ASSESSING SYLLABLE STRENGTH VIA AN AUDITORY MODEL.

M. Allerhand, S. Butterfield, A. Cutler, R. Patterson.

MRC Applied Psychology Unit,
15 Chaucer Road, Cambridge CB2 2EF.

1. INTRODUCTION.

We describe an empirical study of the processing of prosodic information in continuous speech. The results of a previous word boundary perception experiment are correlated with measures of syllable strength derived from an auditory model. The aim is twofold: to evaluate the performance of prototype measures of pitch–strength and loudness as a criterion of syllable strength, and to correlate these measures with human listeners use of syllable strength as a word boundary cue.

One of the outstanding problems in speech recognition is the difficulty of reliable segmentation of continuous speech, which can lead to an explosion of complexity at the lexical access stage of speech recognition. A strategy for locating the points in a continuous speech signal from which attempts at lexical access are most likely to be successful would be very helpful to a recogniser. Ranking the points in a speech signal according to their likelihood as word boundaries is not simply an efficiency issue; it is also crucial for disambiguating phonologically similar phrases, providing evidence to justify the selection of one phrase over another without recourse to very high (eg semantic) levels of processing. It is well-known that most English sentences of reasonable length allow many possible parses with sub-word transcription (Harrington and Johnstone, 1987; Briscoe, 1989). Although many of these will be nonsensical, they will nevertheless be valid at all levels of processing below the semantic level.

Cutler and Norris (1988) suggested that the word boundaries can be significantly disambiguated by exploiting the prosodic probabilities of a language. They have proposed a heuristic strategy based upon the metrical structure of a stress language, such as English, in which syllables may be either strong (containing full vowels) or weak (containing reduced vowels, usually a schwa). The evidence that the English language is structured in a way which facilitates word boundary location using the distinction between strong and weak syllables comes from statistical analyses of the English vocabulary. Cutler and Carter (1987) analysed the metrical structure of English words from a very large (over 98,000 words) corpus, and also their frequency of occurrence in spontaneous British English conversation, to determine the distributions of strong and weak syllables in relation to word boundaries. They found that, on average, a strong syllable has about a three to one chance of being the onset to a new lexical word (such as a noun, adjective, or verb). A weak syllable, on the other hand, is most likely to be a grammatical word (such as an article, preposition, and such words of simple grammatical function). Similar results are reported by Waibel (1988). The metrical segmentation strategy (MSS) exploits the probability that most lexical words begin with strong syllables. The distributions of strong and weak syllables indicate that the MSS is not only able to rank points in a speech signal as word boundaries, but is able to broadly identify the type of word boundary.

Prosodic information therefore provides a link between phonology and syntax. It provides pre-lexical information of word class, dividing the vocabulary broadly into lexical words and grammatical function words. This suggests that people can use prosodic information to disambiguate phonologically similar phrases having different, though equally valid, syntax. Using a study of human segmentation errors, Cutler and Butterfield (1992) found evidence that humans do make use of such a strategy; this work is described in more detail below (section 2.2).

The MSS has not been implemented in a speech recognition system. Practically all work in speech recognition (e.g. using HMMs) uses input data derived from smoothed spectra, and therefore makes no use of prosodic information. The idea of a pre-processor to transform the acoustic speech signal into certain time-varying parameters which are representative of basic auditory features such as pitch, pitch-strength, and loudness, which are known to contribute to perceived syllable strenghth, was proposed by Zwicker and Terhardt (1979). However the system was never fully realized. It required the development of appropriate algorithms for the auditory transformations which have only recently become available.

A computational model of temporal pitch theory was originally proposed by Licklider (1951) based on a multi-channel autocorrelation analysis, but it required the development of a functional model of the cochlea to simulate the spectral analysis and neural transduction which preceed the neural autocorrelation process. Thus it was some time before the first practical implementation of the model appeared (Lyon, 1984). Modern implementations of Licklider's "Duplex Model of Pitch", called "correlograms", have been shown to explain many complex pitch phenomena (Slaney and Lyon, 1990; Meddis and Hewitt, 1991a, 1991b). A similar respresentation based upon a computational model of auditory processing (Patterson et al. 1992), called an "auditory image", has been developed by Patterson and Holdsworth (1992). This representation, originally designed as a model of temporal integration, has been shown to be closely related to the correlogram (Allerhand and Patterson, 1992). Both correlogram and auditory image are representations designed to explicate prosodic information.

Patterson (1987) has shown that a "spiral mapping" of the auditory image can further enhance prosodic information. This spiral mapping (which could equally be applied to the correlogram) presents pitch, pitch–strength, and loudness information in the form of a pattern of spokes radiating from the centre of a spiral, (the details of the operation of the spiral are beyond the scope of this short paper). Recently Allerhand et.al. (1991) developed pattern recognition algorithms to extract three features from the spiral representation of the auditory image which arguably represent three auditory sensations: pitch–strength, loudness, and pitch chroma. Pitch-strength is a measure of the randomness of a signal, or alternatively of the amount of periodicity in noise; a measure designed to have a low value if the signal is mostly noise, and a high value if the signal is mostly periodic (Hall and Soderquist, 1975; Terhardt et al 1982). Loudness is a measure which varies with signal intensity, but is not a simple intensity measure (Stevens, 1972; Zwicker et al, 1991). The algorithm (Allerhand et.al., 1991) gives a pitch–strength measure which is independent of the sound level and the pitch of the speech signal. An evaluation using repetition pitch, generated from iterated ripple noise (Yost, 1980), has shown that this algorithm agrees closely with human performance. The prototype measures of pitch–strength and loudness do not account for variations in timbre.

Given the strong theoretical basis for the application of prosody in speech recognition, and the new algorithms for auditory prosodic feature extraction, we felt it was timely to implement an empirical study.

2. EMPIRICAL STUDY.

This study consists of (a) a measurement experiment (b) correlation of measurements with existing results of a perception experiment. Both measurement and perception experiments used the same speech database. Results are presented showing the measurements as features for discriminating strong and weak vowels and syllables, and showing the correlation between the measurement and the perception data.

2.1 Design of the speech database.

The speech materials used in this study were used in two psycholinguistic studies of the role of rhythm in the perception of word boundaries (Smith, Cutler, Butterfield, and Nimmo-Smith, 1989; Cutler and Butterfield, 1992). They consisted of 48 unpredictable sequences of six syllables. (For a complete list see either of the two references). Each sequence had an alternating stress rhythm of strong (S) and weak (W) syllables. In half the cases the rhythm was SWSWSW (e.g. *soon police were waiting*); in the other half it was WSWSWS (e.g. *conduct ascents uphill*). These manipulations resulted, obviously, in exactly equal numbers of strong and weak syllables in the sequences as a whole as well as in each syllable position. Each of the two rhythmic structures allows very many different possible divisions into words, and each is a very common pattern in English.

Two further factors were varied systematically in the materials. One was where word boundaries occurred with respect to the rhythm. One-third of the sequences had only weak word-initial syllables (e.g. *conduct ascents uphill*; *sons expect enlistment* - note that although in the latter example the very first syllable is strong, the first syllable in each string is irrelevant, since it is necessarily word-initial). A further one-third had only strong word-initial syllables (e.g. *dusty senseless drilling*; *an eager rooster played*); and the remaining third had a mixture of strong and weak word-initial syllables (e.g. *soon police were waiting*; *achieve her ways instead*). Roughly equal numbers of strong and weak syllables were word-initial versus non-word-initial. This divides the set of syllables into four subsets according to type: strong word-initial, strong non-initial, weak word-initial and weak non-initial.

The remaining factor was the nature of the vowel in the strong syllables. These were chosen from a set of three phonetically short vowels (/ɛ/, /ɪ/, /ʌ/) and a set of three phonetically long vowels (/eɪ/, /i/, /u/). One quarter of the utterances contained all long vowels in the strong syllables (e.g. *soon police were waiting*); one quarter contained all short vowels (e.g. *conduct ascents uphill*); and the remaining half contained a mixture of long and short vowels (e.g. *achieve her ways instead.*) The weak vowels were mostly schwa. The 48 sentences were recorded in a sound dampened room by a phonetically trained male speaker of Southern British English.

2.2 The perception experiment.

Cutler and Butterfield (1992) presented the recorded sentences to 18 listeners in a "faint speech" experiment, i.e. at the level of the estimated speech reception threshold for each subject. The subjects were told that they would be listening to "speech that is difficult to hear clearly". Their task was to write down what they thought was said. The measure was the number of word boundary misperception errors which occurred on each syllable, and the main finding of the study was that boundary insertion errors were more common before strong than before weak syllables, while boundary deletion errors were more common before weak than before strong syllables. Thus in the response *the doctor sends her bill* to *conduct ascents uphill*, a word boundary has been inserted prior to every strong syllable (-duct → doc-, -scents → sends, -hill → bill), but deleted prior to one weak syllable (a- → -tor). The number of errors which subjects made in this study provides the perceptual data with which we here compare the syllable strength measures derived from the auditory model.

2.3 The measurement experiment.

Each of the 48 recorded sentences was manually marked to indicate the syllable boundaries. The portions of each sentence before the onset of the first syllable and after the offset of the last syllable were marked off, and between these the six syllables were considered to be contiguous. The guiding principle was to mark syllables within words, (e.g. *trus/ting*), and morphology across words, (e.g. *co/llect/e/nough* rather than *co/lle/cte/nough*). Stop closures were typically included in the region of the syllable preceeding the burst.

The pitch-strength and loudness measures were sampled during stretches of strong and weak syllables drawn from the marked speech database. Four measurements of each syllable were made: the maximum values of pitch-strength and loudness during a syllable, and the area under the pitch-strength and loudness measures. This area measure was defined as the sum of the respective measures which exceeded half the maximum value during a syllable, and was designed to incorporate in a simple way both level and duration of the respective pitch-strength and loudness measures of syllable strength.

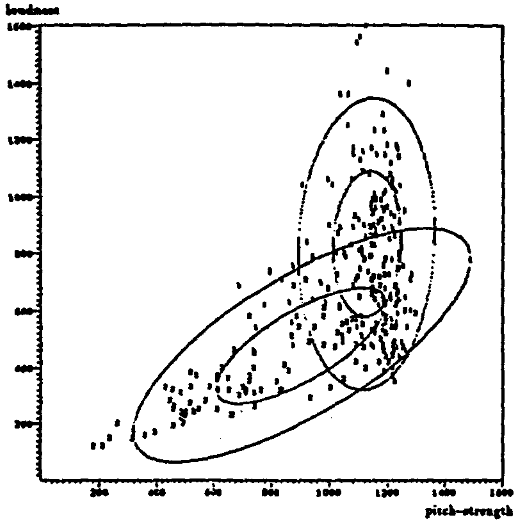

Figure 1. Maximum measures per syllable.

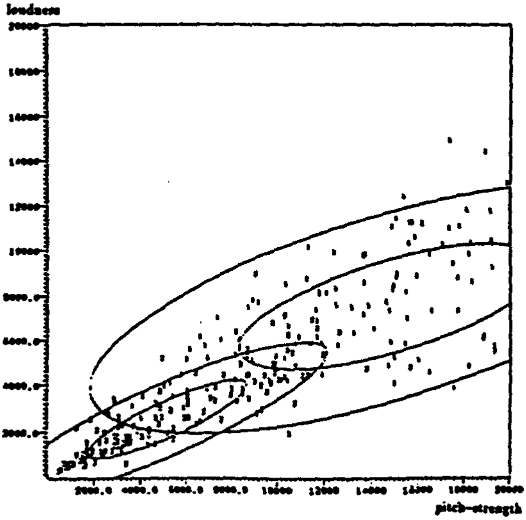

Figure 2. Area measures per syllable.

3. RESULTS.

Figures 1 and 2 show scatter diagrams of coordinate pairs of pitch–strength and loudness values corresponding to each syllable. Figure 1 shows the maximum pitch–strength and loudness values, and figure 2 shows the corresponding area measures. In either diagram coodinate pairs are labelled with "1" or "2" according to whether the measurements originate from a strong or a weak syllable respectively. The diagrams are superimposed with Gaussian contours showing the distributions of the measurements from the strong and weak syllables.

The absolute number of errors made by the 18 listeners on each syllable was tabulated, and correlation analyses were undertaken between these values and the four syllable strength measures. An initial analysis established the validity of the measures across the data set as a whole. Since strong and weak syllables respectively pattern quite differently in the human error data, and the syllable strength measures appear to be distinguishing between strong and weak syllables, we would expect significant correlations between the measures and the human data. Boundary insertion errors were treated in this analysis as positive values, and boundary deletion errors as negative values. This allowed us to predict a positive correlation, since the higher the syllable strength measure (i.e. the "stronger" the syllable), the more likely a boundary insertion error (positive value) and the less likely a boundary deletion error (negative value) should be.

As expected, the results showed that across the whole data set, there was a positive correlation between likelihood of an insertion error and all four measures. Each of the four correlation coefficients was significant at at least the .01 level. Since, as we have already noted, the likelihood of an insertion error was much greater for strong than for weak syllables, this result confirms the patterns shown in Figures 1 and 2; the syllable strength measures are efficiently distinguishing between strong and weak syllables.

The correlation across the whole data set arises from the fact that both syllable strength measures and error data effectively separate strong from weak syllables. It does not, however, follow that the syllable strength measures and the error data vary in parallel within the strong and weak syllable subsets (see section 2.1). That is, it is not necessarily the case that listeners are using the dimensions of variation captured by the syllable strength measures in making the word boundary decisions reflected by their performance in the perception experiment.

To test this, separate correlation analyses were carried out for each of the four types of syllable: strong initial and non-initial, weak initial and non-initial. Thus this analysis produced 16 separate correlation coefficients - four correlation analyses between syllable strength measures and error data, for each of four syllable subsets. None of these correlations reached significance at the .05 level, and so we conclude that the syllable strength measures are not correlated with the error data within the four syllable subsets.

3.1 Conclusion.

The significant correlations across the data set as a whole arise solely from the fact that both syllable strength measures and error data are effectively distinguishing between strong and weak syllables. This analysis, summarised in Figures 1 and 2, shows that the prototype measures of pitch–strength and loudness, as factors contributing to a syllable strength measure, have succeeded extremely well in capturing the distinction between strong and weak syllables. It also suggests that a rudimentary measure of duration, as incorporated into the measures shown in figure 2, can improve the measured separation between strong and weak syllables. It would appear, however, that the measures perform in a different way from the way human listeners make the same distinction.

The expected correlation between perceptual errors in judgement and the syllable strength measures would be: (1) Strong syllables with a relatively low strength measure correlate well with the strong syllables which were more easily mistaken as weak, (and as non-word–initial). (2) Weak syllables with a relatively high strength measure correlate well with the weak syllables which were more easily mistaken as strong, (and as word–initial). However this was not the case. The insignificant correlation between the measures and the perceptual error data suggests that the gradient distinctions captured by the syllable strength measures are not fully capturing the nature of the judgements made by human listeners as to whether a syllable is strong or weak. This in turn accords well with the finding of Cutler and Fear (1991) suggesting that human listeners prefer to group strong versus weak vowels categorically according to vowel quality rather than treating them as varying continuously on dimensions such as duration, pitch–strength and loudness. That is, when listeners are judging whether a syllable is strong or weak, (and accordingly more or less likely to be the initial syllable of a lexical word), they base their judgements, at least in part, upon the spectral quality of the vowel.

The prototype prosodic measures of pitch–strength and loudness do not account for variations in timbre. Although these features successfully separate strong and weak syllables, the results suggest that people may use timbre information as well as purely prosodic information when making segmental decisions. For example, the pitch–strength and loudness (as measured independent of the timbre) of a schwa are not generally sufficient to cue the difference between an initial and a non-initial weak syllable.

Nevertheless, the prototype measures segregate strong and weak syllables, and may incorporate useful lexical segmentation into an automatic speech recognition system. Whether this would gain added efficiency by the incorporation of spectral information as well is the subject for further research.

Acknowledgements.
The work presented in this paper was supported by the UK Medical Research Council and grants from Espirit BRA (3207) and MOD PE (XR 2239 and SLS/42B/663).

REFERENCES.

Allerhand M., Patterson R., (1992) Autocorrelation and Auditory Images. Submitted to *JASA*.

Allerhand M., Patterson R., Robinson K., Rice P. (1991) "Application of the SVOS Algorithm", Research Report V, MOD PE SLS/42B/663, August 1991.

Cutler A., Butterfield S. (1992) Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31, 218-236.

Cutler A., Carter D. (1987) *The predomination of strong initial syllables in the English vocabulary. Computer Speech and Language* 2 pp133–142.

Cutler A., Fear B. (1991) Categoricality in acceptability judgements for strong versus weak vowels. *Proceedings of the ESCA Workshop on Phonetics and Phonology of Speaking Styles*, Barcelona; 18.1-18.5.

Cutler A., Norris D. (1988) The role of strong syllables in segmentation for lexical access. *J. Exptl. Psych: Human Perception and Performance* 14 pp113–121.

deCheveigne A. (1986) A pitch perception model. *Proc ICASSP*-86 897–900

Hall J.W., Soderquist D.R. (1975) Encoding and pitch strength of complex tones. *J. Acoust. Soc. Am.* 58 1257–1261.

Harrington J., Johnstone A. (1987) The effects of word boundary ambiguity in continuous speech recognition. *Proc. 11'th International Congress of Phonetic Sciences*, Vol 3, pp89–92, Tallinn, Estonia.

Licklider J.C.R. (1951) A Duplex Theory of Pitch Perception. *Experientia* VII/4 128–134.

Lyon R.F. (1984) Computational models of neural auditory processing. *Proc. ICASSP* 1984 3.

Meddis R, Hewitt M.J, (1991a) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *J.Acoust.Soc.Am.* 89 6 pp2866–2882.

Meddis R, Hewitt M.J, (1991b) Virtual pitch and phase sensitivity of a computer model of the auditory periphery. *II: Phase* Sensitivity. *J.Acoust.Soc.Am.* 89 6 pp2883–2894.

Patterson R.D. (1987) A pulse ribbon model of peripheral auditory processing. In: *Auditory Processing of Complex Sounds*, W.A. Yost and C.S. Watson (eds), Erlbaum, New Jersey, pp167-179.

Patterson R.D., Holdsworth J. (1992) A functional model of neural activity patterns and auditory images', In: *Advances in Speech, Hearing and Language Processing* Vol 3, W.A. Ainsworth (ed), JAI Press, London (in press).

Patterson R.D., Robinson K., Holdsworth J., McKeown D., Zhang C., Allerhand M. (1992) Complex sounds and auditory images. In: *9th International Symposium on Hearing: Auditory physiology and perception*, Y Cazals (ed), Pergamon, Oxford (in press).

Slaney M., Lyon R.F. (1990) A perceptual pitch detector. *Proc. ICASSP 1990* pp357–360.

Smith M.R., Cutler A., Butterfield S., Nimmo-Smith I. (1989) The perception of rhythm and word boundaries in noise-masked speech. Journal of Speech and Hearing Research, 32, 912–920.

Stevens S.S. (1972) Perceived level of noise. *J.Acoust.Soc.Am.* 51 575–601.

Terhardt E., Stoll G., Seewann M. (1982) Algorithm for extraction of pitch and pitch saliance from complex tonal signals. *J.Acoust.Soc.Am.* 71 679–688.

Van Immerseel L., Martens J.P., (1992) Pitch and voiced/unvoiced determination with an auditory model. *J. Acoust. Soc. Am.* 91 3511–3526.

Waibel A. (1988) *Prosody and speech recognition.* London: Pitman.

Yost W. (1980) "Temporal properties in pitch and pitch strength of rippled noise", in Psychophysical, Physiological and Behavioural Studies in Hearing, eds G van den Brink, F Bilsen, pubs Delft U.P., The Netherlands, 1980.

Zwicker E., Fastl H., Widmann U., Kurakata A., Kuwano S., Namba, (1991) Program for calculating loudness. *J.Acoust.Soc.Jpn.* 12 39–42.

Zwicker E., Terhardt E. (1979) Automatic speech recognition using psychoacoustic model. *J. Acoust. Soc. Am.* 65 487–498.