RECENT ADVANCES IN UTILISING PROSODY IN SPEECH RECOGNITION

Andrew J. Hunt

Speech Technology Research Group
Department of Electrical Engineering,
The University of Sydney

ABSTRACT - In the last few years both the potential for use of prosody in Automatic Speech Recognition and its actual use have grown substantially. The qualitative and quantitative understanding of prosodic features, including stress, pausing, rhythm, and intonation, has improved significantly. At the same time automatic speech recognition systems have reached a level of sophistication at which prosodic features can play a useful and complementary role to conventional recognition techniques. This paper outlines recent research work on the nature and utilisation of prosody and looks at areas of promise. A trend towards more sophisticated processing of prosodic features is observed.

INTRODUCTION

It has long been known that prosodic cues are of great importance in human speech communication. Prosodic features, including intonation, stress, pausing, rhythm and intensity, are known to aid listeners at all levels of speech understanding: e.g. phonetic recognition, lexical access, syntactic and semantic interpretation and pragmatics. It can be argued that computer-human communications will reach high standards by utilising the same cues.

In speech synthesis good prosody is vital for generating natural sounding speech, and is also important in providing effective understanding and acceptance of the synthesised speech. It can be argued that a parallel situation exists in the field of Automatic Speech Recognition (ASR). Many current speech recognition systems can be substantially enhanced by exploitation of prosodic cues and these enhancements will improve the utility of the systems.

Much of the work on utilising prosody in ASR follows from work by Lea in the 1970's. His seminal paper (Lea 80) is a good introduction to prosody and its potential uses in ASR.

This paper discusses the relevant work which has taken place in the fields of linguistics and speech technology. It is observed that the results being achieved have improved as the theoretical understanding of prosodic phenomena improves and as more sophisticated techniques are used for the processing of prosodic features. The paper focuses primarily on work for the English language. The paper is divided into four sections: the first outlines a framework for the processing of prosodic features, and the following sections outline major work on utilising prosody in ASR.

UTILISING PROSODY

The utilisation of prosody in ASR typically takes place in three stages:

[1] Linguistic explanation of a particular prosodic phenomenon or cue,
[2] "Engineering" extraction/measurement of the prosodic cue from the acoustic signal,
[3] Utilisation of the cue in a speech recognition/understanding system.

The first stage is the development of an understanding of a prosodic phenomena in human speech in the linguistic, phonology, or phonetic domains. This understanding can include descriptions of the acoustic signal, perceptual studies of features, and description of the phenomena within a theoretical linguistic framework. The work can be qualitative or quantitative. This work in linguistics and phonetics points to some way in which prosody can be utilised in ASR.

The second and third stages adapt this theoretical framework to an ASR system. The second stage is the extraction or measurement of the prosodic phenomenon in the acoustic signal by either direct or indirect means. This phase requires quantitative analysis of speech and will often need to adapt the linguistic/phonetic theory to a more measurable form. The third stage is utilisation of the extracted feature within an ASR system. Pragmatic steps must often be taken which result in using prosodic features in ways different from the strict linguistic explanations of the phenomenon.

The following three sections consider major representative examples of the use of prosody in ASR within this framework. The sections deal with three "levels" of prosodic processing - phonetic and word recognition, syntactic and semantic analysis, and pragmatic processing. Each section considers separately the linguistic theory (i.e. stage 1), and the application to ASR (i.e. stages 2 and 3).

PHONETIC RECOGNITION AND WORD RECOGNITION

Linguistic Background

Acoustic phonetic research and psycholinguistic research has shown that duration, intensity, pitch and stress all play a part in recognising phonemes and lexical items. As examples, Ainsworth (75) studied the effect of pitch in the perception of vowels. Ladefoged and McKinney (63) measured the intensity characteristics of vowels. The durational characteristics of phonemes has also been closely studied: Klatt (76) and van Santen (92) amongst others have attempted to systematically model duration. Work on perception has shown that intensity and duration are useful cues for phonetic processing (in English at least), but that pitch is of marginal significance as it is confounded by utterance level intonation. Silverman (87) has a good analysis of work on intrinsic pitch from both a production and perception viewpoint, and mathematically models other segmental pitch perturbations.

Studies show that the acoustic correlates of stress for most languages are a combination of intensity, pitch, duration and spectral pattern. In English all four features affect perception of stress (Lehiste 70, Fry 55, 58). The intrinsic intensity, intrinsic pitch and intrinsic duration (or quantity) of phonemes are also important in the perception of stress and pitch accent (Silverman 87).

Utilising Prosody for Phonetic Recognition and Word Recognition

A great deal of work has attempted to use duration and intensity explicitly for phonetic and word recognition; fewer have used pitch. Intensity is easy to measure from the acoustic signal, though the equivalent perceptual correlate of loudness is more difficult to estimate. Pitch can be measured with some accuracy, but there are still weaknesses in most pitch detectors (see Silverman 87 for analysis of certain pitch detection problems). Phoneme duration is more difficult to measure since the boundary and identity of each phoneme must be known. Current phoneme recognition does not appear to be sufficiently accurate for this purpose.

Robinson et al (90) used intensity along with other features for phonemic recognition using recurrent neural networks, but found pitch of marginal benefit. Intensity improved the performance of the HMM-based SPHINX recogniser (Lee 89). Recognisers based on HMMs implicitly utilise coarse durational characteristics of phonemes because training builds an underlying state system which models the temporal characteristics of speech. However, it is accepted that the temporal models of HMMs and the durational probabilities of speech are different and some work has been done on modifying HMMs to better model speech timing (Rabiner 89).

Waibel (84, 87, 88) performed isolated word recognition based on a variety of prosodic cues. He found intensity to be the best cue on its own and that other features such as stress and durational patterns were also effective. From analysis of large dictionaries, Waibel (88) concluded that prosodic cues will be of greatest use in large vocabulary systems where phonetic confusability is greater. This analysis also revealed a significant number of English word pairs which cannot be discriminated using phonetic information alone; however, most of these pairs can be discriminated using prosodic information.

Aull and Zue (85), Bannert (86), Bundgaard (89), Cheung et al (77), Friej and Fallside (88), Hieronymus (89), Vaissierre (88) and Waibel (88) all describe stress recognisers working on different styles of speech for a variety of languages.

Comments

Intensity, duration, pitch and stress are of some use for phonetic recognition and word recognition. Stress will be of more benefit for large vocabulary systems and all four will be more useful in continuous speech where spectral cues to phonetic identity are less clear. However, prosody does not provide outstanding improvements for current systems which already achieve high accuracy.

SYNTACTIC AND SEMANTIC PROCESSING

Linguistic Background

Research into the link between prosodic features and syntactic and semantic cues in spoken English is a contentious field. A number of theories have been developed to try to explain durational features, pausing, and intonation, and to link them to other linguistic knowledge. Links exist between prosodic boundaries marked by pauses, rate changes and syllable lengthening on one hand, and syntactic structure and information grouping on the other hand. Various models of this relationship have been developed (Price et al 91, Gee and Grosjean 83). More prominent syntactic boundaries tend to be marked by more prominent prosodic boundaries but the prosodic and syntactic structures are not identical. Wightman et al (92) investigated the acoustic determinants which marked lengthening associated with prosodic boundaries, which may then be used to predict syntactic structure. These prosodic cues can also be used by listeners to disambiguate syntactically ambiguous sentences (Lehiste 73).

Intonation is the linguistic entity reflecting pitch. The role of intonation for syntactic interpretation appears to be complementary to the temporal cues, but secondary in importance. The placement of pitch accents is affected by word class: content words (e.g. nouns and verbs) are far more likely to carry a pitch accent than function words (e.g. prepositions and articles). However, intonation plays a strong role in semantic processing. Intonation can indicate information status, focus, the given/new distinction and contrastiveness by appropriate placement of pitch accent, boundaries and changes in pitch range (Hirschberg et al 87). These cues have not been used in speech understanding systems to date, but may prove particularly useful.

Silverman et al (sub) investigated a particular semantic feature: the indication of primary focus of utterances by intonational cues. Users responding to queries from an automated system, marked information bearing words with a nuclear pitch accent, and often preceded and/or followed the words with a pause. Detection of the nuclear accent could be used to significantly enhance performance of word spotting systems.

Utilising Prosody for Syntactic and Semantic Processing

The extraction of the relevant parameters of pitch and duration was described in the previous section. Pausing is a relatively easy parameter to measure in the acoustic signal, but the detection of features such as pitch accent and syllable lengthening is more difficult. Pitch accent is marked in essentially the same way as stress, that is by pitch, duration and intensity. Syllable lengthening must be detected as a change from the expected or average length of syllables and individual segments, which is not particularly reliable with current technology. Wightman and Ostendorf (91) describe an algorithm to mark prosodic boundaries and estimate their relative significance based on syllable lengths.

Pausing information has been used by several researchers, including Lea (80) and Vaissiere (89), to locate major syntactic boundaries, working on the theory that longer pauses mark more significant syntactic boundaries. More sophisticated methods have been used by Ostendorf and Veilleux (91) who were able to reliably predict prosodic boundary features from syntactic parsing. Ostendorf et al (91) compared temporal cues with prosodic boundaries predicted from syntax to rank candidate sentences. The results of their automatic system compare favourably with human performance in the disambiguation of sentences. Bear and Price (90) looked at how duration could be used directly in a natural language parser. Rowles and Huang (92) have integrated durational cues and pitch accent information into a combined syntax-semantic parser with promising results.

Comments

Four observations are made on utilising prosody for processing syntax and semantics. First, using temporal and intonational cues requires continuous speech. Pausing information cannot be used in isolated word or separated word systems and intonation and durational information are not well controlled in broken speech. As ASR systems have only recently begun to show robust performance on continuous speech it is only reasonable that utilising prosody for syntactic and semantic processing is only now becoming practical.

Second, it is widely recognised that the temporal and intonational characteristics of speech are strongly influenced by the speaker's context: read and spontaneous speech show measurable differences in prosodic features (Silverman et al - sub). Thus, if the speaker's context is not appropriately handled or modelled, the effectiveness of utilising prosody is reduced.

Third, prosody offers the potential for allowing more freedom in grammars for recognisers. The combination of NLP and prosody may provide an effective means of integrating lexical recognition and higher level processing with prosody providing sensible constraints upon flexible grammars. Stochastic grammars, such as Kupiec (92) and Church (88), may also be able to integrate prosodic information within a statistical framework.

Finally, efforts have been undertaken recently to produce a standard prosodic labelling system (Silverman et al, 1992). This system will enable large prosodic databases to be developed. These databases should provide a basis for ongoing linguistic research and a good base for training prosodic recognition systems.

PRAGMATICS - DIALOGUE MANAGEMENT AND INTENTIONS

Linguistic Background

A variety of work has shown strong links between prosodic features and pragmatics. Intonation is used to indicate speaker intention, attitude and state-of-mind. Intention and attitude are of particular importance in management of human-machine dialogue. Pierrehumbert and Hirschberg (90), Hirschberg et al (87) and Grosz and Hirschberg (92), amongst others, looked at the relationship between speaker intention and pitch accent placement and investigated the way intonation can mark changes in topic, and mark new and given information. Hirschberg and Litman (87) looked at intonational characteristics of cues phrases used in dialogues (e.g. "Now, ...", "But, ...") and found that these cue phrases are distinguished by consistent phrasing and accent and point to shifts in topic and focus. There are also consistent means by which intonation is used between speakers in discourse to control turn-taking (Eggins et al 91).

Utilising Prosody with Pragmatics

To date, the exploitation of this information for ASR is limited, though work is continuing. The amount of data available for studying intonation in dialogue and for developing systems has increased substantially. The ATIS Speech Database provides a large copora from a Wizard-of-Oz simulation of a man-machine dialogue to provide a common database for the DARPA Spoken Language System Task (Hemphill et al 90).

Comments

A few observations are made. First, as discussed in the previous section, systems utilising prosody for pragmatics will be most effective with continuous speech and will also need to model context appropriately. Second, the use of prosody in this domain could enhance the naturalness of the human-machine interface. Studies of man-machine interaction have shown human speaker capacity to adapt to the capabilities of a system, and thus if a machine utilises more sophisticated cues then speakers are likely to use them. Third, interpreting prosodic features should improve response to different user intentions and aid dialogue management by providing high level cues to topic shifts and user intentions. These cues are only occasionally evident in orthographic transcriptions of dialogues and so prosodic analysis may provide information not available elsewhere in speech.

CONCLUSIONS

This paper has aimed to review the current state of use of prosody in speech recognition systems. There is growing interest and research in the area and the coming years offer the possibility of prosody being used to great effect. Prosody is most useful in large vocabulary, continuous speech systems and in systems with less restrictive grammars than are currently used. Prosody is also showing potential for improving dialogue management and interpretation of speaker intentions. Key trends in the field are the development of substantial databases to support research and the use of more sophisticated processing and classification techniques.

ACKNOWLEDGMENTS

The author wishes to thank Dr Julie Vonwiller for her patient help on introducing an engineer to the wide world of linguistics. The author is supported by a Telecom Australia Research Laboratory Fellowship and an Australian Postgraduate Research Award.

REFERENCES

Ainsworth, W., (1975) *Intrinsic and extrinsic factors in vowel judgements*, in G. Fant, M Tatham Auditory Analysis and Perception of Speech (Academic Press: London).

Aull, A., & Zue, V., (1985) *Lexical Stress Determination and its Application to Large Vocabulary Speech Recognition*, ICASSP '85, pp 4111-4114.

Bannert, R., (1986) *From prominent syllables to a skeleton of meaning: a model of prosodically guided speech recognition*, Working Papers, Department of Linguistics, Lund University.

Bear, J. & Price, P., (1990) *Prosody, Syntax and Parsing*, Proc. of the 28th Conf. Assoc. for Computational Linguistics, pp 17-22.

Bundgaard, M., (1989) *An Algorithm for Recognition of Stress in Danish and its Application in an ASR System*, Eurospeech 89, v1, p 526.

Cheung, J., Holden, A., & Minifie, F., (1977) *Computer Recognition of Linguistic Stress Patterns in Connected Speech*, IEEE Trans. on Acoustics, Speech and Signal Processing, pp 252-257.

Church, K.W., (1988) *A stochastic parts program and noun phrase parser for unrestricted text*, Proc. 2nd Conf. on Applied Natural Language Processing, Austin, pp136-143.

Eggins. S, Vonwiller. J, Sefton. P, & Matthiesson, C. (1991) *The Description of Minor Clauses in Information-Seeking Telephone Dialogues*, Research Paper, University of Sydney.

Freij, G.J., & Fallside, F., (1988) *Lexical stress recognition using hidden Markov models*, ICASSP '88, pp 135-138.

Fry, D.B., (1955) *Duration and Intensity as Physical Correlates of Linguistic Stress*, JASA Vol 27, pp 765-768.

Fry, D.B., (1958) *Experiments in the Perception of Stress*, Language and Speech, Vol 1, pp 126-152.

Gee, J.P., & Grosjean, F., (1983) *Performance structures: A psycholinguistic and linguistic appraisal*, Cognitive Psychology, Vol 15, pp 411-458.

Grosz, B., & Hirschberg, J., (1992) *Some Intonational Characteristics of Discourse Structure*, ICSLP '92, forthcoming.

Hemphill, C., Godfrey, J.J., & Doddington, G.R., (1990) *The ATIS spoken language systems pilot corpus*, Proc. of the DARPA Speech and Natural Language Workshop.

Hieronymus, J.L., (1989) *Automatic Sentintial Vowel Stress Labelling*, Eurospeech '89, pp 226-229.

Hirschberg, J., (1990) *Accent and Discouse Context: Assigning Pitch Accent to Synthetic Speech*, National Conf. American Assoc. for Artificial Intelligence, pp 952-957.

Hirschberg, J., & Litman, D., (1987) *Now Let's Talk About NOW: Identifying Cue Phrases Intonationally*, Proc. of the 25th Annual Meeting of the Association for Computational Linguistics.

Hirschberg, J., Litman, D., Pierrehumbert, J., & Ward, G., (1987) *Intonation and the Intentional Structure of Discourse*, Intl. Joint. Conf. on Artificial Intelligence, pp 636-639.

Klatt, D., (1976) *Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence*, JASA, Vol 59, pp 1208-1221.

Kupiec, J., (1992) *Robust part-of-speech tagging using a hidden Markov model*, Computer Speech and Language, Vol 6, pp 225-242.

Ladefoged, P., & McKinney, N.P., (1963) *Loudness, sound pressure and subglottal pressure in speech*, JASA, Vol 35, pp 454-460.

Lea, W.A., (1980) *Prosodic Aids to Speech Recognition*, in W.A. Lea, Trends in Speech Recognition, pp 166-205, (Prentice-Hall: Englewood Cliffs, NJ).

Lee, K.F., (1989) *Automatic Speech Recognition: The Development of the SPHINX system*, (Kluwer Academic Publishers: Boston).

Lehiste, I., (1970) *Suprasegmentals*, (MIT Press: Cambridge, MA).

Lehiste, I., (1973) *Phonetic disambiguation of syntactic ambiguity*, Glossa, Vol 7, pp 107-122.

Ostendorf, M., & Veilleux, N., (sub) *A Heirarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location*, Submitted manuscript.

Ostendorf, M., Wightman, C.W., & Veilleux, N.M., (forthcoming) *Parse Scoring with Prosodic Information: An Analysis/Synthesis Approach*, to appear in Computer Speech and Language.

Pierrehumbert, J., & Hirschberg, J., (1990) *The Meaning of Intonational Contours in the Interpretation of Discourse*, in P.R. Cohen, J. Morgan, M.E. Pollack, Intentions in communication, pp 271-311 (MIT Press: Cambridge, MA).

Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C., (1991) *The Use of Prosody in Syntactic Disambiguation*, JASA, Vol 90, pp 2956-2970.

Rabiner, L.R., (1989) *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, IEEE Proceedings, Vol 77, pp 257-285.

Robinson, T., Holdsworth, J., Patterson, R. & Fallside, F., (1990) *A comparison of preprocessors for the Cambridge recurrent error propagation network speech recognition system*, Intl. Conf. on Spoken Language Processing, pp 1033-1036.

Rowles, C. & Huang, X-M., (1992) *Prosodic Aids to Syntactic and Semantic Analysis of Spoken English*, Proc. of 30th Annual Meeting of the Assoc. for Computational Linguistics, pp. 112-119.

Silverman, K.E.A., (1987) *The structure and processing of fundamental frequency contours*, PhD Thesis, University of Cambridge.

Silverman, K.E.A., Blaauw, E., Spitz, J. & Pitrelli, J.F., (sub) *Towards Using Prosody in Speech Recognition/Understanding Systems: Differences between Read and Spontaneous Speech*, Submitted manuscript.

Silverman, K.E.A., Zue, V., Pierrehumbert, J., Price, P., Ostendorf, M., Beckman, M., & Hirschberg, J. (1992 - forthcoming) *A standard scheme for labelling prosody*, to appear in Intl. Conf. on Spoken Language Processing.

Vaissiere, J., (1988) *The Use of Prosodic Parameters in Automatic Speech Recognition*, in H. Niemann, M. Lang, G. Sagerer, Recent Advances in Speech Understanding and Dialog Systems, pp 71-99, (Springer-Verlag: Berlin).

Vaissiere, J., (1989) *On automatic extraction of prosodic information for automatic speech recognition*, Eurospeech 89, pp 202-205.

van Santen, J.P.H., (1992) *Deriving text-to-speech durations from natural speech*, in G Bailly, C Benoit, T.R. Sawallis, Talking Machines: Theories, Models and Designs, pp 275-285, (Elsevier Science Publishers).

Waibel, A., (1984) *Suprasegmentals in Very Large Vocabulary Isolated Word Recognition*, ICASSP '84, pp 26.3.1-4.

Waibel, A., (1987) *Prosodic Knowledge Sources for Word Hypothesization in a Continuous Speech Recognition System*, ICASSP '87, pp 856-859.

Waibel, A., (1988) *Prosody and Speech Recognition*, (Pitman Publishing: London).

Wightman, C.W., & Ostendorf, M., (1991) *Automatic Recognition of Prosodic Phrases*, ICASSP 91, pp 321-324.

Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P.J., (1992) *Segmental Durations in the Vicinity of Prosodic Phrase Boundaries*, JASA, Vol 91, pp 1707-1717.