Proceedings of The Institute of Acoustics

STAR_PAK: A SIGNAL PROCESSING PACKAGE FOR ACOUSTIC-PHONETIC ANALYSIS OF SPEECH

G. Duncan, J. Dalby, M.A. Jack

Centre for Speech Technology Research,
University of Edinburgh.

OVERVIEW OF THE STAR PAK

Star_pak (Signal Transform Analysis for Recognition PAcKage) is a software suite which provides all front end signal processing requirements for the Edinburgh speech recognition system based on feature extraction techniques. The design of the package is based on processing a speech "frame", i.e. a short-time segment of the order of 25.6ms, which is processed by an array of signal processing techniques which provide a rich acoustic description of the speech signal. The star_pak processing strategy can be viewed as applying firstly a set of kernel transformations, such as Fourier transformations and autocorrelation, which provide a set of primary transform descriptors. Certain of the outputs from these transforms are then further combined to provide secondary descriptions of signal transform domains (see Fig. 1), such as high frequency dominance, voiced/unvoiced, etc. The resulting primary and secondary signal descriptors are stored on a frame-by-frame basis into an output file which is related to the original input signal file and is referred to as a "star map" ("map" = Multiple Acoustic Parameter file).

The internal processing status of the star_pak is contained in a highly composite structure referred to as a "star_field", and this structure is written, shortly after its initialisation, to a file to allow post-processing interrogation if desired. However, the full power of the star_field lies in its ability to be examined and/or altered by a controlling module external to the star_pak. In particular, the use of such an observable and controllable processing structure permits both classical closed-loop control and expert system control of processing modules within star_pak.

Two output files are therefore manipulated by star_pak for each signal file it processes: the star_field, used in both read/write mode, containing a list of values (in mixed floating point/integer/character format) assigned to certain signal processing parameters; and the star_map, in write mode only, which contains the data resulting from the application of signal transforms to successive frames of the time waveform which was input to the package. It is important to note that star_pak is itself devoid of any recognition rules; that is, it does not seek to interpret the signal, but rather to provide a description of it in various transform domains on a per-frame basis.

RULE-BASED INTERPRETATION OF STAR_MAP OUTPUT

In off-line processing and analysis the first task of the star_pak is to create an uncommitted symbol array of empty symbol labels and data elements commensurate with the time length of the digitised speech waveform and the predefined frame length and interframe interval. The star_map is formally created by tagging symbolic labels to the array according to the processing mix identified by the command line interpreter, dependent upon the version of the star_pak which was invoked. This use of symbolic labelling within the star_map provides a convenient data retrieval and browsing device for high-level processing of the signal descriptors in a Lisp environment.

STAR_PAK: A SIGNAL PROCESSING PACKAGE FOR ACOUSTIC-PHONETIC ANALYSIS OF SPEECH

In its most basic form, star_pak provides an "open-loop" front end processing strategy to provide an acoustic description of each analysis frame. That is, signal processing algorithms are provided to satisfy computational requirements for rule-based phonetic interpretation of the speech waveform at the lowest level of the expert system. The investigation of the logical connectivity of elements of the star_map to form any given phonetic interpretation is achieved by applying rule-generating software, known as SEGLAB [1] , to prespecified test segments of speech whose phonetic content has been hand-transcribed by a human phonetic expert. Once rules have been established, then phonetic hypotheses are generated in the operational mode by processing the incoming signal description for each analysis frame.

OPTIMAL CONTROL OF STAR_PAK PROCESSING

Although providing an efficient and flexible framework for "open-loop" front end processing, the power of the star_pak lies in its ability to accept controlling decisions which directly affect the values of processing parameters and the retrieval of frames of signal data, which need not necessarily be time-sequential. The controlling structure in this respect is the star_field (see Fig.1) from which each processing module can be loaded with required parameters prior to processing any single analysis frame.

The extraction of information from any signal necessarily entails the measurement of one or several of its time-varying characteristics. In the case of a communications system processing a well-defined set of signal parameters which have been designed according to commonly-agreed criteria, then the signal recovery system can be as complex as satisfy desired detection error-rate criteria, since all information-bearing elements of the signal are pre-defined. In communications systems where the signal model is either too complex to be solved directly with a numerical algorithm or where the model is not well understood, however, the signal recovery system tends to be arbitrarily complex depending on the properties ascribed to the system, even ignoring the effects of channel signal-to-noise power ratio (SNR). Indeed, recovery of information from a signal generated from such an ill-defined system is dependent not only on the assumed physical properties of the system but also on the assumed protocol of the information carried by the time- varying components of the signal. The speech signal falls very much into the latter category of signals.

The presence of a signal in a communications channel implies that information is being transferred from a source to a destination. It is well-established that for a signal to transfer information, then there should be maximum uncertainty from one instant to the next as to the exact signal value which follows. The signal which transfers information must, at some time or other, be measured at the destination so that information transfer can be achieved. The presence of additive noise, however, increases the uncertainty in the measurement of the signal which is under observation. Very low SNR environments can render the signal virtually undetectable so that no useful signal information is transferred. There is, however, information still being transferred to the destination, but about the noise source. Neither the sequence of values describing the received waveform nor the rate of transfer of information correspond to that expected by the listening system.

That is, the application of "open-loop" front end processing to speech in high noise environments provides descriptions of the signal space which are insufficient for decoding the information contained in the speech waveform. These descriptors cannot however be considered as erroneous since they nonetheless represent epochs of the received signal + noise. They would, however, be wrong if they were used to produce interpretations of the desired signal.

From a signal processing standpoint, in order for a purely algorithmic approach to provide enhanced signal detection, then either (1) the system generating the signal must be observable or estimable, or (2) the signal space must be constrained to a limited set of well-defined states, or (3) the additive noise must be observable as an independent entity and must exhibit stationarity, at the very least in path length. Signal enhancement techniques associated with these conditions are (1) inverse filtering/cepstrum/source-filter deconvolution techniques, (2) matched filtering/cross-correlation, and (3) adaptive filtering/ spectral subtraction. Such signal enhancement techniques encounter severe problems when applied to the speech signal, however.

The techniques in groups (1) and (2) rely on all components of the possible set of signal values (the signal space) being orthogonal to each other, that is, maximally dissimilar. Most importantly, their application to real time problems depends crucially on there being only a very limited, well-established finite set of values in the signal space. Speech, on the other hand, has such a variety of signal values, particularly when viewed from a speaker-independent speech recognition task, that to store, say, all possible correlation references for the techniques in group (2) would be impractical even if real time cross-correlation across the entire signal space were possible.

The techniques in group (3) represent an improvement over those in groups (1) and (2) in that no prior detailed knowledge of the signal source is required; however, the noise source source must be or have been observable over a satisfactory period of time for its characteristics in both time and frequency to be evaluated. If this can be achieved, then either the noise waveform or an estimate of its spectrum can be subtracted from the signal + noise to yield the desired signal alone. However, such noise-cancellation strategies break down if either (a) the noise cannot be observed in isolation or (b) the statistical properties of the noise change at the same rate or faster than the adaptive algorithm can accomodate. Noise which exhibits speech-like qualities, for example, would be difficult to cancel using adaptive techniques. Short-duration sounds such as clicks and thumps would likewise be difficult to cancel effectively. An office environment, with its multiplicity of noise sources with a wide variety of statistical, temporal and frequency characteristics would represent a highly complex and expensive operation in signal processing, both in terms of processing time and hardware development.

Thus, signal processing by itself is likely to achieve minimal effect on the machine detectability of those parts of speech with a low SNR. A complementary approach to improve noise immunity involves an examination of the amount of redundancy in the speech waveform. It is this feature which allows not simply detection but also correction of errors within the message. In the speech recognition task, error detection and correction commences at the first level of processing which interprets the speech waveform, that is, the phonetic level, but must, for maximum correctability, extend into the upper levels of the recognition process into the lexical, syntactical and semantical areas of the speech recogniser. Such a hierarchical method of error detection followed by correction is well-established in digital data transmission systems, where waveform coding to deliberately produce redundancy in the digital data stream provides error detection and, if there is sufficient redundancy, also provides error correction.

The argument for redundancy then, starts with the interpretation of phonetic features in the speech waveform, and as such relates to random single errors in the detection of phonetic features. Longer errors due to, say, random bursts of noise which mask several phonetic features can be overcome by once again exploiting redundancy in the speech waveform, only now the redundancy has to be found at the level of the lexical data and above. Once again, correlates of this strategy are to be found in digital transmission systems, where techniques such as convolutional coding against long-term bursts of noise involve not simply redundancy at the individual bit level, but coding of the message itself in part or in whole. Indeed, the amount of noise immunity can be shown theoretically to be proportional to the amount of time spent encoding the message [2], and consequently decoding of the message takes longer too. However, the disadvantage is that the complexity of the processing which the overall system must accomplish also increases, but exponentially with coding time. Thus, real systems are constrained to reach a practical compromise between an acceptable level of noise immunity and processing time and complexity. A further complication to this approach is that not all talkers utilise phonetic redundancy in the same form or to the same extent. For example, it has been reported [3] that in the analysis and recognition of plosive sounds, formant transitions do not offer good discriminant cues in every case.

All of the above approaches assume that the descriptions of the received signal embody primary and secondary tranformations which have fixed processing parameters. There is no explicit attempt to characterise the noise and hence selectively and intelligently apply signal enhancement according to the perceived signal environment. Particularly in front end processing systems which are applied to speaker independent recognition, different talkers have different styles and modes of speaking which may affect the performance of processing algorithms and hence the subsequent phonetic interpretation of their transform descriptions. Indeed, different portions of the speech signal for the same speaker may require different parametric models for any single analysis technique.

Consider the case of a purely voiced speech uncontaminated by noise. Closed-loop optimal control can be used to govern the choice of parameters in a given spectral estimation technique prior to assignment of formants in order to optimise performance of the spectral estimator. For example, in estimating formants which undergo rapid transition using LPC techniques, the long time width of the autocorrelation window compared to the speed of movement of the formant centre frequency may cause the moving formant to have a much higher estimated bandwidth than it actually has; in remedying this, it is conceivable to have a feedback mechanism within the technique which would reduce the length of the autocorrelation window until there was less than, say, a given percentage change in formant bandwidth from frame-to frame.

However, it is very rare for signals with such high signal-to-noise ratios (SNR) to be forthcoming in a working environment for speech recognition such as an office. Taking the case of a formant tracking system based on LPC spectral estimation. the presence of, say, wideband noise will also cause the perceived bandwidth of a spectral resonance to increase, and this effect can likewise be reduced by decreasing the length of the time window over which the LPC model is calculated. This unfortunately also appears to have the adverse effect of masking steady-state, high-frequency, low-amplitude formants under certain circumstances.

In practice, then, it will be difficult to apply optimal control of variables within the analysis techniques when the effects of noise on the signal are effectively unknown. However, it would evidently be useful to have some means by which the environment affecting the signal could be estimated in order to exercise some control over, for example, spectrum estimation parameters. If it is not possible to implement some form of closed-loop control on the estimation technique itself, then the next best solution would seem to be to evaluate the effects of different types of ambient environment on a given technique and somehow sense the environment so that parameters in any processing technique in the acoustic front end can be altered to provide the optimum description of the short-time signal segment. This processing arrangement is conceptualised in Fig. 2, which illustrates the rôle of the star_field in providing parametric control according to some expert system supervisor.

SUMMARY

Star_pak is a flexible signal processing package which supplies transform descriptions of the speech signal on a frame-by-frame basis. In an open-loop processing architecture, star_pak processing parameters remain fixed with no attempt to mitigate the effects of additive noise. However, the architecture of the star_pak allows for both classical closed-loop control of individual processing techniques and, most importantly, expert system manipulation of processing parameters according to the perceived signal environment. It is considered that the application of intelligent noise cancellation via the star_pak architecture will provide for enhanced robustness in interpretation of the speech signal in low SNR environments.

This work was carried out under a U.K. Science and Engineering Research Council grant.

REFERENCES

[1] A. Blokland, G. Watson, J. Dalby, and H. Thompson, "Seglab: an interactive environment for phonetic segmentation and labelling of speech", Institute of Acoustics Conference on Speech and Hearing, Windermere, (Nov. 1986)

[2] C.E. Shannon, "A mathematical theory of communication", Bell System Technical Journal, v.27(4), pp.623-656 (Aug. 1948)

[3] R. de Mori, L. Lam, and M. Gilloux, "Learning and plan refinement in a knowledge-based system for automatic speech recognition", Technical Report No. SOCS 86.14, McGill University, Montréal (May 1986)

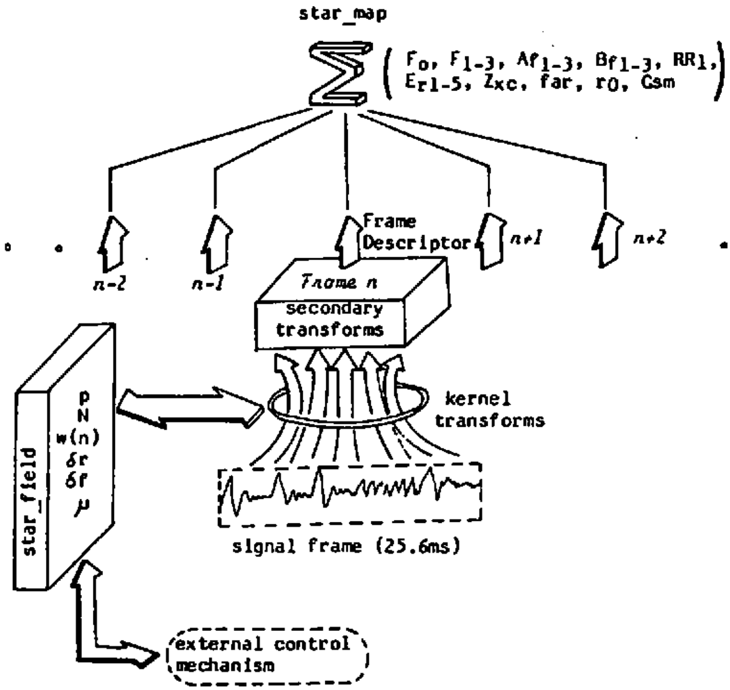

Fig. 1 Conceptualisation of star_pak processing. Star_map elements are: F_0 = fundamental frequency; F_{1-3}, Bf_{1-3}, Af_{1-3} = formant centre frequency, bandwidth and amplitude; r_0 = 0^{th} autocorrelation lag; RR1 = r_1/r_0; Er_{1-5} = energy spectral density ratios; far = frame amplitude range; Zxc = zero crossing rate.
star_field elements are:
μ = preemphasis factor; w(n) = window type; N = FFT length;
p = LPC model order; δr = z-transform radius step;
δf = frequency resolution.

STAR_PAK: A SIGNAL PROCESSING PACKAGE FOR ACOUSTIC-PHONETIC ANALYSIS OF SPEECH

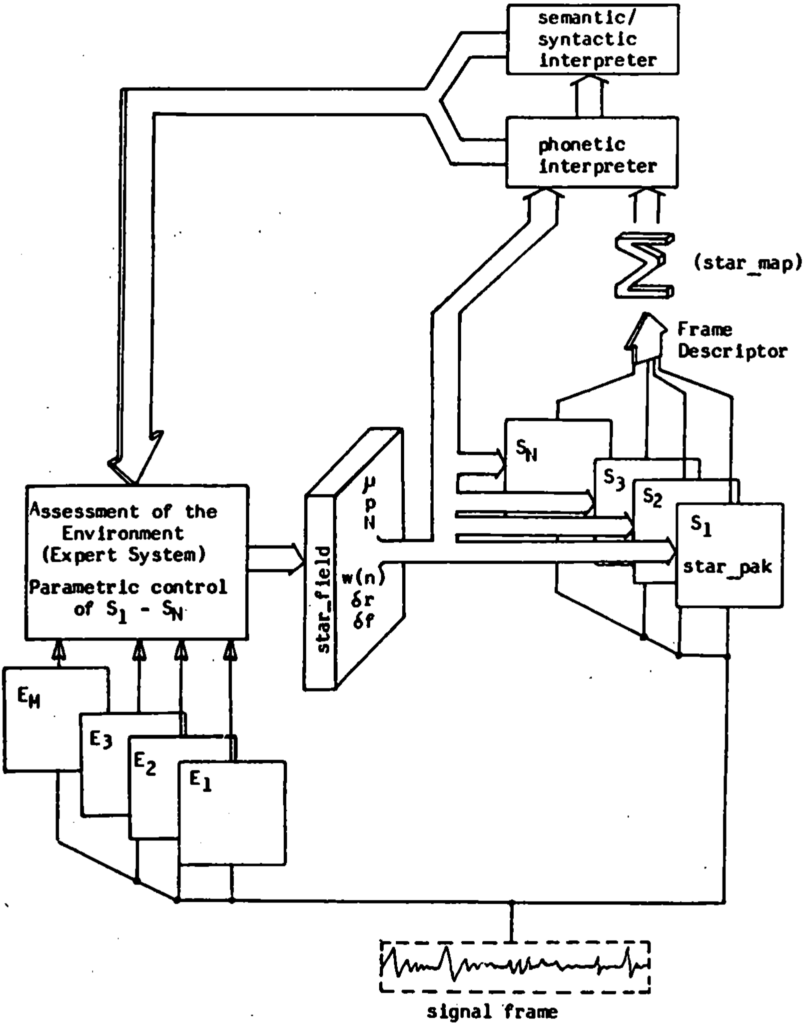

Fig. 2 Adaptive star_pak processing under expert system control.
E_{1-M} = fixed parameter processing modules to provide auxiliary signal data to the expert system.
S_{1-N} = variable parameter star_pak processing modules.