CRITERIA FOR LABELLING PROSODIC ASPECTS OF ENGLISH SPEECH

Paul C. Bagshaw Briony J. Williams

Centre for Speech Technology Research, University of Edinburgh
80 South Bridge, Edinburgh EH1 1HN, Scotland, UK

ABSTRACT

We report a set of labelling criteria which have been developed to label prosodic events in clear, continuous speech, and propose a scheme whereby this information can be transcribed in a machine readable format. We have chosen to annotate prosody in a syllabic domain which is synchronised with a phonemic segmentation. A procedural definition of syllables based on the grouping of phones is presented. The criteria for hand labelling the prominence of each syllable, tone-unit boundaries and the pitch movement associated with each accented syllable, are described. Work to automate this process is presented and experimental results evaluating its performance are included.

I. INTRODUCTION

The need for a large corpus of prosodically labelled English speech is motivated by the use of prosodic events in training speech synthesisers, in automated foreign language pronunciation teaching, and to aid parsers used in speech recognition to disambiguate phonetically similar, but syntactically different utterances.

Speech synthesis requires a mapping from prosodic events to a set of acoustic parameters for their realisation. Parsers and the analysis of language pronunciation, on the other hand, require the reverse mapping to provide descriptors for the acoustic correlates of prosody, and semantic and pragmatic knowledge to be extracted from these correlates. The prosodic labelling of a language corpus must therefore annotate both the linguistically significant features in speech prosody and the inflections of the acoustic parameters.

We aim to transcribe sentential stress (the prominence of syllables in continuous speech) and the pitch movement associated with any accented syllables for such systems. By initially hand labelling these prosodic aspects, a set of acoustic features are sought which will form a mapping for speech synthesis, and at the same time, enable these prosodic events to be labelled automatically given the acoustic features, for parsers and language pronunciation description. The transcription system we propose is intended to be an annotation scheme for linguistically significant prosodic events in English. It is not designed to give a detailed description of every possible inflection in an F0 contour. The set of symbols (see table 1) is designed for use by both a hand transcriber of the prosodic events and for some automated procedure.

The labelling scheme described has been used to transcribe, by hand, prosodic events in a database of 453 utterances from the English language ATR conference-registration dialogues with focus[1]. An acoustic analysis of these labels attempts to establish a correlation between a set of features chosen to characterise the acoustic parameters believed to manifest prosody, and the perceived prosodic events that are transcribed.

Continuous speech is initially segmented into phone units and labelled using a HMM-based automatic segmenter (evaluated in [18]). The phones identified are grouped into syllables. Syllable boundaries are thus synchronised with the phone boundaries. The procedure employed to group phones into syllables is described in section II. Each syllable is labelled by hand as unstressed, stressed (but not accented), stressed and accented (but not nuclear), or as the nuclear accented syllable of a tone-unit. Each syllable immediately preceding a tone-unit boundary is also marked, in order to specify the boundary location. The nuclear accented syllable of a tone-unit is (according to the "British School" of intonational phonology) the final accented syllable in that tone-unit [4]. This definition of nuclear syllables and the criteria used to determine syllable prominence are addressed in section III. Each accented syllable is associated with an additional label that describes the pitch contour movement which marked it. Thus, pitch contour labelling is also synchronised with syllable boundaries. The time location of this movement may occur before, during and subsequent to the domain of the accented syllable. Pitch contour labelling criteria are described in section IV. In section V a set of acoustic features are proposed which we intuitively feel will describe the acoustic correlates of sentential stress. These acoustic features are used to form a tree-based statistical model for a small corpus of hand labelled prosodic events. This methodology is described in section VI. Its application reveals a low correlation between the acoustic features and the events labelled, which poses questions regarding the relationship between the theory and the acoustics of sentential stress. These are discussed in section VII.

II. SYLLABIFICATION

The following procedural definition is used for syllabification.

i) Phones are grouped into syllables on a phonological rather than phonetic basis. Consonantal phones (such as [m, n, l, r, s]) which may result in schwa deletion [9, pp.297–299] [5] and take on the syllabic nucleus, are therefore syllabified as if the vowels were present. Hence, *shortest* in rapid speech is syllabified as /ˈʃ ɔ – t s̩ t/, and *additional* as /ə – ˈd ɪ – ʃ n̩ – l̩/. A glottal stop that may occur before or instead of a word-final stop is treated as an instance of the underlying stop phone, and any glottalised onset to vowels is considered to be part of the vowel.

ii) Syllable boundaries are formed from the boundaries of words considered in isolation. Although in continuous speech, consonants at the end of one word can syllabify with the initial vowel of the following word [13], such resyllabification is not necessary in forming a domain in which to describe prosodic events. Thus, for example, the syllabification of *at all* differs to that of *a tall* even if /t/ is aspirated in both cases and they are phonetically identical. This approach has been adopted because the exact boundaries between syllable nuclei are not of critical im-

[1]The ATR dialogues where spoken by a female bilingual speaker of Japanese and American English.

portance, although identifying the nuclear phone is. Similarly, resyllabification is unnecessary across words that appear to blend together due to vowel deletion, as may be the case in *under a*, which is syllabified as /ˈʌ n – d r̩ – ə/.

iii) The boundaries between syllables are also determined by the presence of a morphological boundary. The boundary between a free morpheme and an inflectional suffix (except **-s**) or a class-II derivational affix is taken to be a syllable boundary. Thus, *hopeless* is syllabified as /ˈh əʊ p – l ə s/ rather than /ˈh əʊ – p l ə s/; and *uninteresting* is syllabified as /ʌ n – ˈɪ n – t ə – r ɛ s t – ɪ ŋ/ rather than /ʌ – ˈn ɪ n – t ə – r ɛ – s t ɪ ŋ/.

iv) On the basis of English phonotactics, any cluster of phones forming the onset or the coda of a syllable must also be a permissible word-initial or word-final cluster. According to this rule, *extra* may be syllabified as /ˈe k – s t r ə/, /ˈe k s – t r ə/, or /ˈe k s t – r ə/.

v) The 'maximal onset (and minimal coda) principle' [16] [3, pp.10–18] arbitrates between competing analyses. According to the principle, as many consonantal phones as possible form a syllable onset. Using this principle, *extra* would be syllabified as /ˈe k – s t r ə/. However, in cases when alternative boundaries are possible, stressed syllables tend to attract consonants more than unstressed ones, particularly in the case of ambisyllabic consonants such as [s, f] [7, pp.19–23]. When this final criterion is applied, the syllabification adopted for the example becomes /ˈe k s – t r ə/.

III. SENTENTIAL STRESS LABELLING

The salience of each syllable within an utterance is labelled as one of $\{u, s, a, n\}$ (see table 1) on the following basis.

Sententially stressed syllables are those that are perceived as salient due to a prominence of energy and/or duration and/or pitch [6] [12, chap 4] within an utterance. The default (and therefore intonationally unmarked) pitch movement in English is a slight downwards trend in pitch [4, 11]. This movement does not give any intonational prominence to the syllable within the declination, even if that syllable is stressed on the grounds of prominent duration and/or intensity. The same situation occurs if a stressed but unaccented syllable is one in a series of gently rising pitch movements. Where there is no pitch discontinuity, there is no accent [4]. An accented syllable must also be a stressed syllable and an accompanying pitch movement must occur during the accented syllable or on a syllable before or subsequent to the perceived accented syllable [8].

Each tone-unit of an utterance will have one peak of prominence in the form of a nuclear pitch movement. The nuclear accented syllable is the syllable on which the one obligatory pitch movement occurs in a tone-unit. This is traditionally believed to be the final accented syllable in a tone-unit [15]. At present, we make use of this traditional definition.

Tone-unit boundaries are marked by placing a diacritic $\{:\}$ on the label $\{u, s, a, n\}$ of the syllable immediately preceding the boundary. The tone-unit boundaries are identified by two phonetic features [4, pp.204–207]. Firstly, the presence of junctural features, such as slight pauses, final lengthening and rhythmic discontinuities, can signal the end of a tone-unit. However, a pause does not necessarily correspond with a tone-unit boundary in spontaneous speech, particularly in cases of disfluency. Secondly, given that the first prominent syllable, for the majority of tone-units in an utterance, is of approximately the same pitch level [4], the boundary may be signaled by some perceivable pitch change. This change can be either a step up from a falling pitch movement, or a step down from a rising pitch movement. It may be difficult to identify such pitch resets when the tone-unit onset is low and the final accent of the previous tone-unit ends with a pitch fall, or the onset is high and follows a tone-unit whose final accent ends with a rise in pitch.

Table 1: Symbols for Sentential Stress and Pitch Movement Labelling

ASCII†	Symbol	Description
u	$\{u\}$	— Completely unstressed
s	$\{s\}$	— Stressed but unaccented
a	$\{a\}$	— Stressed and accented
n	$\{n\}$	— Nuclear accented
| *pipe*	$\{:\}$	— syllable immediately preceding a tone-unit boundary
\	$\{\backslash\}$	— pitch accent is a fall
/	$\{/\}$	— pitch accent is a rise
v	$\{\vee\}$	— accent is a fall-rise
^ *hat*	$\{\wedge\}$	— accent is a rise-fall
l	$\{-\}$	— level tone
<	$\{\leftarrow\}$	— pitch movement is part of the realisation of an accented syllable to the left of this syllable
>	$\{\rightarrow\}$	— pitch movement is part of the realisation of an accented syllable to the right of this syllable
- *minus*	$\{^{-}\}$	— the range of the pitch movement is unusually wide (increased)
_ *underscore*	$\{\_\}$	— the range of the pitch movement is unusually narrow (decreased)
' *apostrophe*	$\{\triangle\}$	— pitch "peak" or level tone pitch is unusually high
, *comma*	$\{\bigtriangledown\}$	— pitch "peak" or level tone pitch is unusually low
[	$\{\|\ \}$	— initial part of $\{\vee\}$ or $\{\wedge\}$ pitch movement is shallow
]	$\{\ \|\}$	— final part of $\{\vee\}$ or $\{\wedge\}$ pitch movement is shallow

†The ASCII characters listed are the prosodic labels used in machine readable data.

IV. PITCH MOVEMENT LABELLING

The pitch contour of an utterance is labelled as a series of pitch movements at (or near) each accented syllable. A pitch movement is either a continuous pitch glide, for example over a long vocalic section of speech, or a discrete pitch jump from one level to another over a series of syllables. Each pitch movement in an utterance is labelled as one of the five categories $\{\backslash, /, \vee, \wedge, -\}$ (see table 1).

A description is associated with each and every syllable labelled as accented (or nuclear accented) to mark the direction of pitch movement on this and any following unaccented syllables. These labels should only be time aligned with an unstressed or an accented (nuclear or otherwise) syllable $\{u, a, n\}$, but not with a stressed (but unaccented) syllable $\{s\}$. (Any stressed syllable corresponding with a time aligned pitch movement label should be marked as an accented syllable.) If the pitch movement is aligned with an unstressed syllable $\{u\}$, a diacritic is applied to the pitch movement label in order to indicate whether the pitch movement is part of the realisation of the nearest accented syllable $\{a, n\}$ to the left $\{\leftarrow\}$ or the nearest one to the right $\{\rightarrow\}$. There may be more than one pitch movement associated with an accented syllable; for example, if there is a rise-fall pitch movement in the realisation of an accented syllable but the rise occurs on a preceding unstressed syllable and the fall occurs on a succeeding unstressed syllable. The uses of these diacritics enable the inflections of the F0 contour to be described while maintaining a transcription of the perceived pitch movement.

Pitch range markings are used to describe the extent of the movement in a pitch glide and the distance between levels of a pitch jump, but not for level tone. If the pitch range is distinctively wider or narrower than expected for a particular contrastive effect, it is marked with a diacritic {‾, _} on the pitch direction labels. Diacritics are also applied to these labels if the "peak" part of a pitch movement (the initial part of a fall {\}, the final part of a rise {/}, and the mid-section of a fall-rise or rise-fall {∨, ∧}) or the pitch of a level tone {−} is unusually high {△} or low {▽} for the particular speaker. In order to describe occurrences of pitch fall-rise and rise-fall with a particularly shallow rise or shallow fall, two further diacritics are included. These are used to represent, for example, fall-shallow rise as {\/}.

V. ACOUSTIC FEATURES

A set of acoustic features must be extracted from the raw speech waveform in order to automatically identify syllable prominence and pitch movements. In our preliminary stages of producing an automatic prosodic labelling algorithm, eighteen features are used to describe what are we believe to be the acoustic correlates of stress (duration, intensity and fundamental frequency).

The energy and fundamental frequency of the speech waveform (sampled at 20kHz) are measured for 20ms frames of speech at 5ms intervals so that values are synchronised with the cepstral coefficients and lower three formant frequencies used in the auto-segmentation process. The fundamental frequency (F0) is determined using a slightly enhanced version of the pitch tracker described in [14]. In order to measure the signal energy, each frame is passed through a Blackman-Harris window and an amplitude spectrum is calculated using a 512-point FFT. The total energy for the frequency range of 50Hz–2kHz is determined by summation of the corresponding frequency bins. Each frame energy value is then expressed in decibels with respect to the maximum frame energy for the utterance. This process forms an utterance-normalised sonorant energy contour. Both the raw F0 contour and the energy contour are smoothed using a 3-point non-linear median filter and a 5-point hanning window [17].

The phone given by auto-segmentation which forms the nucleus of a given syllable is identified by the following procedure. The phones in the syllable are split into two groups on the basis of whether or not they are a member of the set of vocalic phones and potentially syllabic consonantal phones (currently, all vowels plus [l, m, n, r]). Each phone is associated with the maximum sonorant energy within its tenure. If there are phones in the syllable which are members of this set, then the one whose associated energy is greatest, is selected as the syllable nucleus. Otherwise, none of the syllable phones are [vowel, l, m, n, r] and the phone with the greatest maximum sonorant energy is selected. The duration associated with any syllable in determining its prominence is the duration of its nuclear phone – this will be referred to as the "syllable duration". Using the duration of the entire syllable or the duration of all consecutive sonorants in the syllable as this measure has not yet been investigated.

Each syllable in an utterance is characterised by the maximum soronant energy within its tenure (syllable energy), its "syllable duration", the maximum F0 value within its tenure, the F0 values at the beginning and at the end of the syllable, and an F0 slope in Hz per second which describes the rate of change in F0 through any voiced regions of the syllable. The syllable energy and "syllable duration" are Z-score normalised to eliminate phone-specific effects [1]. For each phone type, the mean and population standard deviation of the syllable energy/duration is determined. Then, for each token of that phone type, the syllable energy/duration is normalised by subtracting the mean and dividing by the population standard deviation. Hence, for each syllable, there are six acoustic features extracted – phone-normalised duration, phone-normalised energy, maximum F0, start-time F0, stop-time F0, and F0 slope. In automatically establishing the prominence of any syllable in an utterance, these six features for the current, previous and next syllable are used, giving a total of eighteen features per syllable.

Table 2: Confusion Matrix of Sentential Stress Labelling by Hand and by Automation – cyclic exclusion of each utterance during training

		Automatic Label			
		a,n	s	u	total
Hand Label	a,n	889 (12.3%)	72 (1.0%)	849 (11.7%)	1810 (25.0%)
	s	237 (3.3%)	72 (1.0%)	673 (9.3%)	982 (13.6%)
	u	567 (7.8%)	142 (2.0%)	3731 (51.6%)	4440 (61.4%)
	total	1693 (23.4%)	286 (4.0%)	5253 (72.6%)	7232 (100.0%)

Misclassification error rate = 2540/7232 (35.1%)

Table 3: Confusion Matrix of Sentential Stress Labelling by Hand and by Automation – all utterances used during training

		Automatic Label			
		a,n	s	u	total
Hand Label	a,n	1143 (15.8%)	44 (0.6%)	623 (8.6%)	1810 (25.0%)
	s	240 (3.3%)	113 (1.6%)	629 (8.7%)	982 (13.6%)
	u	334 (4.6%)	51 (0.7%)	4055 (56.1%)	4440 (61.4%)
	total	1717 (23.7%)	208 (2.9%)	5307 (73.4%)	7232 (100.0%)

Misclassification error rate = 1921/7232 (26.6%)

The F0 features are also normalised so that each movement is independent of its absolute F0 values. Our intuition suggests that F0 change is the significant factor, not the absolute F0 values. Normalisation of the nine F0 parameters (the maximum F0, start-time F0, and stop-time F0 for the current, previous and subsequent syllables), is performed by determining the minimum value of these parameters and subtracting it from each. The change in F0 through the syllables is therefore described independently of the absolute height of the F0 movement.

VI. APPLICATION OF A TREE-BASED STATISTICAL MODEL

The sentential stress and pitch movements associated with accented syllables have been hand labelled in the ATR database of 453 utterances using the symbols given in table 1. The prosodic transcription was done by only one labeller.

The automatic prosodic labelling algorithm is still in its infancy and so the acoustic features described in section V are being used only to identify any given syllable in an utterances as either unstressed, stressed or accented (nuclear or otherwise). Distinguishing pitch movement types has not yet been incorporated.

The acoustic features are used as parameters to a tree-based statistical model (using "S" [2]). The model is trained on all but one of the utterances in the database. The tree classifies each hand-transcribed sentential stress label on the basis of the features given. This tree is then used to predict the labels for the utterance that was not included in the training set. These automatically generated labels are compared with those given by hand. This process is repeated in a cyclic fashion for all the utterances and the comparisons are summed. The confusion matrix (table 2) indicates the number of occurrences that each hand-transcribed label is predicted as accented $\{a, n\}$, stressed $\{s\}$ or unstressed $\{u\}$ using this process.

In order to give an indication of the dependency of the automatic labels on the method used, table 3 shows a similar confusion matrix generated when the test utterance is included in the training data.

VII. DISCUSSION

The misclassification error rate of 26.6% is quite promising given that the selection of the acoustic features that have been used is based on intuition. This, however, may not be the only contributing factor to erroneous classifications. It could be that the acoustic features are in fact closely related to the prosodic events labelled, but that the tree-based statistical model is not the most appropriate method to classify these events given the acoustic features (this is supported by the considerable difference between tables 2 & 3). Alternatively, the acoustic features presented could be insufficient to characterise the prosodic events. For example, it is likely that representing F0 movements across a three-syllable window is restrictive, given that such movements can clearly span many or part of syllables. It may be that the labelling scheme is an inadequate system for describing sentential stress and the pitch movements as perceived by the transcriber. This can be illustrated by the fact that sentential stress is not a simple binary distinction between stressed and unstressed. In ambiguous cases, the transcriber uses linguistic knowledge not evident in the acoustics. For example, the syllable in question will be marked as sententially stressed only if it can be lexically stressed. This may lead to every occurrence of schwa being marked as unstressed regardless of the acoustic evidence. With such linguistic knowledge unavailable to the tree-based model, confusions will inevitably arise between the hand labels and automatic labels.

It is most likely that the classification errors are due to some combination of all these factors, although the extent to which any one factor effects the error rate is difficult to determine. The correct-classification rate of 73.4% is, however, close to the percentage of correlating labels between two hand labellers – in the prosodic labelling of the Lancaster/IBM spoken English corpus, transcribers achieved 72% agreement for seven categories of sentential stress labels $\{\backslash, /, \vee, \wedge, -, s, u\}$ and 83% agreement for the categories "accented"/ "stressed"/ "unstressed" [10].

ACKNOWLEDGEMENTS

Thanks to Keith Edwards, Sally Bates, Alex Monaghan, Nick Campbell, Jim Hieronymus, and Bob Ladd for their valuable assistance. This work has been supported by ATR Interpreting Telephony Research Laboratories, Kyoto, Japan.

References

[1] W.N. Campbell. Evidence for a syllable-based model of speech timing. In *Proc. International Conference on Spoken Language Processing*, volume 1, pages 9–12, Kobe, Japan, 1990.

[2] L.A. Clark and D. Pregibon. Tree-based models. In J.M. Chambers and T.J. Hastie, editors, *Statistical Models in S*, chapter 9, pages 377–419. Wadsworth & Brooks, Pacific Grove, California, 1992.

[3] E. Couper-Kuhlen. *An Introduction to English Prosody.* Edward Arnold (Publishers) Ltd., London, 1986.

[4] D. Crystal. *Prosodic Systems and Intonation in English.* Cambridge University Press, Cambridge, U.K., 1969.

[5] J.M. Dalby. *Phonetic Structure of Fast Speech in American English.* PhD dissertation, Indiana University Linguistics Club, Bloomington, Indiana, 1986.

[6] D.B. Fry. Duration and intensity as physics correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27(4):765–768, 1955.

[7] E.C. Fudge. *English Word-Stress.* George Allen & Unwin, London, 1984.

[8] E. Gårding and Gerstman. The effect of changes in the location of an intonation peak on sentence stress. *Studia Linguistica*, 14:57–59, 1960.

[9] A.C. Gimson. *An Introduction to the Pronunciation of English.* Edward Arnold, London, second edition, 1970.

[10] G. Knowles and P.R. Alderson, editors. *Working with Speech: The Computational Analysis of Formal British English Speech.* Longman, London, 1994.

[11] D.R. Ladd. Peak features and overall slope. In A. Cutler and D.R. Ladd, editors, *Prosody: Models and Measurements*, chapter 4, pages 39–52. Springer-Verlag, Heidelberg, Germany, 1983.

[12] I. Lehiste. *Suprasegmentals.* The Massachusetts Institute of Technology Press, Cambridge, Massachusetts, 1970.

[13] I. Maddieson. Phonetic cues to syllabification. In V.A. Fromkin, editor, *Phonetic Linguistics (essays in honor of P.Ladefoged).* Academic Press Inc., London, 1985.

[14] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. Signal Processing*, ASSP–39(1):40–48, 1991.

[15] J.D. O'Connor and G.R. Arnold. *Intonation of Colloquial English.* Longman, London, second edition, 1973.

[16] E. Pulgram. *Syllable, Word, Nexus, Cursus.* Mouton, Den Hague, 1970.

[17] L.R. Rabiner, M.R. Sambur, and C.E. Schmidt. Applications of non-linear smoothing algorithms to speech processing. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP–23(6):552–557, 1975.

[18] M.S. Schmidt and G.S. Watson. The evaluation and optimization of automatic speech segmentation. In *Proc. 2nd. European Conference on Speech Communication and Technology*, volume 2, pages 701–704, Genova, Italy, 1991.