Speech *F*0 estimation with enhanced voiced-unvoiced classification

TAMÁS BÁRDI

Department of Information Technology, Péter Pázmány Catholic University
bardi.tamas@itk.ppke.hu

Keywords: voicing determination algorithms, low-pass filter, basic parameter extraction

Pitch detectors for speech signal can only work correctly if the fundamental frequency estimation is linked with a reliable voiced-unvoiced decision. A pitch detection algorithm is presented with an enhanced voicing detection method, which gives less error rate than concurrent methods. This pitch detector is based on the well-known autocorrelation method with some modification. The robustness of the algorithm on voicing decision was evaluated over a database of speech recorded together with a laryngograph signal.

Although modern pitch perception models state that the subjective pitch of a sound is not always one to one relation with its fundamental frequency (*F*0), in speech signal processing *F*0 estimators are commonly known as *pitch detection algorithms* or PDA, pitch and *F*0 are treated often as synonyms. A reliably estimated pitch contour of a speech waveform can be useful for a wide range of application. Speech *F*0 variations plays important role in prosody analysis such as discriminating statements and questions. Automatic speech recognition in tonal languages such as mandarin Chinese or Vietnamese also needs a good pitch detector.

Many pitch determination methods have been proposed [10] in the literature and the most comprehensive review is that of Hess [7]. Most of them are moderate in performance but there are some outstanding. For example Bagshaw's eSRPD method [3,4] estimates *F*0 with less than 1% gross frequency error where voiced excitation exists in speech. But it detects the presence or absence of a voiced excitation with 3-4% error.

It is common in speech sciences that linguistically meaningful pitch can exist only where voicing exists. Hence the solution of voicing problem is a premise for the solution of the pitch determination problem. Voiced/unvoiced (V/UV) distinction is a must for speech recognition, since there are words differing from each other only in a voiced or unvoiced consonant, for example 'too' and 'do'.

Voicing determination algorithms (VDA) can be realized implicitly as a part of the PDA but also as a standing alone application. Several VDA has been proposed in the literature [7,12], some of them deserve attention unsparing theoretical invention and but mostly with not a persuasive performance. The rate of V/UV errors is usually higher than the *F*0 estimation error rates in PDAs. Atal and Rabiner presented a multi-parameter solution based on pattern recognition approach [5,6,7]. It gives 4% decision error but it solves a stronger task namely the voiced/unvoiced/silent (V/U/S) classification instead of voiced/unvoiced (V/UV) decision.

The present paper introduces an enhanced method for voicing detection built in a PDA. Our algorithm is based on the well-known Autocorrelation Function (ACF). Using our VDA the decision error falls nearly to 2%. Using Fast Fourier Transform to compute ACF our algorithms can be implemented with less than 2 megaflop per second computational cost assuming 8 kHz as sampling frequency.

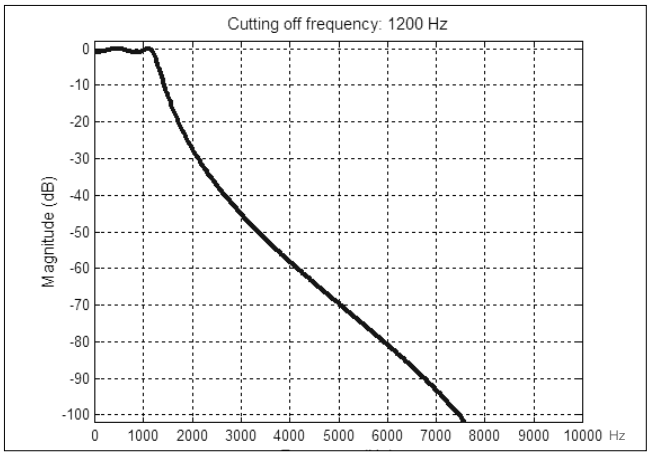
Next sections of this paper follow the modular structure of the algorithm. Section 2 describes our unique preprocessor. It was designed above all to help V/UV distinction, and it plays an important role in achieving the error rate mentioned. After preprocessing typically 30-50 ms long windows of speech are sent to the basic extractor, which is described in section 3. This part computes the ACF and extracts parameters for V/UV decision and *F*0 estimation from it. We use there a special trick namely the "skeletonization" to reduce '*F*0 on the upper limit' type estimation errors.

Our very simple but efficient built-in VDA is in section 4. V/UV decision is based on two parameters, they are compared with thresholds. This two threshold method is essential attaching that good decision error rate. In the literature generally PDAs involve postprocessor which smoothes pitch contours. We do not apply postprocessor now, because this paper focuses only to the reliability of the voicing determination.

1. Preprocessing of speech signal

The usual realization of a PDA is subdivided into three main building blocks: 1) preprocessor, 2) basic extractor, and 3) postprocessor. The main task of the preprocessor is increasing the ease of pitch extraction or voicing determination.

The basic extractor normally works on 20-50 ms long windows of the speech signal. But distinguishing between the steps of preprocessing and basic extraction has just formal importance very often. When windowing



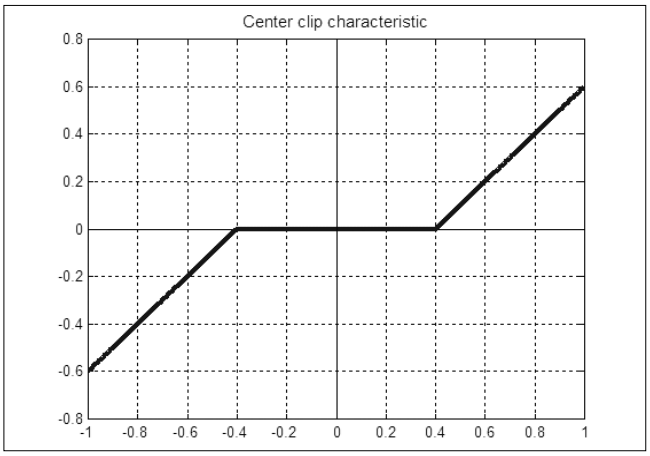


Fig. 1. Characteristics of low-pass filter and center clip applied in preprocessor

step precedes the preprocessor in the execution order of the algorithm we can not examine their work really separately and a lot of computations are duplicated if the windows are overlapped. Windowing before preprocessing makes impossible to listen by ear the output of the preprocessor connectedly. In contrast with that we suggest running the preprocessor on the complete speech signal, after then taking out windows form the output signal and sending them to the basic extractor. In this case we can make sensible an inner state of the algorithm. Creating sensual checkpoints inside a complicated speech processing system can help to optimize its parameters empirically. Our preprocessor is partly "optimized" by ear: fine tuning it we adjusted some parameters until we felt that the output sounds good.

In our preprocessor we use low-pass filter and center clipping. Those are both common in the literature of PDAs [6,9,11]. The characteristics of low-pass filter (Chebishev I type) and center clip used in our method are shown in *Fig. 1.*

The technique of adaptive center clipping applies time-varying clipping level which is adjusted according to the signal amplitude. Generally the varying clipping level is a fixed percentage of an envelope of the speech signal computed some way. The original innovation in our method that it combines the two step: the amplitude envelope is computed from the original speech signal and the low-pass filtered signal is center-clipped with 40% of the envelope. This method removes almost everywhere the speech segments with clearly stochastic excitation such as voiceless consonants. The output signal becomes zero where the low frequency component of the input signal represents the rate of total energy not high enough.

Fig. 2, 3 and *4* show the work of our preprocessor.

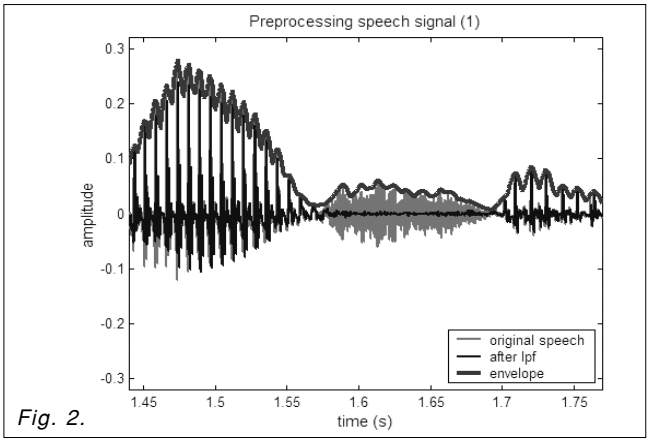


Fig. 2.

The original speech with its envelope and the low-pass filtered signal.

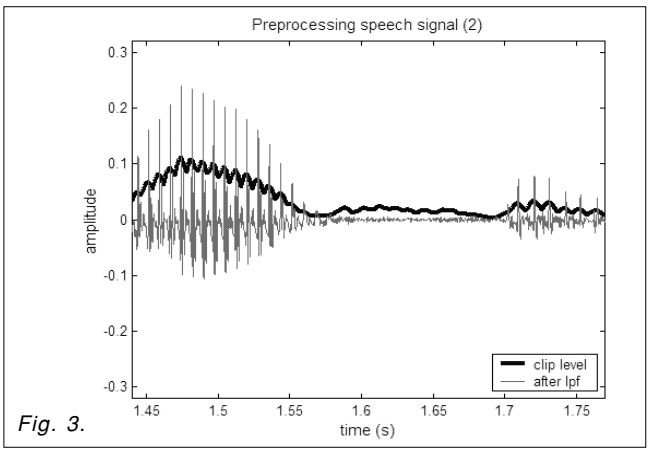


Fig. 3.

Low-pass filtered speech and the computed center clip level.

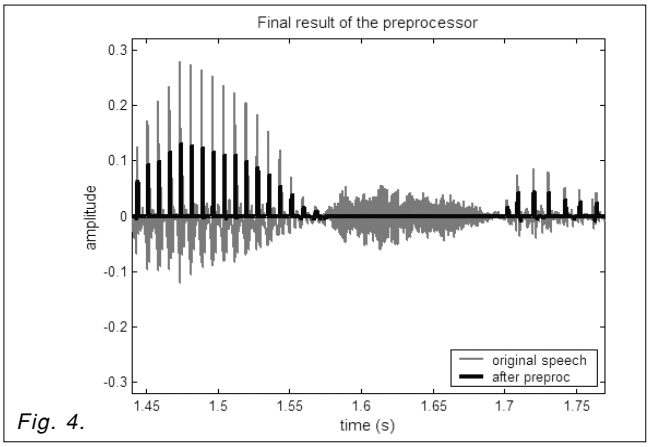


Fig. 4.

Speech signal before and after preprocessing.

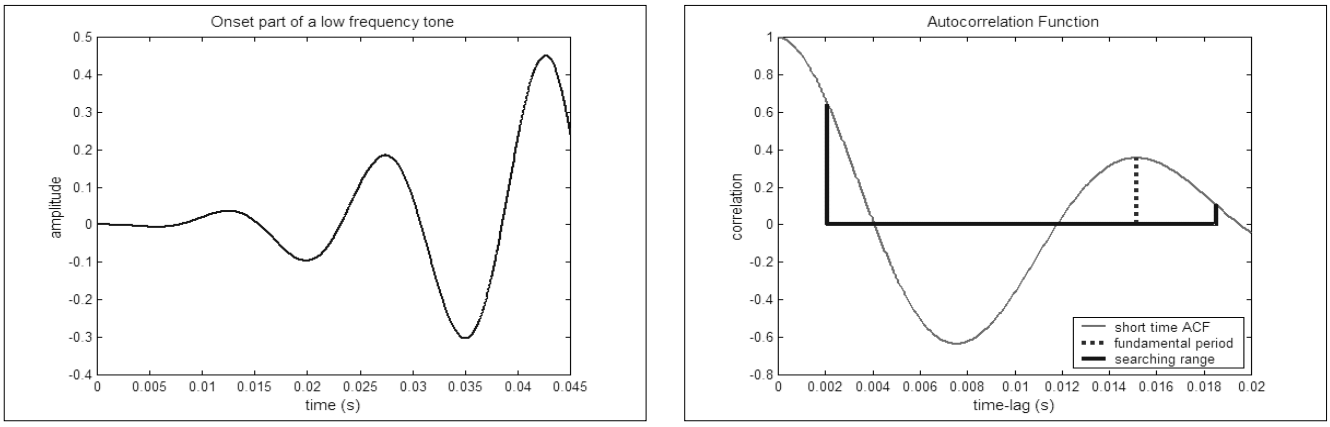


Fig. 5. Onset part of a low frequency tone (67 Hz) and its autocorrelation. The value of ACF at the fundamental period is lower than at the limit of the searching range.

2. Basic parameter extraction

This part of our PDA computes the Autocorrelation Function of the actual signal window and then the algorithm searches for the "best" local maximum of the ACF. The value of the selected peak serves as the main voicing decision parameter and its time lag is the estimation of the fundamental period. But how could we find the "best" peak? As you can see below, the "best" maximum is not so far definitely the global one.

First of all note that all in our formulas the time related symbols (τ, t, u, W) are meant in seconds and signals are meant continuous in time and amplitude hence we use integrals instead of sums. Signal amplitude is meant as the rate of maximal amplitude that can be processed in the system, so that $-1.0 \le x(t) \le 1.0$. These notations make our discussion independent of sampling frequency and bit-rate. Our integral type formulas can easily be converted to sums for concrete applications when sampling frequency and bit-rate are known.

Instead of the biased definition of ACF, which is common in signal processing, we use the unbiased definition, and we apply artificial biasing on it. (W denotes the window length; we set it to 32 ms for this investigation.)

$$r_t(\tau) = \frac{\int_{t-W/2}^{t+W/2} x(u)x(u-\tau)\,du}{\int_{t-W/2}^{t+W/2} x(u)^2\,du} \quad (\tau, t, u, W \text{ in sec}) \quad (1)$$

and the artificial biasing (its degree can be tuned through the gr coefficient):

$$r_t^{biased}(\tau) = r_t(\tau) \cdot (1 - gr \cdot \tau) \quad (2)$$

Computing ACF on the biased way it shows shrinking with increasing values of τ, which gives gain for the fundamental period against its multiples. Although this shrinking can be useful and attractive, its rate can only be tuned by adjusting W. De Cheveigné suggests [5] computing ACF on the unbiased way using fixed window length for all τ time lags and after then applying artificial bias on it. That enables us to adjust the rate of shrinking and window length independently.

For the onset part of a low frequency voice the maximum of the ACF frequently occur at the limit of the searching range. This phenomenon causes the "*F*0 on the upper limit" type errors. That can be seen in *Fig. 5.*

To avoid this sort of error we suggest using the skeleton function or "fishbone" method in other words. The skeleton of a function takes the values of the original function at its local maxima or minima and takes zero otherwise. For our purposes the most suitable definition of local extreme is an intermediate level one between the strict and non-strict version. *Fig. 6.* shows how we do mean local extreme.

Definition: $f: \mathbf{R} \to \mathbf{R}$ real function has local extreme at x, if f not stricly monoton and not plain at x.

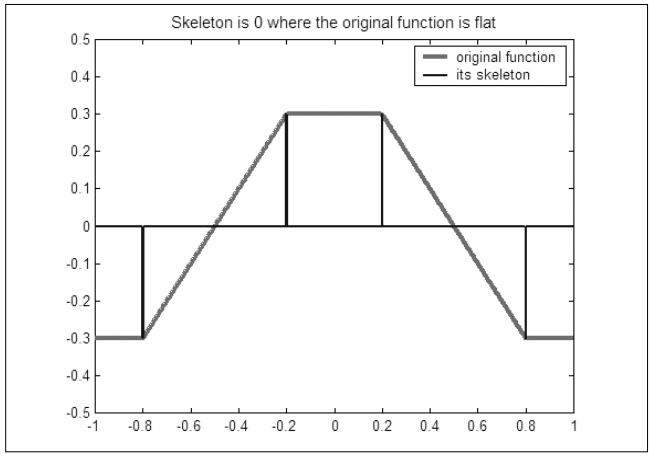
Definition: $g = skeleton(f)$ if and only if

$$g(x) = \begin{cases} f(x) & \text{if } f \text{ has local extreme at } x \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Despite the artificial bias, for the release part of a clear voiced sound ACF tends to have higher peaks with increasing time lags as it can be seen in *Fig. 7.*

This symptom occurs only if the ACF is almost one or even greater than one at the fundamental period. We suggest applying a so called preference level to avoid this problem. Then our algorithm picks the first

Fig. 6. Skeleton function takes 0 where its original is plain.



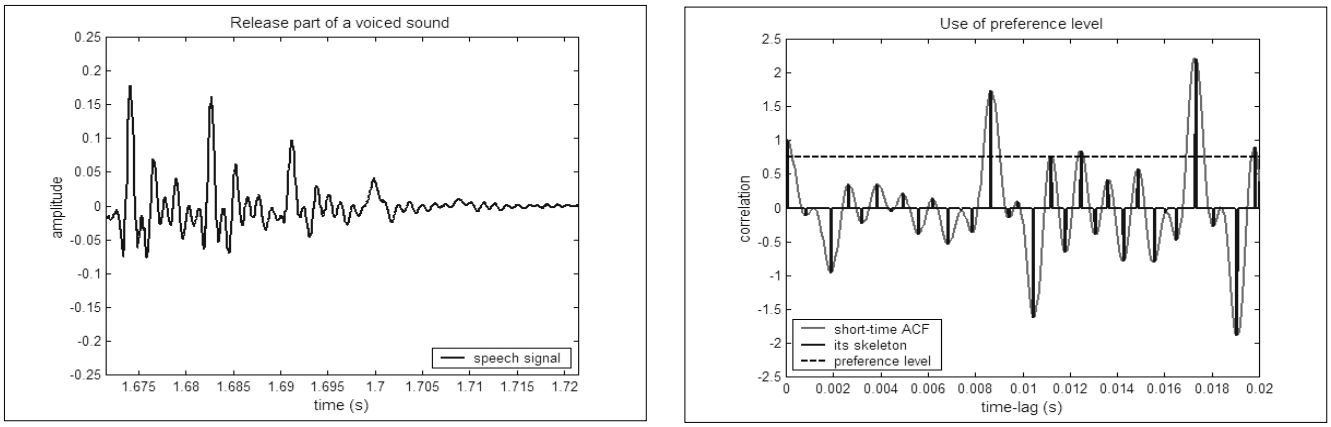


Fig. 7. Release part of a voiced sound and its autocorrelation.

peak that exceeds the preference level. If there is no peak exceeds the preference level the highest peak is chosen. We used 0.75 as preference level chosen it empirically.

Summarizing our basic parameter extractor now we list the algorithm's steps in the correct order:

Step 1: Compute unbiased ACF as in Eq (1).

Step 2: Skeletonization: $sr_t(\tau) = skeleton(r_t(\tau))$.

Step 3: Constrain the *F*0 searching range:

Let $[F0_{min};F0_{max}]$ the searching interval,

$$srl_t(\tau) = \begin{cases} -0.5 & \text{if } \tau < 1/F0_{max} \\ sr_t(\tau) & \text{if } 1/F0_{max} \le \tau \le 1/F0_{min} \\ -0.5 & \text{if } \tau > 1/F0_{min} \end{cases} \quad (4)$$

Step 4: Bias the skeleton:

$srl_t^{biased}(\tau) = (1 - gr \cdot \tau) \cdot srl_t(\tau)$ with $gr = 1.75$ (5)

Step 5: *F*0 estimation:

Step 5/A: Applying the preference level:

$$\tau^* = \min\{\tau : srl_t^{biased}(\tau) \ge 0.75\} \quad (6)$$

Step 5/B:

If *5/A* did not succeed choose the highest peak:

$$\tau^* = \arg\max\{srl_t^{biased}(\tau)\} \quad (7)$$

and the estimated fundamental frequency:

$$F0^* = 1/\tau^*. \quad (8)$$

Step 6: Get voicing decision parameter:

$rm_t = srl_t(\tau^*)$ from the unbiased skeleton (9)

Fig. 8. shows an example for the algorithms work.

3. Voiced-unvoiced decision

Our VDA uses rm_t (9) as decision parameter and the logarithm of the signal energy on the analyzed window:

$$p_t = 10 \cdot \log_{10}\left(\frac{1}{W}\int_{t-W/2}^{t+W/2} x(u)^2\,du\right) \quad \text{(dB)} \quad (10)$$

Consequently from the definition:

p_t=0 dB for a full-scaled square wave.

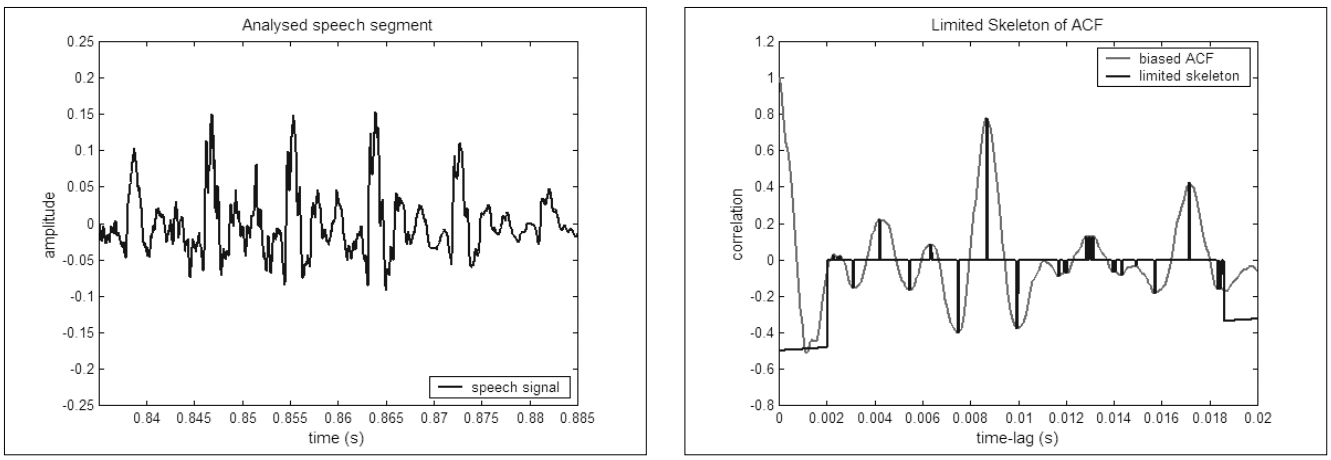
Parameters are compared with threshold so the voicing indicator function is:

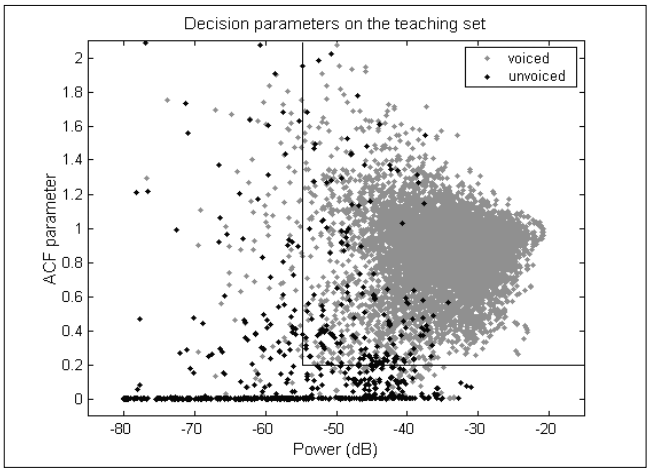
$$voicing(t) = \begin{cases} 1 & \text{if } (rm_t > rmth) \,\&\, (p_t > pth) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Where *pth* and *rmth* are the thresholds.

And now the only question is where to put these thresholds. Tuning procedure of thresholds is linked with the evaluation of voicing decision error rate. We divided the evaluation speech database into 2 parts: 1st half is the teaching set and 2nd half is the control set. The teaching set is for optimizing thresholds on it and the control set is for evaluating our VDA with the optimized thresholds. This evaluation method is correct only if the control set is disjoint from the teaching set. Partitioning both male and female speech into the teaching and control sets provides the maximum speaker independency of the optimization.

Fig. 8. The global maximum of srl(?) shows the fundamental period of the speech window.





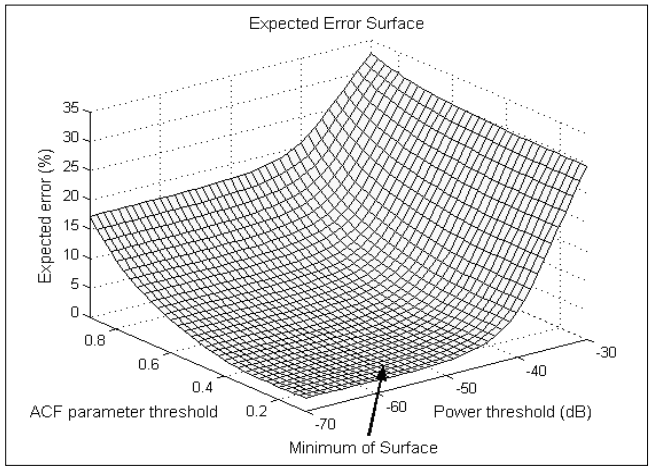


Fig. 9/a. Distribution of decision parameters. Fig. 9/b. Expected error surface.

Voicing decision parameters were extracted using *W*=32 ms window length and *F*0 searching range was between 55 and 480 Hz. *Fig. 9/a.* shows their distribution on the teaching set. Light points come from the voiced segments and dark points come from voiceless segments. The two perpendicular lines depict the two threshold classification method. As it can be seen they do not separate the voiced and voiceless sets perfectly.

Expected Error Surface can be derived from the distribution as a function of threshold pairs. The value of the surface at *(x,y)* represents the voicing decision error on the teaching set itself choosing *(x,y)* as thresholds. Minima of surfaces represent the optimal threshold. *Fig. 9/b.* shows the surface.

Optimized thresholds are: *pth* = –55.2 *dB* and *rmth* = 0.23. The value of error surface at that point is 1.95%. Applying these thresholds on the control set the error rate is *2.13%.* This error rate is the tested performance of our algorithm.

4. Summary

Surveying our algorithms we think that three original trick help us to achieve the 2.13% error rate. First is the combination of low-pass filter and center clip in preprocessor, the second is using skeleton in the basic extractor and the third is considering signal energy in voicing determination. The signal energy indicates voicing much more significantly after preprocessing than before. Precise formulation and correct execution order are also essential.

5. Evaluation database

Our algorithms were tested on the Fundamental Frequency Determination Algorithm (FDA) Evaluation Database recorded at University of Edinburgh, Centre for Speech Technology Research and authored by Paul Bagshaw.

This database is available via ftp from the URL: http://www.cstr.ed.ac.uk/~pcb/fda-eval.tar.gz

It contains 0.12 h speech, 50 English sentences each spoken by one male and one female speakers. 37% out of the total time are voiced segments and 63% are voiceless (silent and unvoiced consonants together). Synchronously with speech signal laryngograph signal was also recorded, which was the basis of labeling voiced and unvoiced segments.

Acknowledgements

The author would like to thank to Dr. György Takács for his instruction and help, to the leaders of the Doctoral School of Department of Information Technology, Péter Pázmány Catholic University for their support and trust, and to Dr. György Lajtha for his help.

References

[1] B. S. Atal and L. R. Rabiner
"A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition"
IEEE Trans. Acoust., Speech, Signal Processing, - Vol. ASSP-24, pp.201–212, 1976.

[2] B. S. Atal and L. R. Rabiner
"Voiced-unvoice decision without pitch detection"
J Acoust. Soc. Am., Vol.58., 1975.

[3] P. C. Bagshaw Automatic prosodic analysis for computer aided pronunciation teaching PhD Thesis, Univ. Edinburgh, 1994.

[4] P. C. Bagshaw, S. M. Hiller and M. A. Jack
"Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching"
Proc. 3rd European Conf. on Speech Comm. and Technology, Vol.2., pp.1003–1006, Berlin, 1993.

[5] A. de Cheveigné and H. Kawahara
"YIN, a fundamental frequency estimator for speech and music"
J Acoust. Soc. Am., Vol.111., Apr 2002.

[6] J. R. Deller, J. H. L. Hansen and J. G. Proakis
Discrete-Time Processing of Speech Signals, Macmillan, New York, 1993.

[7] W. A. Hess:
Pitch Determination of Speech Signals,
Berlin, Springer-Verlag, 1983.
[8] L. R. Rabiner:
"Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone quality speech"
Bell Syst. Tech. Journal, Vol.56, pp.455–482, 1977.
[9] L. R. Rabiner:
"On the Use of Autocorrelation Analysis for Pitch Detection" IEEE Trans. Acoust., Speech, Signal Processing,
Vol. ASSP-25, pp.24–33, 1977.
[10] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal:
"A Comparative Performance Study of Several Pitch Detection Algorithms"
IEEE Trans. Acoust., Speech, Signal Processing,
Vol. ASSP-24, pp.399–418, 1976.
[11] L. R. Rabiner and R. W. Schafer:
Digital Processing of Speech Signals,
Prentice Hall, Engelwood Cliffs NJ, 1978.
[12] L. S. Smith:
"A Neurally Motivated Technique for Voicing Decision and F0 Estimation for Speech"
Centre for Cognitive and Computational Neuroscience, Tech. Report, Vol. CCCN-22,
University Stirling, Scotland, 1996.

News

At Supercomm, **ECI Telecom** will display the latest enhancements to ist XDM platform that make it the first multi-service provisioning platform (MSPP) to fully integrate CWDM and DWDM. Fully integrating CWDM, LDWDM, Ethernet and SONET onto the same platform provides carriers and service providers with a solution that seamlessly connects and manages an optical network from the metro edge to the regional core with simplified operations, reduced costs and ent-to-end performance monitoring. The XDM converged platform with end-to-end multi-layer management allows for seamless traffic connectivity from subtending CWDM or SONET edge rings to metro core DWDM rings with performance monitoring for all services. Additionally, the XDM Build as you Grow architecture enables more freedom to choose the right technology for each service and application.

Veraz Networks will demonstrate on-the-fly services creation, customization, and services management for providers and their customers. The demonstration will highlight the ability to actually create, customize, deploy, and provision services ll without the need for new software releases. This on-the-fly automation reduces the time traditionally required to take services from concept to revenue generation from months or years to hours. The demonstration will introduce Veraz´s built-in service management capabilities. It will demonstrate how providers can create services, group services into service bundles, provision and update service bundles. The service bundles can be made available hierarchically. With Veraz´s solution, service providers can create and customize new services for customers instantly to be able to meet and respond to individual needs faster than ever before.

Mr. Houlin Zhao, Director of ITU´s Telecommunication Standardization Bureau (TSB) commended the Chairman for his leadership, his ability to steer the work of the Assembly to a successful conclusion and for having achieved sound results consensually. „We agreed new tools, resolutions, decisions and guidelines that will make ITU-T more efficient and much stronger." – Zhao told delegates.
The main highlights of the Assembly include:
– A next-generation networks (NGN) focus spanning the work programme of all study groups
– The creation of a new Study Group on NGN
– The adoption of new resolutions on Internet-related issues
(ENUM, spam, internationalized domain names, country code top level domain (ccTLD) names)
– The adoption of a resolution on cybersecurity
– The adoption of measures aimed at enhancing a greater involvement of developing countries in standardization activities
– A group to oversee the sector´s seminar and workshop programme and to monitor the market for new topic areas
– The inclusion of a gender perspective in the work of the ITU-T with the adoption of a resolution on gender mainstreaming
The setting up of 13 Study Groups with their areas of responsibility and the designation of their chairmen and vice-chairmen. WTSA also designated the chairman and vice-chairmen of the telecommunication standardization advisory group (TSAG). A request for a study on the economic effect of call-back and other similar calling practices in developing countries an how they impact on their ability to develop their telecommunication networks and services.