ROBUST PITCH DETECTION IN A NOISY TELEPHONE ENVIRONMENT

Joseph Picone[1], George R. Doddington, Bruce G. Secrest[2]

Texas Instruments Inc.
P.O. Box 225016 MS 238
Dallas, Texas 75266

ABSTRACT

While many readily available pitch tracking algorithms are capable of accurately tracking pitch on studio quality speech data, robust performance in real operational environments is still an elusive goal. In this paper, three pitch detection algorithms are evaluated over a database consisting of speech data collected over a wide range of telephone lines including long distance exchanges. The speech material contained in the database consists of excerpts from typical telephone conversations, collected at the receiving end of a two party exchange.

Subjective and objective evaluations were conducted on three pitch tracking algorithms: an improved version of the Integrated Correlation [1,2] pitch tracker, the Gold-Rabiner [3,4] parallel processing algorithm, and the NSA LPC-10 DYPTRACK Version 43 [5-8] algorithm. A comparative analysis of these algorithms indicates that the integrated correlation pitch tracker provides significantly better performance, mainly due to its ability to make accurate voicing decisions in noisy environments. In addition, intelligibility tests demonstrate that synthetic speech intelligibility is correlated with the degree of accuracy of pitch estimation. This result reinforces our belief that accurate pitch tracking is crucial to the operational acceptance of the speech quality produced by the LPC pitch-excited vocoder.

INTRODUCTION

The problem of pitch detection, like many other problems in speech analysis, remains unsolved mainly because the traditional model of fundamental frequency used in the LPC vocoder is an oversimplified, and frequently inadequate model. Pitch detection, in fact, is more art than science, since pitch detection is a problem which defies a closed-form mathematical solution. A particular pitch detection algorithm is successful not for its ability to produce a pitch estimate for a periodic signal, but rather, for its ability to track pitch through the quasi-periodic sections of speech which exhibit anomalous pitch behavior. Pitch detection can best be summarized as the task of finding periodicity in a non-periodic waveform.

Most pitch tracking algorithms fail because of aperiodic vocal cord vibrations of the speaker, often occurring during the onset of voicing, or at the end of a phrase or sentence during which the vocal effort has been substantially reduced. It is not unusual to see an aperiodic sequence of pulses at the onset of voicing for a word beginning with a vowel. Such anomalies often manifest themselves as irregular pitch periods or as regions where every other pitch pulse varies widely in amplitude, as shown in Fig. 1. Even visual inspection of the waveform can sometimes fail to yield a clear estimate of the fundamental frequency. The goal of a good pitch tracker should be to produce a perceptually acceptable estimate of the fundamental frequency in these aperiodic regions. Though one might not think a pitch excited vocoder could accurately represent such a signal, a pitch track may be constructed which produces an acceptable quality of synthetic speech.

The methodology used to evaluate pitch detector performance presented in this paper focuses on measuring performance in an actual operational environment, the telephone network. Characteristics of this channel include noise introduced both at the microphone and in the acoustic channel, nonlinearities introduced by the microphone, and anomalous pitch behavior frequently observed in conversational speech. A speech database was collected by directly coupling an analog tape recorder to a telephone handset and recording the incoming portion of the two party exchange. Over 4 hours of speech material was excised into a representative database which consists of 60 utterances totaling 3 minutes of speech collected from 38 different speakers. The database consists 39 utterances collected from adult male speakers and 21 utterances collected from adult female speakers. Reference pitch tracks which attempt to optimize the quality of the corresponding LPC synthetic speech were constructed using an interactive pitch editor which utilizes an interactive graphics editor to manipulate the pitch track and a real-time LPC synthesizer for speech playback.

Using this database, both subjective and objective evaluations were conducted for three pitch detectors. The objective evaluations used an objective measure developed by Secrest and Doddington [1,2] which is known to have a high correlation with subjective listening tests [2,8] on studio quality speech data. This measure essentially tabulates voicing errors and pitch frequency errors, weighted in a perceptually meaningful manner. The subjective evaluations, described in Section III, measured intelligibility of the synthetic speech produced by each pitch detector, as well as the intelligibility of the original digitized speech, and the intelligibility of the reference pitch tracks.

In any comparison of pitch detection performance, it is essential to benchmark an algorithm against current state of the art. In this case, we haven chosen a version of the Gold-Rabiner (GR) pitch detector [3,4] used in the Texas Instruments Speech Command System, and the NSA LPC-10 (NSA) Dyptrack algorithm [5-8]. The third algorithm included in the comparison is an enhanced version of the integrated correlation pitch detector (IC) first introduced by Secrest and Doddington [1,2]. This is an LPC residual-based pitch detector which uses a single dynamic programming algorithm to perform the voicing decision and pitch

1. J. Picone is currently with AT&T Bell Laboratories, Indian Hill Main - Room 6C-336, Naperville-Wheaton Road, Naperville, Illinois 60566.

2. B. G. Secrest is currently with Standard Oil Production Company, 2 Lincoln Centre, Dallas, Texas 75240-6251.

1442 CH2396-0/87/0000-1442 $1.00

frequency estimation. An overview of the algorithm is given in the next section, while a more detailed description of the algorithm can be found in [1,2,12]. In this paper, we show that the IC algorithm achieves significantly better performance than either the Gold-Rabiner or the NSA algorithm.

THE INTEGRATED CORRELATION PITCH TRACKER

A block diagram of the IC pitch tracker is shown in Fig. 2. The IC pitch tracker operates on a filtered version of the LPC residual signal. A dynamic programming algorithm finds an optimal pitch contour from a set of pitch period candidates produced from a correlation function computed on the filtered residual signal. There is no separate voicing decision; voicing is integrated into the dynamic programming optimization by augmenting the pitch period candidate vector at each frame with an unvoiced hypothesis. Several key features of the system are described below.

Experimental results indicate that the ability to make accurate voicing decisions diminishes when processing the LPC residual [2], due to the absence of the spectral-slope information. Also, because the spectrum of the residual is flat, it is not uncommon to find a small correlation in the residual signal during voicing intervals. For this reason, the spectral slope of the speech signal is restored by adaptively filtering the speech signal, creating a residual signal which retains the slope cues of voicing found in the original speech signal.

Sibilant areas of speech, though unvoiced, frequently exhibit a strong periodicity in a narrowband range of frequencies above 3 kHz. When recorded through a carbon-button microphone, as in a telephone environment, these sections of the speech display a periodic waveform, and sound something like a whistle. A strategy to combat the strong correlations produced by these signals is the use of the averaging filter prior to correlation computation. The averaging filter significantly reduces the correlation of all periodic signals above 2 kHz, while only slightly reducing the correlation of periodic signals below 2 kHz. The modified correlation computation shown in Fig. 2 reduces the correlation of sibilant type sounds relative to normal voiced speech, and is referred to as sibilant suppression.

A critical issue in the use of any correlation function is the placement of the window over which the correlation function is to be computed. Performance can be significantly improved by shifting the location of window to maximize the correlation values during voicing, as well as avoid the occurrence of a pitch pulse at the edge of the window. We assume that higher energy portions of the speech signal are more correlated than lower energy areas, and use an energy function derived from the energy contour to adjust the position of the window.

Once the correlation function has been computed, pitch candidates are generated by searching the correlation function for peaks. A dynamic programming algorithm is used to find an optimal path through the set of pitch period candidates. The candidate optimal pitch tracks are kept in a circular buffer, and delayed for several frames, a process we call "curing". The length of the circular buffer controls how much delay is used in making a firm decision about the pitch. The pitch track is allowed to cure for several frames, and eventually settles. If the delay is sufficiently large, any changes in the optimal pitch track at the beginning of the circular buffer due to changes in the optimal path at the current frame are probably indicative of an unusually difficult pitch track. Experimental results indicate that a delay of 60 ms is sufficient to achieve asymptotically good results.

A COMPARATIVE EVALUATION

Formal evaluations were conducted on the three pitch trackers described above using the conversational speech database. Some statistics for the speech database are shown in Fig. 3. Fifty three percent of the total speech material was classified as voiced speech. Objective tests were conducted on the three algorithms using the objective measure described in [1,2]. Speech intelligibility tests were also conducted on the synthetic speech produced by these pitch trackers, and correlated with the objective results.

In the objective evaluations, all three pitch trackers were constrained to operate at a 20 ms frame period, a frame period which will typically allow the LPC pitch excited vocoder to operate at a bit rate of 2400 bits/s. The combined results of the objective evaluation are shown in Fig. 4(a). The objective measure corresponds to percent frame errors, that is, the percentage of frames which differ from the corresponding frames in the reference pitch contour, weighted in a perceptually meaningful manner [2].

The objective results are broken down by type of error in Fig. 4(b). The objective measure tabulates errors in three classes. The first class of error, the gross pitch error (GPE), representing a voiced frame classified as voiced, tabulates errors which occur when the reference pitch track and the candidate pitch track differ in pitch frequency. The second class, termed the voiced to unvoiced (V-U) error, represents a voiced frame detected as unvoiced by the pitch tracker. The third class of error, termed unvoiced to voiced (U-V), represents an unvoiced frame classified as voiced. The results in Fig. 4(b) indicate that the IC algorithm does a significantly better job of voicing classification.

Subjective evaluations were also conducted on this same database. A frequently accepted measure of vocoder performance is the intelligibility index, typically measured using the Diagnostic Rhyme Test [9]. However, this test is designed to use a particular database of consonant vowel consonant utterances which do not stress the capabilities of a good pitch detector. Though the objective measure used in this study is known to correlate with subjective speech quality measurements, there remains the question of whether improved speech quality results in higher intelligibility (for instance, at bit rates above 9600 bits/s, enhancements in speech quality do not usually result in higher intelligibility).

The intelligibility test presented in this study was conducted using naive listeners, and attempts to measure conversational speech intelligibility. The subject's task involved listening to a vocoded utterance in the database, and transcribing its contents into the computer. Listeners were allowed to listen to an utterance as many times as desired and randomly edit their responses, with no time restrictions imposed.

In addition to the three pitch trackers mentioned above, two other baseline conditions were evaluated. First, the vocoded speech using the reference pitch tracks, representing the ultimate quality achievable by any pitch tracking algorithm, was evaluated. All vocoders used the same LPC synthesizer and LPC information. The LPC vocoded speech was synthesized using an unquantized LPC 10th order model with coefficients computed at a frame rate of 20 ms. The autocorrelation method of LPC analysis was used with a 30 ms analysis window, a hamming window, a preemphasis of 1.0 for analysis, and a preemphasis of 0.9375 for synthesis. Second, the original digitized speech data was included in the evaluation, from which a baseline intelligibility of the database is derived. Recalling that the speech data consisted of excerpts from conversations, excised such that

utterances generally appear out of context, the listening task proved to be quite challenging.

The transcriptions provided by the listeners were scored for accuracy using an algorithm described in [10,11]. This scoring algorithm optimally aligns the listener's transcription with the known contents of the utterance using phonemic transcriptions, and scores the errors in terms of word substitution, insertion, and deletion errors. The combined intelligibility score is defined as the sum of each of these three classes of errors. Performance can also be tabulated at the phoneme level, and is typically consistent with the word error rates.

An intelligibility test such as this requires that each utterance is presented to each listener exactly once. Since the entire test involved evaluating the five conditions described above over a database of 60 utterances, each subject is presented with 12 utterances from each of the 5 conditions. The order of presentation is randomized such that after every 5 subjects, each utterance from every condition has been presented exactly once. Further, after each group of five listeners, a different randomization is used, so that each subset of five listeners had a different permutation of the test. A total of twenty five listeners were used to generate the results presented in this paper, allowing every utterance for each condition to be presented exactly five times.

The results of the subjective evaluation are shown in Fig. 5. The objective measure ranks the IC pitch tracker first, the NSA algorithm next, and the Gold-Rabiner algorithm last. The speech intelligibility results indicate the same ordering. Observe that the IC pitch tracker produces intelligibility close to that of the reference pitch tracks.

The correlation between the objective scores and the intelligibility scores is 0.99. The low intelligibility score for the original digitized speech is a product of the carbon-button microphone distortions, the lack of contextual information in the speech material, and the low signal to noise ratios of certain portions of the database. The low intelligibility score of the reference pitch track data is an indication of the limitations of the low rate vocoder model, especially when channel distortions are introduced into the LPC analysis. Note that vocoded speech intelligibility is also a function of the performance of the pitch tracker, reinforcing our belief that improved pitch detection is an important factor in the operational acceptance of the pitch excited vocoder.

CONCLUSIONS

Aperiodic vibrations of the vocal cords, typically occurring while the speech effort level is low, still pose the greatest challenge to the pitch extraction algorithm. The majority of pitch errors observed in this database tend to be a result of poor voicing decisions. Given that the voicing boundaries are correctly identified, the dynamic programming algorithm used in the IC pitch extractor will usually produce an acceptable pitch contour. The majority of the errors made by the NSA algorithm and the Gold-Rabiner algorithm can be classified as voiced frames being detected as unvoiced frames.

The integrated correlation pitch detection algorithm has been shown to provide significantly better performance in a realistic operating environment, the telephone channel. Speech intelligibility was shown to be highly correlated with pitch detector performance. While the reference pitch tracks were considered the most intelligible, the IC algorithm's performance approached that of the reference pitch tracks.

ACKNOWLEDGMENTS

The authors are indebted to W. H. Russell of Bolt, Beranek, and Newman, Inc. for providing an implementation of the NSA pitch tracking algorithm. This implementation was identical to the version evaluated in [8], and represents all improvements through April 1, 1984. The authors are also indebted to W. Anderson of Texas Instruments, Inc. for providing an implementation of the Gold-Rabiner algorithm.

REFERENCES

[1] B. G. Secrest and G. R. Doddington, "An Integrated Pitch Tracking Algorithm For Speech Systems," in Proc. 1983 IEEE Int. Conf. Acoust., Speech, Signal Proc., pp. 1352-1355, April 1983.

[2] B. G. Secrest and G. R. Doddington, "Postprocessing Techniques For Voice Pitch Trackers," in Proc. 1982 IEEE Int. Conf. Acoust., Speech, Signal Proc., pp. 172-175, May 1982.

[3] L. R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1975.

[4] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1978.

[5] W. R. Bauer and W. A. Blankinship, "DYPTRACK - A Noise Tolerant Pitch Tracker," NSA Internal Report, Dec. 11, 1973.

[6] T. E. Tremain, J. W. Fussell, R. A. Dean, B. M. Abzug, M. D. Cowing, and P. W. Boudra, Jr., "Implementation of Two Real-Time Narrowband Speech Algorithms," in Proc. 1978 Eascon, Washington D.C., pp. 698-708, September 1978.

[7] T. E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," Speech Technology, pp. 40-49, April 1982.

[8] V. R. Viswanathan and W. H. Russell, "Subjective and Objective Evaluations of Pitch Extractors For LPC and Harmonic Deviation Vocoders," BBN Report No. 5726, Bolt, Beranek, and Newman Inc., Cambridge, MA, July 1984.

[9] W. D. Voiers, "Diagnostic Evaluation Of Speech Intelligibility," Benchmark Papers on Acoustics: Speech Intelligibility and Recognition, ed. by M. E. Hawley, Dowden, Hutchinson, and Ross, Stroudsburg, PA, Vol. 11, pp. 250-275, 1977.

[10] J. Picone, K. M. Marshall, G. R. Doddington, and W. F. Fisher, "Automatic Text Alignment For Speech System Evaluation," IEEE Trans. Acoust., Speech, and Signal Proc., Vol. ASSP-34, No. 4, pp. 780-785.

[11] K. M. Marshall, J. Picone, and W. F. Fisher, "Phonetic String Alignment," Journal of the Acoust. Soc. Of Amer., Vol. 77, No. S1, p. S12, Spring 1985.

[12] J. Picone, G. R. Doddington, and B. G. Secrest, "Evaluation of the Integrated Correlation Pitch Detector In A Noisy Telephone Environment," submitted for publication to Speech Communication, in December 1986.

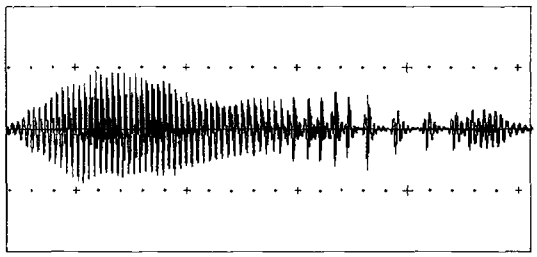


Fig. 1. Waveform of the word "drown" spoken by a female demonstrating anomalous pitch behavior.

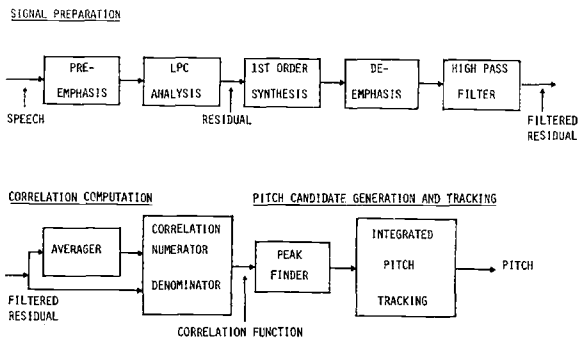


Fig. 2. The Integrated Correlation Pitch Tracker

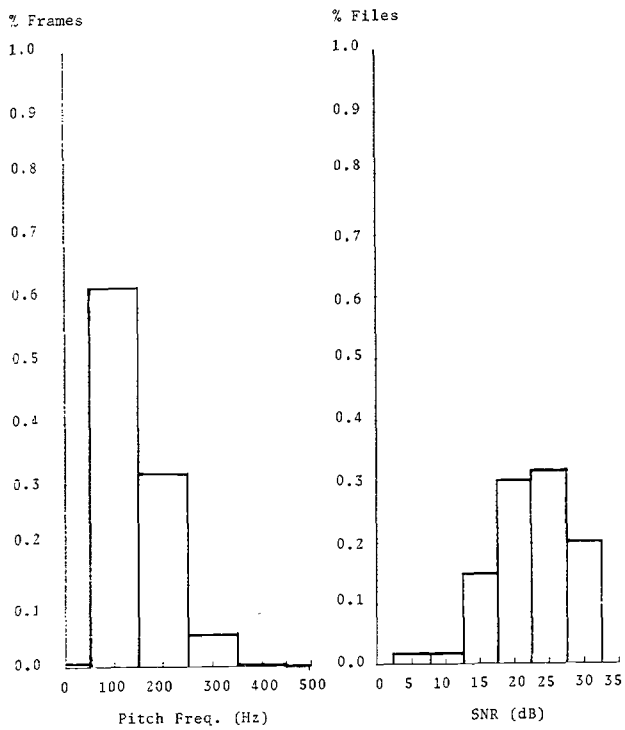
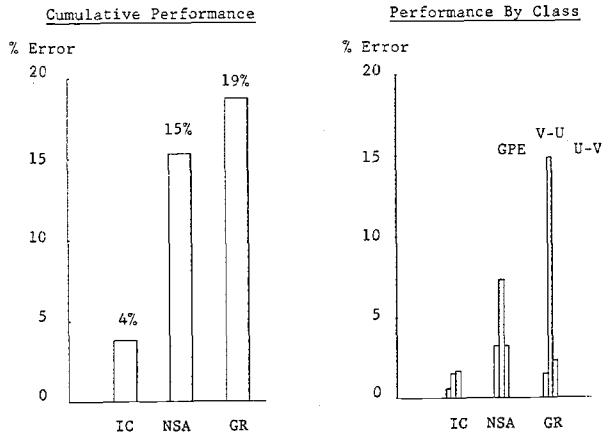


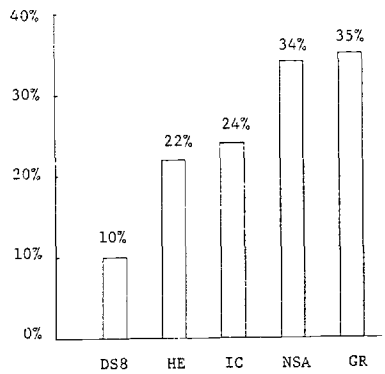
Fig. 3. Pitch and SNR Histograms For The Telephone Database



IC : Integrated Correlation
NSA: NSA LPC-10 Dyptrack
GR : Gold-Rabiner

GPE: Gross Pitch Errors
V-U: Voiced To Unvoiced Errors
U-V: Unvoiced To Voiced Errors

Fig. 4. Objective Performance Evaluation



DS8: 8 kHz Digitized Speech
HE : Hand-Edited Reference
IC : Integrated Correlation
NSA: NSA LPC-10 Dyptrack
GR : Gold-Rabiner

Fig. 5. Subjective Performance Evaluation