. 5, 6, and 7. These windows were designed using $D(e^{j\omega}) = 1$, $W(e^{j\omega}) = 1$ for $\omega \in [0, 0.012\pi]$ and $D(e^{j\omega}) = 0$, $W(e^{j\omega}) = 1000$ for $\omega \in [0.0425\pi, \pi]$. The examples show how the relaxation of phase linearity constraints leads to a better magnitude response – in the form of lower side-lobes in this case. Note that these windows are equiripple and do not have the rapidly decaying height of side-lobes property.

The modified linear programming method based on the work of Burnside and Parks (1995) was used to solve the nearly linear phase problem given by Eqs. (3)-(7). This type of window was not used in speech recognition experiments described in this paper. The arbitrary phase problem (Eq. (8) replacing Eq. (6)) was solved using the general-purpose optimization procedure "SOLVOPT"[1]. Due to the diversity of possible design cases the use of general optimization methods was most appropriate. It provided a framework for efficient manipulation of different criteria and desired properties.

A drawback of the asymmetric window design is the complexity of the design process that requires a solution of a significantly more difficult minimax approximation problem. This, however, does not increase the complexity of an SRS because the window can be precomputed. It is also difficult to find the optimal design specifications (desired passband and stopband for instance). Since the difference in the side-lobe heights of symmetric and asymmetric windows increases with the main-lobe width, we used the above described $D(e^{j\omega})$ that gives a main-lobe that is approximately 3 times wider than the one of the corresponding Hamming window. Certainly, some additional research in this field is needed, particularly in finding more efficient design methods and optimal design specifications.

3. Practical evaluation

The main motivation for the work presented in this paper is to evaluate the contribution of windows with certain time-frequency properties to speech recognition performance and to its inherent robustness. In this section the windows are analyzed by a practical evaluation in a reference testing environment. This will give empirical evidence of window influence on the performance and on the inherent robustness of SRS, but should be treated with caution. Enhancing the speech signal representation by itself does not help much if further stages in the recognition process (classification stage in this case) are not able to utilize the advantage. This means that the practical evaluations can provide only partial answers. Also, the generality of conclusions is arguable because they depend on specific parameters used in a practical evaluation. However, it seems that this is currently the only possibility and that more definite answers are a matter of further research and evolution in this field.

Two groups of equiripple windows were used in our experiments. The symmetric linear phase Remez3 windows were designed using the Parks-McClellan method. The asymmetric arbitrary phase Solvopt3 windows were designed as described above. The stopband weighting function $W(e^{j\omega})$ was set to 10, 100, and 1000 giving the six windows Remez3_10, Remez3_100, Remez3_1000, Solvopt3_10, Solvopt3_100, and Solvopt3_1000. Note that Figs. 5, 6, and 7 show as examples Remez3_1000 (linear phase) and Solvopt3_1000 (arbitrary phase) windows. The reason for including the symmetric Remez3 windows is to demonstrate that lower side-lobes can improve robustness also for symmetric windows when compared to the standard Hamming window.

3.1 Reference testing environment

The reference testing environment consists of two speech recognition systems, based on different approaches, and of two bilingual speech databases. Both isolated and connected digit recognition were used and a variety of common additive and convolutional distortions were simulated to evaluate the inherent robustness of SRS.

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

3.1.1 Speech Recognition Systems

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".

The first one is based on the statistical approach (Rabiner, 1989). It recognizes one word at a time. Whole-word continuous models are used with 8 internal states, mixtures of 8 Gaussian densities and diagonal covariance matrices. It is implemented using the HTK software package[1]. This approach is currently the most frequently used type of speech recognizer.

The second system uses a neural network in the form of a multi-layer perceptron with 200 internal neurons. It is capable of recognizing whole utterances of concatenated words. Context dependent speech units are used. A simplified form of Viterbi search procedure is used 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[2].

The architectures for both systems were left unchanged (as much as reasonably possible). Utterances were converted into sequences of feature vectors consisting of the "standard" set of normalized[3] MFCC and corresponding delta features. In the HTK recognizer, 12 MFCC features together with the logarithm of the frame energy and the corresponding 13 delta features were used giving a total of 26 features. The CSLU recognizer uses only MFCC features without delta features, although the vector at time *t* is actually a concatenation of 5 vectors at *t-60ms, t-30ms, t, t+30ms, t+60ms.* This sums up to a final 13*5=65 features.

A sampling frequency of 8000Hz was used in both systems. The window of length 32ms was shifted in steps of 10ms across the speech signal. In all cases Word Error Rate (WER) was measured.

3.1.2 Speech databases

All experiments were carried out on two different speech databases: one in English and one in Slovenian.

SLO-DIGITS[4] database (Rozman and Kodek, 2000) was used in both SRS. It consists of 780 Slovenian adult speaker utterances recorded over public telephone lines with their inherent noise. Simple 13-word vocabulary (digits from "0" to "9" and words "yes", "no" and "stop") was used. Each utterance consisted of all 13 words in random order. Speakers were selected on the basis of parameters like age, gender, and location. In practical evaluations 234 speakers were used for training the recognizers. The test and validation sets consist of 156 speakers each. The same sets were used in both SRS. Main characteristics of this database are lower quality of recordings and a variety of different dialects. An isolated digit version and a connected digit version of SLO-DIGITS were created and used in the experiments.

To enhance the variability of evaluation conditions the English "Numbers 95" database[5] was also used with the CSLU 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 to 6 digits per utterance). There were 1368 speakers in the training set, 555 speakers in the validation set and 1168 speakers in the test set.

---

[1] *URL: http://htk.eng.cam.ac.uk/.*

[2] *URL: http://cslu.cse.ogi.edu/toolkit/.*

[3] *Cepstral mean subtraction was performed.*

[4] *In Slovenian this database is named "ŠTEVKE".*

[5] *URL: http://cslu.cse.ogi.edu/corpora/numbers/index.html.*

3.1.3 Inherent robustness evaluation

It should be stressed that both recognizers were trained on the "clean" training set that did not contain any added noise. The different window functions were tried and the inherent robustness was evaluated in terms of a system's performance on noisier, simulated conditions that were not present in the training phase. No additional adaptation was performed prior to testing.

The following seven additive noise recordings derived from the NOISEX database (Varga et al, 1992) 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").

Additive noises were also combined with a convolutional distortion. Lowpass filtering that simulates an acoustic obstacle[1] between speaker and microphone that is common in hands free speech was used.

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.

Tables 1 and 2 give the recognition rates for the CSLU recognizer. A slight degradation in "Clean" conditions is quite usual for robustness enhancement techniques. On the other hand, performance increases can be observed in the additive noise groups for asymmetric windows and an even greater improvement in the case of additional lowpass distortion. Similar conclusions can be drawn from the results with the HTK recognizer in Table 3. The only difference is a slight degradation in additive noise test groups. To see if this was caused by the HTK recognizer in combination with the SLO-DIGITS database an additional set of tests using RASTA filtered MFCC + delta features (Hermansky and Morgan, 1994) was done. Performance is consistently better (Table 4) which means that the observed degradation is not related to HTK or SLO-DIGITS. It can also be concluded that some feature sets (in our case RASTA features) better utilize different signal representations than others.

If we take a closer look at the performance on individual additive noise distortions in Table 5, we can see that a significant degradation occurs in the case of "artificial" additive band limited white noise ("Pass 900"). It seems that in this case the main-lobe width plays a more important role than initially expected. Since the noise is limited around 900 Hz, a wider main lobe causes its spectral smearing into near spectral bands that are important for recognition. This is further confirmed by the fact that the same effect does not happen in two similar wideband distortions: White and Pink noise.

Generally speaking, the robustness improvements on additive and lowpass distortions are well beyond our initial expectations. They confirm our initial assumption that the leakage reduction reduces the effect of noise without adversely affecting the clean recognition.

Another conclusion follows from the results. For both Remez3 and Solvopt3 windows the lower side-lobes almost consistently result in better robustness. This confirms that the height of side-lobes is indeed a very important property.

---

[1] *4th order lowpass digital Butterworth filter was used with f_c= 800Hz.*

It is also interesting to point out that the Hamming window has a narrower main-lobe in comparison to other windows. There are conditions where this seems to be important (Tables 3 and 5) despite the great difference in side-lobes height.

4. Conclusion

Our results show that a considerable increase of the inherent robustness can be obtained with the non-standard window functions. It should be stressed that replacing a window function is a simple procedure that does not increase the time or space complexity of a recognizer. We are currently performing tests in which the window is changed after the learning process. Results (as yet unpublished) show improvements that are comparable to those described above. This fact is in our opinion very important because it shows that an improved robustness can be obtained by simply applying a different window to an existing working SRS without any additional training. Not a small achievement for such a simple modification.

It is also reasonable to believe that in the future the speech recognizers will use better signal representation even more efficiently. Since VOIP systems are emerging fast the shorter time delay advantage will also gain in importance.

References

Burnside, D., Parks, T. W., 1995. Optimal design of FIR filters with the complex Chebyshev error criteria. IEEE Trans. on Signal Process. 3 (43), 605-616.

Fletcher, H., 1953. Speech and Hearing in Communication. Krieger, New York.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 4 (2), 578-589.

Hermansky, H., 1997. Should Recognizers Have Ears? In: Proc. ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, France, pp. 1-10.

ITU - International Telecommunications Union, 1996. Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP). ITU-T Recommendation G.729.

Milner, B., Shao, X., 2006. Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end. Speech Communication 6 (48), 697-715.

Parks, T.W., McClellan, J.H., 1972. A program for the design of linear phase finite impulse response digital filters. IEEE Trans. on Audio and Electroacoust. 3 (20), 195-199.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE 2 (77), 257-286.

Rozman, R., Kodek, D.M., 2000. Speech data base ŠTEVKE and robustness of speech recognition systems. In: Proc. of the conference Language Technologies, Ljubljana, Slovenia, October 2000. pp. 75-78.

Rozman, R., Kodek, D.M., 2003. Improving speech recognition robustness using non-standard windows. In: Proc. of Eurocon International conference on Computer as a Tool, Vol. 2, Ljubljana, Slovenia, September, 2003. pp. 171-174.

Varga, A., Steeneken, H.J.M., Tomlinson, M.J., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. CD-ROM available from the Speech Research Unit, DRA Malvern, UK.

Figure Legends

Fig. 1. Typical parameterization process in a speech recognition system.

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.

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

Fig. 4. Time delay effect in the spectrogram of the 3 tone sound using: (a) Hamming window, (b) ITU Hamming-Cosine window.

Fig. 5. Windows designed with FIR methods (N=256).

Fig. 6. Magnitude responses of windows designed with FIR methods (Fs=8000Hz, N=256).

Fig. 7. Group delay in main-lobe of windows designed with FIR methods (Fs=8000Hz, N=256).

Fig 1

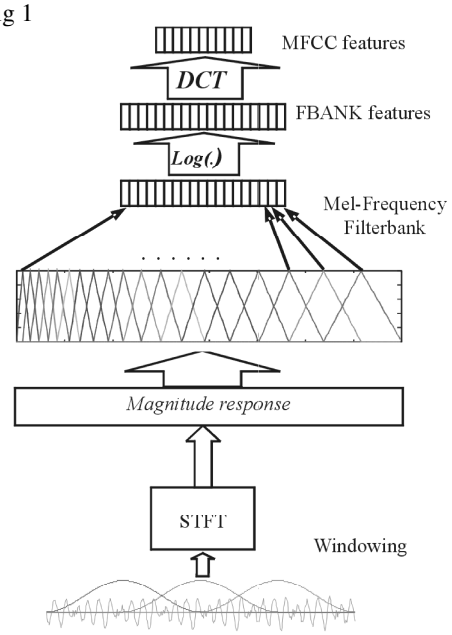


Fig 2

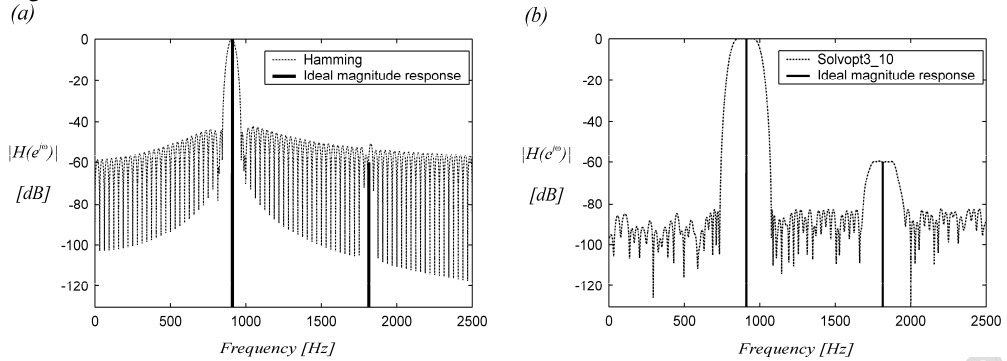


Fig 3

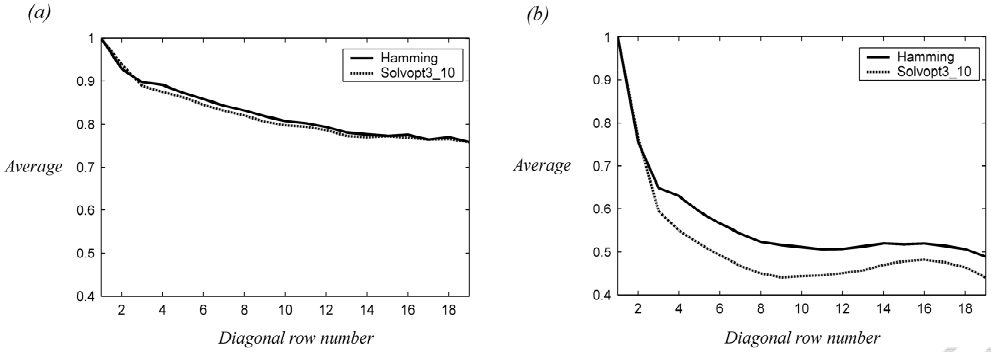


Fig 4

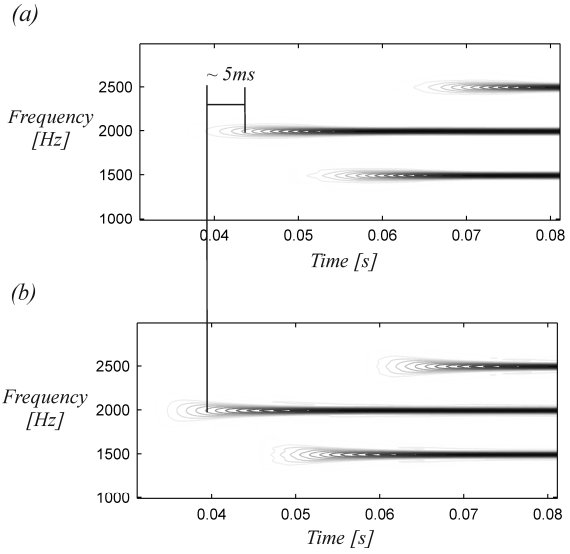


Fig 5

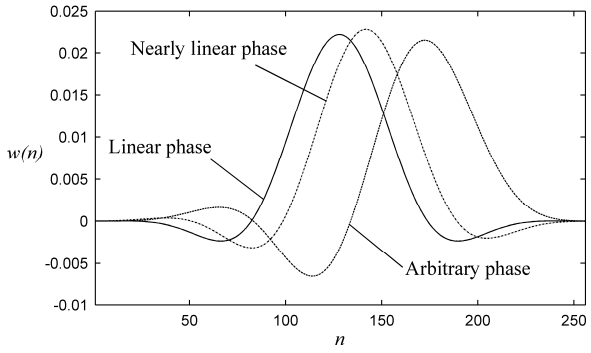


Fig 6

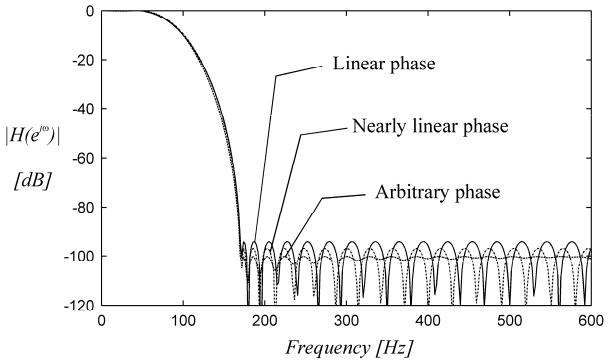
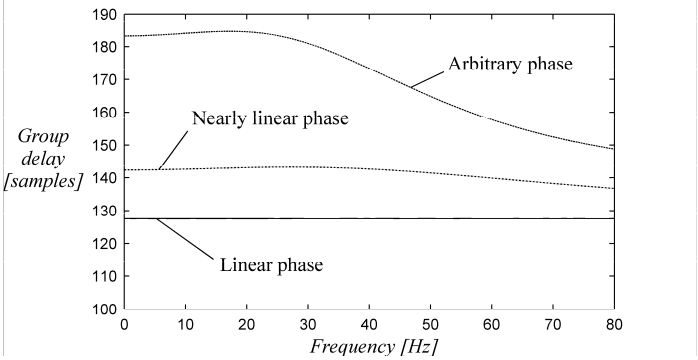


Fig 7



Tables

WER[%]	Clean	Additive [SNR]			Additive+LP [SNR]			Mean
		12 dB	6 dB	0 dB	12 dB	6 dB	0 dB	
Hamming	2,6	16,6	31,2	56,6	35,9	52,2	73,0	38,3
Remez3_10	3,1	16,3	31,4	58,3	26,6	40,9	64,8	34,5
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

Table 1. WER on connected digit task. CSLU SRS on Numbers 95 database was used.

WER[%]	Clean	Additive [SNR]			Additive+LP [SNR]			Mean
		12 dB	6 dB	0 dB	12 dB	6 dB	0 dB	
Hamming	5,9	13,7	26,7	46,7	45,7	58,0	69,4	38,0
Remez3_100	5,7	13,8	26,6	47,7	21,6	32,8	52,2	28,6
Solvopt3_100	5,0	12,3	24,2	44,6	20,9	31,3	49,8	26,9

Table 2. WER on connected digit task. CSLU SRS on SLO-DIGITS (connected) database was used.

WER[%]	Clean	Additive [SNR]			Additive+LP [SNR]			Mean
		12 dB	6 dB	0 dB	12 dB	6 dB	0 dB	
Hamming	4,5	12,9	22,0	37,3	38,2	48,7	61,5	32,2
Remez3_100	5,4	15,4	24,8	39,1	22,8	31,4	44,3	26,2
Solvopt3_100	5,1	14,9	24,2	39,5	23,0	30,8	43,3	25,8

Table 3. WER on isolated digit task. HTK SRS on SLO-DIGITS (isolated) database was used.

WER[%]	Clean	Additive [SNR]			Additive+LP [SNR]			Mean
		12 dB	6 dB	0 dB	12 dB	6 dB	0 dB	
Hamming	4,3	13,8	24,6	44,1	34,3	46,4	64,1	33,1
Remez3_100	4,6	12,9	23,2	42,3	21,6	30,8	50,3	26,5
Solvopt3_100	4,2	12,6	22,5	40,6	21,1	29,9	47,6	25,5

Table 4. WER on isolated digit task. HTK SRS with RASTA+delta features on SLO-DIGITS (isolated) database was used.

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
Remez3_100	10,5	11,1	13,9	13,8	14,1	10,8	22,2	13,8
Solvopt3_100	9,6	9,5	12,0	12,6	11,6	9,7	21,2	12,3

Table 5. WER on connected digit task and different noise types (SNR=12dB). CSLU SRS on SLO-DIGITS (connected) database was used.

Graphical abstract

Using asymmetric windows in automatic speech recognition

Robert Rozman*, Dušan M. Kodek
University of Ljubljana, Faculty of Computer and Information Science, Laboratory for Architecture and Signal Processing, Tržaška 25, 1001 Ljubljana, Slovenia

E-mail addresses: rozman@fri.uni-lj.si (R.Rozman), duke@fri.uni-lj.si (D.M. Kodek)

** Corresponding author:*
E-mail address: rozman@fri.uni-lj.si (R. Rozman), Tel.: +386 1 4768374, Fax: +386 1 426 46 47

Abstract:
This paper considers the windowing problem of the short-time frequency analysis that is used in speech recognition systems (SRS). Since human hearing is relatively insensitive to short-time phase distortion of the speech signal there is no apparent reason for the use of symmetric windows which give a linear phase response. Furthermore, phase information is usually completely disregarded in SRS. This should be contrasted with the well-known fact that relaxation of the linearity constraint on window phase results in a better magnitude response and shorter time delay. These observations form a strong argument in favor of the research presented in this paper. First, a general overview of the role that windows play in the frequency analysis stage of SRS is presented. Important properties for speech recognition are highlighted and potential advantages of asymmetric windows are presented. Among them the shorter time delay and the better magnitude response are most important. Two possible design methods for asymmetric windows are discussed. Since little is known about window influence on SRS performance the design methods are first considered from a frequency analysis point of view. This is followed by practical evaluations on real SRS. Expectations were confirmed by the results. The proposed asymmetric windows increased the robustness of elementary, isolated and connected speech recognition on a variety of adverse test conditions. This is particularly true for the case of a combination of additive and low pass convolutional distortions. Further research on asymmetric windows and on the parameterization process as a whole is suggested.