

Knowledge-Based Pitch Detection

Webster Dove, Cory Myers
Alan Oppenheim, Randy Davis, Gary Kopec

Massachusetts Institute of Technology
Research Laboratory of Electronics Dept. of Elec. Eng. and Comp. Sci.,
Room 36-615 Cambridge, MA 02139

Knowledge-based signal processing combines the symbolic manipulation facilities of artificial intelligence with the numerical methods of signal processing. This paper considers pitch detection as a problem for developing techniques for signal/symbol interaction. Specifically an overview of the design of the Pitch Detector's Assistant program is discussed, along with some strategies followed in the program, for the use of information derived from both signals and symbols.

Introduction

Knowledge-based signal processing systems attempt to tightly integrate techniques from the disciplines of artificial intelligence (AI) and signal processing[1]. The primary motivation for developing knowledge-based signal processing systems is the anticipated advantage from combining AI capabilities—symbol manipulation and knowledge representation—with the numerical and mathematical tools of signal processing. Of course, many systems already exist that incorporate and exploit principles from both fields, such as the Hearsay-II speech understanding system developed at Carnegie-Mellon University[2], the Surveillance Integration Automation Project (SIAP) system developed at Systems Control, Inc., of Palo Alto, Calif.[3], and various image understanding systems[4]. In general, any system that has a signal as its input and generates symbolic information as its output (or conversely, as in speech synthesis from text) must perform both signal processing and symbol manipulation.

Important limitations result from the way existing hybrid systems combine signal and symbol processing, however. In general, a large problem such as automatic surveillance or speech recognition must be partitioned into a number of smaller subproblems before it becomes tractable. Because separate communities have traditionally been responsible for advances in each field, the signal/symbol distinction is customarily selected as a convenient and well-defined

dimension along which to factor problems. This choice leads to an architecture in which the signal processing is localized in one set of subsystems, the symbol processing in another. Unfortunately, this partitioning has led to systems with only minimal interaction between the signal processing and symbol processing components. To achieve higher levels of performance in hybrid systems of this type, more interaction across the signal/symbol boundary is necessary.

We have recently started a research program aimed at a tighter coupling between symbol processing and signal processing, and the specific project described in this paper is one component of this program. Our choice of specific applications to focus on has been motivated by several considerations. Clearly the problem should be meaningful, in the sense that if good performance is achieved it will be of interest. Furthermore the size of the problem should be chosen conservatively so that both the signal and symbol aspects can be thoroughly explored. Problems on either side of the signal/symbol boundary offer by themselves substantial challenges for research. Since our concern is to generate ideas that straddle the boundary, we need to make the problem small enough that we avoid dividing the work along that boundary. The problem should also have a history in at least one of the two communities so that there will be the sense that signal processing by itself (or symbol processing by itself) has gone as far as it can. This makes it more likely that new approaches will be given serious consideration. In addition it helps ensure that novel levels of performance on the task are indeed due to novel ideas about AI and signal processing interactions, rather than due to insights specific to one or the other.

One example of a problem area consistent with these criteria is pitch detection and in this paper we describe our initial work directed toward the development of a Pitch Detector's Assistant program (PDA), to assist in the task of manual pitch tracking.

There are several reasons why the task of manual pitch tracking is an attractive vehicle for exploring SP-AI collaboration. First, it is a relatively modest prob-

lem in terms of the amount of signal processing, AI and speech expertise required. As a result, it is possible for a collaborator from any one of these areas to be an active contributor to work related to the other two. As noted earlier, size is an important problem selection criterion --- we require a problem small enough that it does not require partitioning in order to be tractable.

Second, the problem of pitch detection is important in speech processing, yet remains only partially solved after considerable effort from the signal processing community. Although most pitch detection algorithms perform well when the signal-to-noise ratio is high, no algorithm works well in the presence of multiple speakers, high background noise or severe channel distortions. There is some consensus that significant improvements in such environments can only be realized by greater sophistication in the use of speech-related knowledge.

A third attraction of the pitch detection problem is that most currently available algorithms already contain both signal processing and AI components. This reflects the fact that the design of any practical pitch detector requires both mathematical formalisms and informal heuristics. Typically, development begins with a mathematical model of periodic waveforms that suggests a way of measuring periodicity. An algorithm based on this model is tested on actual speech, noticed to work well most of the time and then modified to correct observed shortcomings. Unlike the initial formulation, these modifications are not motivated by a mathematical model of speech. They are instead introduced to handle situations encountered in experience with real data, situations not predicted by the mathematical models.

A final aspect of manual pitch tracking is that it involves both perceptual and cognitive skills. On the one hand, visual pattern matching and texture discrimination appear to be involved in detecting periodicity and identifying unvoiced regions. On the other hand, knowledge about the properties of speech waveforms is often used in classifying difficult segments.

Manual pitch tracking takes a long time (perhaps 30 minutes per second of speech) and it requires an understanding of speech to resolve ambiguities. The *Semiautomatic Pitch Detector (SAPD)* for hand editing[5] provided an environment which displayed three different "views" of the data that could be used for determining period; the low-pass filtered time waveform, the autocorrelation of it, and the real cepstrum of the full bandwidth signal. The user was expected to choose the correct pitch on a frame by frame basis (or mark the frame as unvoiced).

The PDA program attempts to improve on this by generating much of the pitch track automatically. As is discussed below in greater detail, the system uses the utterance waveform and a phonetic transcription as

input, with both numeric and symbolic information contributing to the analysis. In addition, the program does not process the utterance in time sequential order, thereby permitting the initial coarse analysis to guide later processing.

The PDA program provides a testbed on which to develop ideas for using symbolic information in pitch detection, not a means for performing optimal pitch detection. Although good pitch detection performance is certainly a desirable behavior, the goal is to learn about the interaction between signal processing and symbol processing. Thus in the near term we have used basic signal processing and symbol processing mechanisms. The absolute pitch detection performance of the system may be compromised somewhat, but more attention can be focused on signal/symbol interaction. This initial effort can then serve to guide more complex implementations.

The PDA Program

The Extended Speech Model: The most common speech excitation model in use today assumes vocal tract excitation to take one of two forms, either a periodic pulse train (for voiced speech) or random noise (for fricated/aspirant speech). This choice is convenient for vocoder design because it simplifies the generation of input to the spectral shaping portion of the vocoder. The PDA program extends the excitation model to include aperiodic voiced and simultaneous voiced and fricated excitation. Although such conditions are rare, they are not absent and are a common source of error in pitch detectors. Specifically, diplophonic speech frequently leads to an unvoiced classification because the periodicity detectors in the system fail and the only alternative to periodic voiced excitation is noise. Since more than just periodicity can determine voicing in the PDA program (the system can claim a segment is voiced without finding a period), the excitation model has in effect been extended. The question of how aperiodic voiced excitation should be treated for synthesis has yet to be examined.

Symbolic Information Sources: Non-numerical information in the PDA program can come from two sources, signal processing and a phonetic transcription. Signal processing of the waveform can lead to statements of the form:

interval	property	value
(10000 11000)	high-frequency energy	large

These represent signal to symbol transformations, and are the primary way signals contribute to the classification of regions. In its initial implementation, the user is expected to provide a phonetic transcription for the

utterance containing statements of the form:

interval	property	value
(1000 1500)	phoneme	/a/

These statements also provide input to the processing modules which determine classification.

The phonetic transcription need not be complete and it is not assumed to be perfectly accurate, (otherwise voicing information could be totally determined from it). It does, however, provide additional information for the interpretation of the utterance and is unusual in the context of signal processing because it is not inherently a numeric signal.

Strategy

One strategy used in the PDA program is initial analysis over regions where the results are obvious. We assume that many regions of the data will submit to simple analysis, and that identifying those regions is itself an easy problem¹. These "Islands of Certainty" then provide a foundation for the more complex analysis necessary in the intervening regions.

To implement this strategy, the PDA program examines several inexpensive transformations of the data together with simple inferences drawn from symbolic sources. Those regions which unambiguously indicate some classification are marked so. Others await later processing. In this way the PDA program gets a reliable start on the problem with modest effort.

A second strategy used in the PDA program is the deferral of decisions wherever more information could cheaply and reliably be collected. Actual pitch period analysis, for example, is deferred until intervals of data are located which have been classified voiced. Simple inexpensive means are used to determine those intervals where a more complex processing should be applied. Two useful results are achieved by this. The program saves computational effort for those situations where it is warranted and the process of evaluating credibility is simplified. This is because the results of simpler processing are more easily understood and easier to interpret.

This point is best expressed by example: consider using the results of several simple pitch detectors as a measure of voicing, compared to making a voicing decision using a low-frequency energy threshold. The Gold-Rabiner pitch detection algorithm [6] uses the first mechanism and complicated processing is necessary to make the decision. Determining whether the low-frequency energy is above a threshold is much simpler. Obviously it will not work as well as the

¹ In the sense that simple processing may be used to solve it.

more complex scheme, but it will work some of the time and it is inexpensive.

Design

Initially, simple processing is used to classify the data. Currently three parameters are calculated: low-frequency and high-frequency energy under a rectangular window, and the first-difference energy also under a rectangular window. These parameters are very inexpensive to calculate, and their relationship to the classifications that we wish to make are obvious.

Locating Obvious Silence: The PDA program initially locates obvious silence, since the system must start somewhere, and silence identification appears to be the least dependent on other analysis.

The PDA program uses both signal processing and symbol processing to locate silence. From a signal processing standpoint, silent regions should have low energy (both high and low frequency) and they should also have low first-difference energy. If we assume that the noise in the utterance is stationary, then the energy balance of all silent regions should be similar as should be the first-difference energy. From a non-numerical standpoint, the start and end of the utterance are very likely to be silent, and certain phonemes (stops, and pause) indicate silence.

Before making signal processing decisions, the lowest values of the parameters over the entire utterance are determined. This is followed by the hypothesis of silent regions based on signal processing and symbol processing as outlined above. Regions may be hypothesized as silent by any of the above parametric indicators or by symbolic inference, but only those which do not violate any of the above conditions are accepted as obviously silent.

Silence detection requires non-causal² decision making unless apriori information about the noise exists. The Gold Rabiner pitch detector uses a fixed amplitude threshold to define silence. This assumption inevitably leads to non-silent regions being classified as silent. The more thorough (and time consuming) approach taken by the PDA program prevents such errors.

Classifying Obviously Voiced Regions: The classification of voiced regions takes place after silence classification for two reasons. First, silence is viewed as simpler and therefore both easier to detect and less likely to be in error. Second, the background noise energy defines the voicing thresholds.

The procedure for voiced classification is similar to that described for silence determination. Regions can be hypothesized as voiced if their low-frequency

² Causal order means purely left to right processing order.

energy is sufficiently large everywhere in a 40 ms interval and their high-frequency energy and first-difference energy are stable over the interval. They can also be hypothesized as voiced if there is a continuent voiced phoneme covering the interval³. The only regions that are actually accepted as obviously voiced are those that do not contradict either criterion. For example, an interval which is hypothesized as voiced by the signal processing procedure, but which is marked as a fricative phoneme will be rejected.

Fricative and Aspirant Regions: These regions are classified as above using both signal processing and symbolic processing techniques. Note that because our model accepts voiced and fricated speech, these regions are not mutually exclusive with those marked voiced. However, for such a region to be accepted the phonetic marking must be consistent with the observed signal properties.

Region Growing: After the above processing is complete, the PDA program has a substantial base of information from which to draw to analyze the remaining regions. For example, suppose there are several neighboring silent regions in the utterance. Perhaps they could be connected into a single region. The following two criteria must be satisfied for this to be possible: first, there must not be any intervening phonetic markings that indicate otherwise; and second, the energy measures must not exceed 4 times the noise energy in the intervening region (the 4x factor is an arbitrary choice).

Pitch Determination: Once no more region classification can be performed, those regions which are found to contain voiced speech are analyzed by periodicity detectors. Several simple pitch detectors are used to hypothesize the period every 10ms in the voiced region. Pitch is only estimated where they all agree.

Having completed preliminary pitch analysis, more complex methods are used on segments where the analysis failed. An interpolated approximation of the pitch is available⁴ from adjacent segments for which the simple analysis worked. This approximation can be used by algorithms which require an a priori estimate of pitch. They can also serve to verify the results for the segment in question.

Symbolic information serves to determine the regions over which pitch detection is performed and when interpolation of pitch is valid. It can also suggest which type of pitch processing is best. For example, nasals are best analyzed using spectral techniques, because their time waveform usually has peaks every half-cycle which cause false triggering of time domain pitch detectors. Finally, symbolic classification of voicing guarantees that a voiced region is marked so even if the pitch detectors are not able to find a period. This can be very important in regions where the quasi-stationarity assumption is not valid.

Conclusions

The program goal of the PDA program is to analyze the utterance where the results are obvious. If the PDA program can analyze large portions of the utterance, the user is able to process more data in the available time. We aim toward a system which can completely analyze most utterances, but much more needs to be done to the PDA program before that is possible.

A variety of extensions are possible at this stage. Arbitrary parameters of the sort presented above can frequently be linked to information about the speech. The system should be able to use numerical and symbolic methods to choose such parameters in an effective way. As presently constituted, PDA does not have to come to conclusions in regions of the utterance where it is unsure. Clearly, for it to fully analyze the utterance, mechanisms for dealing with uncertainty will have to be developed.

³ Voiced stops are excluded from making such hypotheses because they may not be entirely voiced

⁴ Interpolation of pitch is only permitted within a contiguous voiced region.

As a knowledge-based signal processing project, PDA reflects two processing strategies that should be useful whenever a variety of sources of information are available for solving a problem. First, initially only draw firm conclusions where there is no possible error. Those conclusions can provide additional help analyzing the more difficult problems⁵. Second, defer decisions where additional information might easily be gathered. In PDA, the deferral of pitch analysis until voiced regions have been found prevents the improper interpolation of pitch estimates into regions where the utterance might not even be voiced.

Finally, PDA demonstrates that symbolic information can be of value in many stages of this problem and that the combination of signal and symbol processing leads to an effective program even when comparatively simple signal processing techniques are applied. The use of more complex signal processing algorithms in difficult places, guided by the preceding simple analysis, should prove more effective still.

References

- [1] R. Davis, G. Kopec, A. V. Oppenheim, "Artificial Intelligence and Signal Processing: Overview and Experiments", to be published
- [2] L. Erman, R. Hayes-Roth, V. R. Lesser, D. R. Reddy, "The Hearsay-II Speech Understanding System: Integrating Knowledge to Resolve Uncertainty", *Computing Surveys*, vol.12, pp.213-254, June 80
- [3] H. P. Nii, E. A. Feigenbaum, J. J. Anton, A. J. Rockmore, "Signal-to-Symbol Transformation: HASP/SIAP Case Study", *AI Magazine*, vol.3, pp.23-35, Spring 82
- [4] M. Brady, "Computational approaches to Image Understanding", *Computing Surveys*, vol.14, no.1, Mar.82
- [5] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A Semiautomatic Pitch Detector (SAPD)", *IEEE Trans. Acoust., Speech, Signal Processing*, vol.ASSP-23, pp.570-574, Dec.1975.
- [6] B. Gold and L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain", *JASA*, vol. 46, pp.442-448, Aug.1969

This work has been supported in part by an Amoco Foundation Fellowship and in part by the Advanced Research Projects Agency monitored by ONR under contract N00014-81-K-0742 NR-049-506

⁵ This concept is frequently referred to as Islands of Certainty in AI literature. See for example [2].