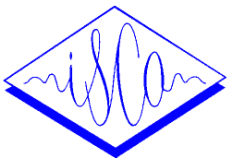

ITRW on
Speech and Emotion
Newcastle, Northern Ireland, UK
September 5-7, 2000

TECHNIQUES FOR THE PHONETIC DESCRIPTION OF EMOTIONAL SPEECH

Peter Roach

School of Linguistics and Applied Language Studies,
University of Reading, U.K.

p.j.roach@reading.ac.uk

ABSTRACT

It is inconceivable that there could be information present in the speech signal that could be detected by the human auditory system but which is not accessible to acoustic analysis and phonetic categorisation. We know that humans can reliably recognise a range of emotions produced by speakers of their own language on the basis of the acoustic signal alone, yet it appears that our ability to identify the relevant acoustic correlates is at present rather limited. This paper proposes that we have to build a bridge between the human perceptual experience and the measurable properties of the acoustic signal by developing an analytic framework based partly on auditory analysis. A possible framework is outlined which is based on the work of the Reading/Leeds Emotional Speech Database. The project was funded by ESRC Grant no. R000235285.

1. THE NEED FOR CODING

The detailed study of large amounts of emotional speech presents the researcher with two main requirements: one is a theory of emotions and their categorisation, and the other a method for transcribing those aspects of speech which are believed to be relevant in the signalling and recognition of emotions. This paper is concerned exclusively with the latter, though some of the work reported here has been concerned with a transcription system which entails both (Greasley et al, 1995 [1]).

In devising a transcription for the description of emotional speech, certain requirements need to be satisfied. The principal requirements are presented below.

(i) The transcription system must allow exhaustive and unambiguous coding of all features of speech which could possibly be relevant in the signalling and recognitions of emotions in speech.

(ii) The system should ideally use features which are capable of being defined in measurable acoustic terms.

(iii) Transcription, storage and retrieval of emotional speech data using the transcription system should be made as efficient and ergonomically practical as possible.

(iv) Inter-transcriber reliability should be given high priority.

(v) As far as possible, the transcription system should be compatible with existing systems.

In the following sections, we consider how the above requirements may be met.

2. PHONETIC FEATURES

The view of the territory in which we are working was for much of the twentieth century obscured by a persistent and highly misleading view of the role of prosody in speech. This view had two linked assumptions:

(i) intonation is used by speakers to convey their emotions and attitudes

(ii) intonation consists of variations in pitch which may be observed scientifically by measuring fundamental frequency.

It was recognised at least as early as the 1950's that this view was fundamentally incorrect, yet it persists in the contemporary literature. In the transcription of psychiatric interviews it was recognised that a description based solely on pitch variation was incapable of capturing the rich variety of relevant phonetic features (Trager, 1958 [2]; Pittenger et al, 1960 [3]). At the outset of the work on the Survey of English Usage, Crystal and Quirk (1964 [4]) recognised that a full transcription of the data that was likely to capture all relevant phonetic information must use an analytic framework that went far beyond the simple intonation descriptions of the time (e.g. the first edition of O'Connor and Arnold [5]). This point of view was set out in much fuller form in Crystal (1969, [6]). While at this time theories of prosodic and paralinguistic features were being developed in the context of linguistic description, other researchers were developing descriptive frameworks with a more sociological orientation. Laver (1968, [7]) is a good example of this movement; other work reprinted in Laver and Hutcheson (1972, [8]) also makes valuable reading. It is surprising that some of the most influential work of recent times in the field of prosody such as the work on Discourse

Intonation of Brazil et al (1980 [9]), on comparative intonation (Hirst and di Cristo, 1999 [10]) and on the Autosegmental approach to intonation (Silverman et al [11]) has to a large extent turned its back on the concepts and frameworks developed in earlier times despite the powerful arguments adduced for the incorporation of such information in an adequate descriptive framework. It should, however, be pointed out that Hirst and di Cristo go to some lengths to justify this procedure (pp. 3-7), and their discussion makes valuable reading. We return to this issue in Section 6 below.

In the next section we look in more detail at the range of features that are relevant in the description of emotional speech.

3. PROSODIC AND PARALINGUISTIC FEATURES

Any attempt to establish a framework for the phonetic description of emotional speech will inevitably run into the problem of distinguishing between *prosodic* and *paralinguistic* features of speech. It is unfortunate that though it is widely recognized that some such distinction must be made and some line must be drawn, the arguments for how the distinction should be made and for where the line should be drawn seem to vary from one work to another. The issue is in essence a simple one: we can agree that some components of what we may in a neutral way refer to as the suprasegmental aspect of speech should form part of the phonology of the language, while others, though still lying within the range of vocal variables that are under voluntary control, are clearly not phonological. As is often found in cases where we find it difficult to make a distinction which we feel is needed, it is most likely that we are dealing with a continuum. Let us look at some examples from each end of the continuum. Most people would probably agree that such phenomena as the intonational difference in English between a rising, a falling and a falling-rising tone form part of a system of contrasts that is near enough to the inter-phoneme contrasts of segmental phonology to count as phonological. A similar view would probably be taken of the “stress” contrast between ∪import (noun) and im∪port (verb). At the other extreme, the use of different voice qualities such as a “whining” quality used in pleading and a “forthright” quality used for giving orders, while clearly distinct and under the voluntary control of the speaker, could not form part of such a system. Knowing how to deal with the extremes of the continuum does not, unfortunately, equip us to deal with the extensive grey areas we find around the middle of the continuum, and taking a poll of published views does not seem to help. Brown (1990, [12]) defines paralinguistic features of speech as

> "those which contribute to the expression of attitude by the speaker. They are phonetic features of speech which do not form an intrinsic part of the phonological contrasts which make up the verbal message: they can be discussed independently of the sequences of vowels and consonants, of the stress patterns of words, of the stressing of lexical rather than grammatical words, and of intonation structure which determines where the tonic syllable falls" (p.112).

In this, Brown appears to be placing paralinguistic features outside the domain of linguistic contrasts, and later equates them with facial expression and body language. However, her list of paralinguistic features include many which are classed as prosodic by Crystal and Quirk. Crystal and Quirk *(op. cit.)* state

> "We are using the expressions "prosodic" and "paralinguistic" to denote a scale which has at its "most prosodic" end systems of features (for example, intonation contours) which can fairly easily be integrated with other aspects of linguistic structure, while at the "most paralinguistic" end there are the features most obviously remote from the possibility of integration with the linguistic structure proper (tremulous voice or clicks of annoyance, for example). Since, therefore, both expressions have this "more or less" character, there is no question of a sharp division between the two, and it would be prejudging the results of future careful research to make a clear-cut list of features undoubtedly playing a role in linguistic patterning and another list of features undoubtedly "beyond" the limits of describable linguistic structure." (p.12).

Crystal *(op. cit.)* refined this position further:

> "The distinction between 'prosodic' and 'paralinguistic' features of utterance can be made partly on phonetic and partly on functional grounds. From the phonetic point of view, *prosodic* features may be defined as vocal effects constituted by variations along the parameters of pitch, loudness, duration and silence. This then excludes vocal effects which are primarily the result of physiological mechanisms other than the vocal cords, such as the direct result of the workings of the pharyngeal, oral or nasal cavities: these are referred to as *paralinguistic* features." (p. 128).

Roach (2000, [13]) criticizes Crystal's definition:

> "This does not seem to me to fit the facts. In my view, `paralinguistic' implies `outside the system of contrasts used in spoken language' - which does not, of course, necessarily mean `non-vocal'. I would therefore treat prosodic variables as linguistic, and consequently part of intonation, while vocal effects like laughs or sobs are non-linguistic vocal effects to be classed with gestures and facial expressions"

At least, for the purposes of this paper, we need not concern ourselves with the phonological status of the variables we identify, since we can consider the establishment of a system for recording characteristics of emotional speech as entirely a phonetic exercise. In setting up the transcription system for our corpus of emotional speech, therefore, we did not give specific attention to questions of what was within and what was outside the normal range of linguistic contrastivity. We took as our starting point the framework presented in Crystal and Quirk *(op. cit.)* and subsequently elaborated by Crystal (1969, [6]), but made substantial modifications to it that will be explained in following sections.

4. SETTING UP A PRACTICAL NOTATION SYSTEM

Perhaps the biggest danger facing the researcher who wishes to annotate material is that of over-complexity. Crystal and Quirk *(op. cit.)* criticize the work of some of their predecessors:

> "(a) The degree of detail involved in the analytic procedure makes progress so slow as to preclude coverage of a sufficiently large corpus of spoken material to provide a reasonable statistical basis for descriptive statements.
> (b) The narrow transcription, reflecting the refined analysis, is too complex typographically and too difficult to read and analyse by reason of the indiscriminate massing of relatively irrelevant detail which obscure the basic patterns. (…)
> (c) There is no agreement as to the degree of delicacy to which the description of any given paralinguistic features should be taken. (…)
> (d) There is little perceptible order in the presentation of the observed phenomena in terms of their linguistic significance, and insufficient exemplification to show the extent of systematisation in the material
> (…)
> (f) There is much disagreement in method and terminology between the various authors. (…)
> (g) The terminology of description is insufficiently defined in terms of relatively objective data, acoustic or articulatory." (pp. 21-2).

Although these criticisms are undoubtedly valid, it has to be said that the eventual outcome of Crystal's work was not itself immune from some of the same criticisms, particularly with regard to points (a), (b) and (c) above.

One specific point of criticism made by Roach *(op. cit.)* of the Crystal system and many others is the failure to distinguish, within the range of prosodic features, between those which occur intermittently and those which are continuously present. It is proposed that we should distinguish between SEQUENTIAL - those components of intonation that are found as elements in sequences of other such elements occurring one after another (never simultaneously), and PROSODIC - components that are characteristics of speech which are constantly present and observable while speech is going on. Examples of the former are tones (pitch-accents), pauses and boundaries, while examples of the latter are such features as tempo, pitch range and loudness.

This distinction is observed in the system we have developed, in which we have as different components of a transcription (1) sequential prosodic information (based on the ToBI system), (2) other prosodic features, (3) paralinguistic features and (4) non-linguistic features. The principles of this system have been published in Roach et al (1998, [14]), and are summarized below.

4.1 The ToBI System for Intonation Transcription

The transcription system for our emotional speech corpus was intended to be multi-linear. It seemed advisable to use a system which would identify the sequential elements of the intonation while leaving continuous prosodic features for separate annotation. Since we were dealing with a corpus stored on computer which was to be transcribed on computer, we clearly needed a computer-readable system, and only two possibilities presented themselves to us. One was the British system used for the Spoken English Corpus (Knowles et al, 1995 [15]), later used in modified form for the computer-readable version named MARSEC (Roach et al, 1994 [16]). Contrary to a widespread belief, this system is significantly different from the "Standard British" model of intonation usually credited to O'Connor and Arnold (1973, [17]), in that it carries with it no concept of nucleus, head and tail within the tone unit. It simply marks every accented syllable with a tone mark selected from an inventory of ten, and separates intonation units with either a major or a minor boundary. This system satisfies many of the strictures on earlier transcription systems expressed by Crystal and Quirk *(op. cit.)*, being simple, easy to use, and tested and agreed between expert transcribers. In MARSEC, standard ASCII characters were used to represent the various tone marks within the text, though not always in an easily recognizable way. This system would have been quite suitable for use in our emotional speech corpus, but the attractions of the ToBI system seemed to outweigh it. For a critical review of ToBI (Tones and Break Indices), see Cruttenden (1997 [18]). This system, though designed originally for transcribing American English, has been used with apparent success for transcribing other accents of English, including British English, and for some other European languages. This is not a suitable place for a detailed account of ToBI, but some of its major points are the following:

- It was designed for use in computer-based transcription (specifically, in the Waves environment, which was used for our research).
- It was developed by a committee of experts whose transcriptions were carefully checked for consistency Pitrelli et al, 1994 [19]
- It is to some extent based on contemporary phonological theory (this is not relevant for our purposes).
- There is an excellent training package for researchers new to intonation transcription.

For our corpus we chose to use ToBI transcription to deal with the sequential elements. This provided us with a "skeleton" representation which marks the points at which important information-bearing prosodic events are located in relation to the time-course of the sound recording and the text. The prosodic events include the placement of pitch-accents (tones), and of boundaries between constituents (break indices). We did not know at the time of planning whether any relationship would be found between the occurrence of basic ToBI elements and emotions.

4.2 Representing other Prosodic Features

We begin by looking at the Crystal and Quirk framework, in which prosodic features comprise tempo, prominence, pitch range, rhythmicality, tension, pause and intonation. The features of tempo, prominence and pitch may vary in a *simple* way (i.e. they will have a specific value for a stretch of speech, such as *fast* or *slow* for tempo), or in a *complex* way (i.e. the variation is from one value of a feature, such as *crescendo* in loudness).

Tempo refers to speech rate and has two different manifestations according to whether it is perceived over polysyllabic stretches or on single syllables. On polysyllabic stretches, "simple" tempo, *i.e.* fast or slow speech, is divided into four marked speeds: *allegrissimo*; *allegro*; *lento*; and *lentissimo*. "Complex" tempo refers to accelerating and decelerating speech rate, and is again referred to in musical terminology as either *accelerando* or *rallentando*. On the single syllable, tempo is either *clipped* or *drawled*.

Features of **prominence** (perceived loudness) can again be either simple, that is quiet or loud, (*pianissimo*, *piano*, *forte*, or *fortissimo*) or complex (*crescendo* or *diminuendo*).

The description of **pitch range** on individual syllables is complicated by the effect of prominence and the intonation system, but on polysyllabic stretches and on the nuclear syllable is either simple (*low* or *high*) or complex (*monotone*, *narrow* or *wide*).

Rhythmicality depends on three discrete factors: perceived regularity of stresses, being either *rhythmic* or *arhythmic*, sharpness of pitch variation and prominence, classified as either *spiky* or *glissando*, and variation in prominence alone without reference to pitch (*staccato* or *legato*).

Tension refers to the precision of articulatory gestures and may be either *slurred*, as in drunken speech, *lax*, *tense*, or *precise*.

Pauses may be silent or voiced and are classified on perceived deviation from a speaker's norm as having four degrees of duration: *brief*; *unit*; *double*, or *treble*.

The above scheme was considered too complex for large-scale transcription work on our corpus, and instead the following scheme was adopted:

Pause
: Crystal & Quirk's categories of pause are replaced by the Break Index tier of the ToBI transcription.

Pitch
: High / low
Wide / narrow.
These annotations are used for pitch range, while as mentioned above, ToBI notation is used to transcribe pitch direction and the pitch of accented syllables.

Loudness
: Loud / quiet
Crescendo / diminuendo.

Tempo
: Fast / slow
Accelerating / decelerating
Clipped / drawled (on single syllables).

The above features were transcribed on separate transcription tiers from the ToBI transcriptions. The format of the transcriptions is described later.

4.3 Representing Paralinguistic Features

This requires conventions for two things: quasi-continuous aspects of voice quality, and intermittent vocal effects. The former has been the subject of much research in recent decades, and we should begin by recognizing that much has changed since the time of publication of the Crystal and Quirk monograph *(op.cit.)*. In particular, work by Laver on voice quality has greatly enhanced our understanding of this aspect of speech (1980, [20]).

Paralinguistic features were divided by Crystal & Quirk into two types: voice qualities and voice qualifications. Voice qualities were said to be due to different modes of phonation: *normal voice*, *falsetto*, *whisper*, *creak*, *huskiness*, and *breathiness*. Laver's taxonomy of simple modes of phonation overlaps considerably with that of Crystal & Quirk and consists

of: *modal voice*, *falsetto*, *whisper*, *creak*, *harshness*, and *breathiness*. However, Laver's description of voice quality includes not only these various modes of phonation, but all the variation, intended or unintended, a speaker is capable of, given physiological restrictions of the vocal organs, taking into account the longitudinal articulatory settings (raised and lowered larynx voice, labial protrusion, and labiodentalised voice), latitudinal settings (labial, lingual, faucal, pharyngeal, and mandibular settings), velopharyngeal settings, and phonatory settings (including compounds of modes of phonation). The categorisation used for our corpus has expanded somewhat on those of Crystal & Quirk and of Laver:

Falsetto
Creak
Whisper
Rough
Breathy
Ventricular
Ingressive
Glottal attack.

The categories *ventricular*, *ingressive*, and *glottal attack* have been added because it was felt they are needed to describe adequately the vocal effects of the data. In accordance with Ladefoged (1971 [21]) and Laver *(op.cit.)*, the terms *creak* and *creaky voice* are used synonymously. Laver's compound phonation types are not explicitly adopted, but the Waves system of time-linked labels makes it possible to transcribe simultaneously occurring voice quality effects by the use of overlapping transcriptions.

We also need to transcribe some intermittent vocal effects that we put under the general heading of *fluid control and respiratory reflexes*. Reflexes are often an involuntary indication of genuine emotional stress. Extreme emotional states produce altered patterns in respiration, the endocrine system, and the metabolism in general. As a response to such changes, reflex behaviour can occur in the vocal tract. For the purposes of our study, the following reflexes are considered relevant, both as conscious signalling of and as involuntary reaction to emotional arousal:

Clearing the throat
Sniff
Gulp
Audible breathing
Yawn.

There exists the possibility that such reflexes are not involuntary, but are being consciously used to convey a particular emotional state.

We also make use of Crystal and Quirk's notion of *voice qualifications*. Under "voice qualifications", we include the following terms:

Laugh
Cry
Tremulous voice.

This is a simplification of Crystal & Quirk's system, including *giggle* under *laugh,* and *sob* under *cry*.

5. A MENU OF FEATURES FOR TRANSCRIPTION

Having worked out the above framework, it was necessary to implement it in the form of a useable transcription system. We were committed to working with on-screen computerised transcription, and the Waves environment allows for multiple transcription windows time-linked to the sound recording and the text. Windows for the acoustic signal, the fundamental frequency trace and the transcribed text of the utterance are obvious requirements in the screen display. In addition to these, three windows are normally used for ToBI transcription: one for marking pitch-accents, one for break indices and one for notation of non-linguistic information (e.g. indicating extraneous noises on the recording). An additional window was used in our work for coding other features. The prosodic and paralinguistic features are seen as having points of time at which they start and finish, possibly overlapping with each other, and it is necessary to signal both the beginning and end points. The menu shown in Table 1 was devised for use in our project, from which selections could be made by mouse.

This menu system has proved to be quick and convenient to use. The complex and multi-level transcriptions were stored, and were then available for study and statistical processing. Techniques for doing this using UNIX tools are described by Stibbard (2000, [22]).

6. DEFINING THE MEANINGS OF FEATURE LABELS

It is necessary to conclude by discussing one of the most difficult aspects of phonetic labelling. We need to answer the question of whether the features we use refer to objectively measurable physical properties, or are impressionistic labels that depend on the analyst's intuitions. There is a historic tension between these two, and the issue is carefully dealt with in Ladd (1996 [23]). Ladd is writing from the perspective of a writer addressing the issue of the phonological status of representations of prosody, and thus has a different perspective from the present paper. He points out that the search for acoustic correlates of intonational categories has been something of an unnecessary distraction in past work. He writes:

> "Because of the general lack of agreement and the notable absence of instrumental evidence for impressionistic descriptions, adherents of the instrumental approach have often felt that their work is somehow more rigorous or more scientific, or at the very least more complete. (…) Such criticisms largely miss the point. The difference between the two approaches is not primarily one of methodology, nor one of completeness, but of theoretical assumptions. …

START	END
high(	high)
low(	low)
wide(	wide)
narrow(	narrow)
loud.(	loud.)
quiet.(	quiet.)
crescendo.(	crescendo.)
diminuendo.(	diminuendo.)
fast:(	fast:)
slow:(	slow:)
accelerating:(	accelerating:)
decelerating:(	decelerating:)
clipped:(	clipped:)
drawled:(	drawled:)
precise:(	precise:)
slurred :(	slurred:)
falsetto{	falsetto}
creak{	creak}
whisper{	whisper}
rough{	rough}
breathy{	breathy}
ingressive{	ingressive}
+nasal{	+nasal}
-nasal{	-nasal}
glottal-attack{}	
clear-throat[	clear-throat]
sniff[	sniff]
gulp[	gulp]
click[]	
breath-in[	breath-in]
breath-ex[	breath-ex]
breath[	breath]
laugh:[	laugh:]
tremulous:[	tremulous:]
cry:[	cry:]
yawn:[	yawn:]

TABLE 1: Menu for labelling prosodic and paralinguistic features.

> "It is important to recognise that the impressionistic descriptions involve phonological categories that could *in principle* be related to instrumentally validated acoustic or articulatory parameters." (p.13)

In the experimental study of the relationships between emotion and features of speech, it seems obvious that for most purposes there would be no point in using phonetic labels that could not in principle be related to physically measurable parameters. Yet in fact it is clear that in much research we are compelled either to work with a very impoverished set of descriptive terms which can be reliably calculated without the need for human judgment to intervene (e.g. long-term average spectrum; pitch range), or to use a richer representation in which some of the terms are only *in principle* definable acoustically, and are in reality at present only usable with human intervention. It is, for example, extremely unlikely that a competent ToBI transcription could be successfully carried out unaided by a computer, though reasonably accurate recognition of accented syllables, pauses and specific tones by neural networks or HMM's has been demonstrated by various researchers. In the case of other prosodic features introduced above, some are easier to measure than others. Absolute speed of utterance (tempo) can be measured, either after identification of segments in the utterance, or when the text of the utterance is known (Arnfield et al [24]), but complex features such as *accelerating* would be much harder to recognize reliably. Loudness (intensity) can be measured, but since this is so heavily dependent on the recording environment it is of limited reliability.

Differences in voice quality are perhaps the most difficult to define. There have been attempts at physical definitions in the past (e.g. Catford 1994 [25], Laver 1980 [20]); the most comprehensive framework for description is to be found in Laver (1994 [25]). While there is some hope that such features may be automatically measurable and detectable in the future, that certainly does not seem to be the case at present.

Most researchers who work in the field of emotional speech do so in the hope that eventually it will be possible to produce convincing-sounding synthetic speech with appropriate emotional characteristics, and that automatic speech recognition systems will be able to detect the emotional state of the speaker. The way for us to approach that goal is to use phonetic descriptions which can be reliably used by human analysts and attempt to work towards systems which are capable of being trained to recognize and synthesize the relevant features. These must form the bridge which connects the traditional impressionistic descriptions of speech with the objective measurements and categorization that we need.

7. REFERENCES

[1] Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P. , Arnfield, S. and Horton, D. (1995) 'Representation of prosodic and emotional features in a spoken language database', *Proceedings of the XIIIth International Congress of Phonetic Sciences,* Stockholm, Vol. 1 , 242-245.

[2] Trager, G.L. (1958) 'Paralanguage: a first approximation', *Studies in Linguistics* **13**: 1-12.

[3] Pittenger, R.E, Hockett, C.F. and Danehy, J.J. (1960) *The First Five Minutes,* New York: Ithaca.

[4] Crystal, D. and Quirk, R. (1964) *Systems of Prosodic and Paralinguistic Features in English,* Mouton.

[5] O'Connor, D.J. and Arnold, G.F. (1961) *The Intonation of Colloquial English*, (First Edition) Edward Arnold.

[6] Crystal, D. (1969) *Prosodic Systems and Intonation in English*, Cambridge University Press.

[7] Laver, J. (1968) 'Voice quality and indexical information', *British Journal of Disorders of Communication,* **3**, 43-54.

[8] Laver, J. and S. Hutcheson (1972) *Communication in Face-to-Face Interaction,* Penguin.

[9] Brazil, D., Coulthard, M. and Johns, C. (1980) *Discourse Intonation and Language teaching,* Longman.

[10] Hirst, D. and di Cristo, A. (1998) Intonation Systems, Cambridge University Press.

[11] Silverman, K., Beckman, M.E., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992). "ToBI: A standard for labeling English prosody". Proceedings of the Second International Conference on Spoken Language Processing, 286-270.

[12] Brown, G. (1990) *Listening to Spoken English*, (Second Edition), Longman.

[13] Roach, P. (2000) *English Phonetics and Phonology* (Third edition), Cambridge University Press.

[14] Roach, P. , Stibbard, R., Osborne, J., Arnfield, S. and Setter, J. (1998) 'Transcription of prosodic and paralinguistic features of emotional speech', *Journal of the International Phonetic Association,* **28**, 83-94.

[15] Knowles, G., Alderson, P. and Wichmann, A. (1996) *Working with Speech*, Longman.

[16] Roach, P.J., Knowles, G.O., Varadi, T. and Arnfield, S.C. (1994) 'MARSEC: A MAchine-Readable Spoken English Corpus', *Journal of the International Phonetic Association,* vol.23.2, pp. 47-54.

[17] O'Connor, D.J. and Arnold, G.F. (1973) *Intonation of Colloquial English* (Second Ed.)*,* Longman.

[18] Cruttenden, A. (1997) *Intonation* (Second Edition), Cambridge University Press.

[19] Pitrelli, J., Beckman, M.E. and Hirschberg, J. (1994) 'Evaluation of prosodic transcription labeling reliability in the ToBI framework', *Proceedings of ICSLP 1994*, Yokohama, 1, 123-6.

[20] Laver, J. (1980) *The Phonetic Description of Voice Quality*, Cambridge University Press.

[21] Ladefoged, P. (1971). *Preliminaries to Linguistic Phonetics*. Chicago: University of Chicago Press.

[22] Stibbard, R. (2000) (this volume)

[23] Ladd, D.R. (1976) *Intonational* Phonology, Cambridge University Press.

[24] Arnfield, S., Roach, P., Setter, J., Greasley, P. and Horton, D. (1995) 'Emotional stress and speech tempo variation', *Proceedings of the ESCA-NATO Workshop on Speech under Stress,* Lisbon, pp.13-15.

[25] Catford, J.C. (1964) 'Phonation types', in D. Abercrombie et al *In Honour of Daniel Jones,* Longman, 26-37.

[26] Laver, J. (1994) *Principles of Phonetics*, Cambridge University Press.