LEVELS OF LABELLING

W.J. BARRY and A.J.FOURCIN

Abstract

A multi-level, multi-tiered approach to the labelling of speech recordings is proposed. Five levels of segmental labelling are defined, with differing degrees of abstraction from the acoustic signal, ranging from the physical level of acoustic parameters to the segmental sound structure of the citation forms of the words in the utterance. The issue of prosodic labelling is also addressed and related to the levels suggested for segmental labelling.

1. Introduction

The present widespread initiatives (UK, Europe, USA, Japan) in speech database recordings rely on labelling for their full theoretical and applicational exploitation. Although labelling issues have been under discussion for some time now (Autesserre et al., 1989; Williams, 1987; Hieronymus et al., 1990), the wider theoretical implications of speech-data labelling have not been addressed. This discussion is a first attempt to do so. It also tries to state clearly in terms acceptable to both linguists and engineers which concepts are important for establishing a theory of labelling.

Terminological discrepancies (both in the sense of different words being used for the same concept and the same word being used for different concepts) are probably the source of most misunderstandings between engineers and linguists, and between linguists working in and around speech and language technology. This discussion cannot escape from "terminology", but it attempts to define terms in a way which might lead to a consensus preventing the automatic interpretation of the terms from the point of view of a particular "school". Speech technology is a highly practical area of research that borrows from many but still needs to define its own theories. The terms introduced in this discussion are defined against a background of computer technology. They are based on constraints such as the computer representation of the lexicon, and computer memory rather than on present-day phonological theories and possible mental representation of linguistic structure.

2. General principles

a. "Labelling" of speech recordings is the temporal definition and naming of parts of an utterance with reference to the acoustic signal. These "parts" may be temporally discrete or over-lapping, and may be defined in acoustic, phonetic or higher level linguistic terms.

b. It should always be borne in mind that all labelling is an abstraction derived

from a particular analytic and theoretical standpoint.

c. The purpose of such labelling is to enable the research and applicational exploitation of database recordings in the development of more advanced speech technology systems and the furthering of explicit speech knowledge.

d. The relation between acoustic events in the signal and any of the many possible linguistic representations of an utterance is complex, non-linear, and only partially understood.

e. Different levels of labelling will enable any utterance to be temporally defined within different theoretical frames of reference, and these theoretical frames to be cross compared by means of their common reference, the acoustic signal.

f. Each level may have a number of different "tiers", in that separable aspects of the same labelling level can be individually represented.

g. To provide maximum potential for future use, database labelling should be:
i) transparent, in that the criteria applied are explicit
ii) flexible, in that new representational tiers can be specified
iii) extendable, in that new phenomena can be incorporated in each tier

3. Labelling levels

This section will illustrate the concept of "labelling level" with a number of examples.

3.1 Physical Level

Definition: Any labels that are defined solely with reference to physically defined properties or events in an utterance.

Discussion: This level is most clearly in need of separate tiers of representation, since each physical representation is based on a particular data-acquisition or -analysis procedure, and is theoretically separate. Thus, data acquired with nasality detectors or with the use of palatography (providing two-dimensional information on tongue-palate contact) would sensibly provide different tiers of labelling, one giving [± NASAL] segments (or conceivably more degrees of nasality), the other giving spatially defined categories of tongue contact. Acoustic parameters are likely to be the most frequently used representations, since the acoustic speech signal is primary to all speech technology applications. The three dimensions of the acoustic speech signal: time, frequency, and amplitude can all be used, individually or in combination, or indeed in many different transformations, as a basis for defining events which can be categorised and used as acoustic labels. It depends on the functional aims of the labelling whether each such dimension, transformation, or combination is defined as a separate tier.

The events are located in time, as overlapping or discrete portions of the signal. Different analyses provide different divisions of the signal.

An example utterance is described in table I and illustrated in figure 1.

Table I Example of acoustic breakdown of utterance "twelve times ten", and resulting segmentation.

	Acoustic event	No. of segments
"Twelve times ten"	- periodicity:	3
	- high frequency noise	5
	- nasality(pole-zero analysis)	2
	- silence	4 (onset, offset +2)
	- broad-band impulses	3
	- spectral shifts:	12 - 16 (depending on resolution)

The choice of acoustic descriptors exemplified in the above example is, in practice, always the product of the analyst's view of what is likely to be useful in differentiating speech sounds. It reflects his/her knowledge about the acoustic structure of speech. Although the above analysis is expressed in general terms, only "nasality" being exclusive to a speech signal, the acoustic events can be defined in a manner which is much more closely linked to general phonetic description. In such cases the labelling would be compatible with parametric descriptions of speech in the tradition of Jespersen's analphabetic transcription (Jespersen,1920) and recent work by Browman and Goldstein (1986), which are like "orchestral scores" of articulator movements, or Kelly and Local's approach (cf. Kelly & Local, 1989 for a general discussion of phonetic representation), which specifies the domain of particular (articulatory or acoustic) properties independent of the segmental structure.

3.2 Acoustic Phonetic Level

Definition: Any label that describes events in the acoustic signal in terms of general phonetic descriptors.

Discussion: Categories that could be used at this level are e.g.:
Stop closure (voiced or voiceless)
Stop release
Aspiration
Fricative noise
Glide
Nasal
Glottal onset, offset, irregularity
Back, front, close, open vowel etc.

Figure 2 shows an example utterance "in arithmetic" labelled in this way. The acoustic segments, in chronological order are:

1. glottal onset, 2. front half-close vowel, 3. nasal, 4. central vowel, 5. glide, 6. front half-close vowel, 7. voiced broad-band fricative, 8. voiceless broad-band fricative, 9. devoiced nasal, 10. central vowel, 11. voiced stop closure, 12. voiceless stop closure, 13. release burst, 14. aspiration, 15. front half-close vowel, 16. glottal offset, 17. stop closure, 18. release burst, 19. aspiration

Note that none of these terms makes any claims about the linguistic function or distinctiveness of the events identified.

However, an economical way of labelling such acoustic phonetic events with a view to linking their occurrence to other levels of labelling, is to define the assumed "phonemic" identity being signalled by the phonetic event and specify it further with the acoustic property. This is the course chosen for the initial labelling of the speech material recorded within the UK National Speech Database project (SCRIBE) since it allows the automatic derivation of "narrow phonetic" and "broad phonetic" labelling (see sections 3.3 and 3.4 below) from the acoustic-phonetic segment string. Mixing levels in this manner is a purely practical measure, and does not affect the acoustic-phonetic basis of the segment definitions. However, it does also retain a differential identity, necessary for analysis purposes, of segments that might be considered acoustically identical, such as stop-closure silence or stop-closure voicing, but which are known to have different durational characteristics depending on place of articulation and context (Lehiste, 1970).

Note also, that the acoustic definition of the events is often dependent on more than one parameter and that the events are discrete. The primary criterion for determining a stretch of the speech signal for the allocation of an acoustic phonetic label is acoustic homogeneity. However, it must be stressed that the definition of beginning and end of a stretch of signal is ultimately arbitrary in most cases. In fact, sound categories such as semi-vowels and diphthongs involve spectral change by their very nature, and "homogeneity" in such cases is stretched to cover a "sameness of change", such as a rising or falling second formant (for /w/ or /j/ for example). Such "acoustic-phonetic" labels cannot be determined without knowledge of what the "sound" is in phonetic-phonological terms. This need introduces the conceptual basis for the next more abstract level, at least as far as present-day phonetic description is concerned, namely "narrow phonetic labelling."

3.3 "Narrow Phonetic Labelling" level

Definition: Labels that characterise the phonetic quality of speech sounds in terms of a set of phonetic transcription symbols such as IPA or its computer-

compatible equivalent[1].

Discussion: The placement of a phonetic label in relation to the acoustic signal will sometimes coincide with one "acoustic-phonetic" segment (e.g. a vowel, a fricative, a nasal etc.) but will correspond to a grouping of acoustic-phonetic segments in other cases (e.g. a plosive consonant, an affricate, a partially devoiced sonorant following /p,t,k/).

One major theoretical problem with this level of labelling is the need for a categorical decision on the part of the labeller whether or not speech sound is "present". The ultimate arbiter is the labeller's perceptual impression. Unfortunately, the symbol string cannot easily indicate the relative prominence of a particular sound. However, it does differentiate as much as possible with respect to properties modifying the base sound (i.e. with diacritics indicating e.g. voicing/devoicing, nasality, rounding, spreading, etc)

Even more than is the case with "acoustic-segment" labelling, the beginning and end-points of "narrow-phonetic" label segments are arbitrary. The perception of a particular sound is the product of acoustic properties contained both within a stretch of the signal that is given the label and in the signal surrounding that stretch. A particular stretch of signal is usually selected to represent the sound because it is considered to contain properties that are primary cues to hearing that sound. This principle is most true of continuant sounds such as vowels, fricatives, nasals etc. Plosives, affricates, and other complex sounds have no such "core" properties, and the best compromise is sought in defining the parts of the signal that "belong" to such sounds. For example, the stop closure, the burst, and the prevocalic aspiration of a /p/, /t/ or /k/ are included, but not the pre-closure transitions. The, perhaps debatable, reason is that the transition signals the place of articulation not the stop nature of the sound. Why, then, is the aspiration included, since that will signal the identity of the following vowel? At this stage, practical needs come to the aid of theoretical argument: the vowel is identifiable without the aspirated transition (in fact, if preceded by the aspiration, the vowel sounds as if it is preceded by some sort of consonant), whereas the differentiation of initial /p, t, k/ and /b, d, g/ depends quite critically on the aspiration.

Theoretical difficulties for narrow phonetic labelling can occur in cases where a perceived sequence of sounds does not match a corresponding sequence of acoustic phenomena.

The utterance "Point zero" (see figure 3) is heard as [pʰOIn?zI@r@U] but examination of the [n?z] part of the signal reveals that the nasal segment is divided into 2 parts by the glottal irregularity represented by [?]. That should

[1] The transcription symbols used in this paper are based on those agreed within the ESPRIT Project 2589 (SAM) as computer-compatible equivalents of the IPA symbols usually used to represent the distinctive sounds of English (cf. Fourcin et al. 1989).

call for a labelling sequence [n?nz] or possibly [?n], but this does not correspond, even in very careful auditory examination of the utterance, to what is heard. There is no theoretically sound solution to the problem at this level of labelling; it requires a more abstract level which relates the acoustic reversal to the correct structural sequence.

3.4. "Phonemic" labelling level

Definition: Labels that represent the functionally distinctive sound units of the language.

Discussion: All Speech Technology applicational goals underlying the labelling of a speech corpus are directed towards either identifying the words and their grammatical structure from the signal (= recognition tasks), or producing as natural or intelligible as possible a signal from a string of words (= synthesis). The so-called "phonemic" structure is one traditional way of representing the distinctive sound shape or citation form of words, and it is for this reason that "phonemic labelling" has been popularly regarded as the sine qua non of speech database labelling; it is the "mediator" between the signal and the lexicon. However, actual labelling practice has not generally been of the "citation phonemic" kind (see 3.4.1 below).

The theoretical and practical aspects of "phonemic" labelling should not be overlooked:

1. The "phonemic" structure of a word is an analytic construct, derived from the auditory comparison of carefully pronounced single words.

2. The distinctive sounds making up the "ideal" pronunciation of a word may not be pronounced when the word is spoken by a particular person on a particular occasion. This may well be the case even if the word is spoken in isolation (e.g. "Reckon" is likely to be pronounced ["rek=N], with a syllabic velar nasal, whereas its ideal "phonemic" form is /"rek@n/), with a schwa + alveolar nasal, but it is much more likely in continuous speech). Variants like this are very well known; after all, they underlie many of the problems of continuous speech recognition and natural speech synthesis. A phonemic level of labelling cannot cater for variants of this type, however. A speaker does not choose between them as he or she might choose between two forms of the word "either"; it is an unconscious variation that is dependent on the situation and style of speech.

3. The relation between the "phonemic" structure and the speech signal, and therefore between the phonemic structure and other levels of labelling is very important in speech technology for two reasons:
i) as a mediator for lexical access
ii) as a source of knowledge underlying the development of phonological rules

4. A conventionalised "phonemic" or "citation form" level of labelling should be agreed which is "fitted" to the phonetically labelled signal. This "fitting"

process will often entail the allocation of several phonemes in a string to one phonetic segment (in cases of elision, or unsequenced acoustic events such as the example of [n?nz] shown in figure 3).

Example: "Shut up and come"

Phonetic	[SVt Vp =m kVm]
Phonemic	/SVt Vp {nd kVm/

5. The decision on the citation phonemic form of words cannot always be automatic. Dictionaries may give alternative pronunciations, either in the form of freely selectable alternatives, or as stylistic variants. For example, "either" as /i:D@/ or /aID@/; "inventory" as /"Inv@ntri/ or /In"vent@ri/; "lenient" as /"li:nj@nt/ or /"li:ni@nt/. The "fitting" process must therefore take into consideration which underlying form the speaker intended.

3.4.1 A variety of labelling, intermediate in its level of "abstraction" between "narrow phonetic" and "citation phonemic" has been commonly used in UK speech database labelling work, and is often referred to as "phonemic labelling". This is a level which only employs symbols that have a phonemic status (i.e. they distinguish words in English: e.g. /m/vs/n/: map vs. nap; /p/vs/t/; carp vs. cart etc), but uses them to indicate non-phonemic, i.e. continuous-speech phenomena such as the reduction and assimilation of "and" illustrated in point 4 of section 3.4.

This level is most economic in that it maximises phonetic information with minimal symbol complexity. To distinguish it from "pure" (citation) phonemic labelling, we will call it "Broad Phonetic". This is in line with the terminology agreed by the eight partner countries in the ESPRIT Speech Technology Assessment (SAM) project.

Examples:	good boy	in bed	bread and butter
phonetic	[gUb bOI];	[Im bed]	[bred @m bVt@]
vs. phonemic	/gUd bOI/;	/In bed/	/bred {nd bVt@

The purpose of this level of labelling is to capture the sequence of actually produced speech sounds in terms of the categories that are used to represent words in the lexicon. For purposes of recogniser training or assessment, it offers a means of reducing the level of abstractness in the symbol representation while retaining the same limited inventory. This is, of course, at the expense of greater phonetic (and therefore also acoustic) variation across the tokens of any one category compared to narrow phonetic labelling. But such differences in variation must be seen in terms of relative degree, not of presence vs. absence. Assimilatory processes that occur in continuous speech are not categorical changes, and cannot be captured in any absolute sense however narrow the phonetic representation attempts to be. For example, the [m] in [Im bed]

possibly has [n] resonances as a result of double articulation (bilabial + apico alveolar) despite the clear /m/ signal in the closing transitions of the /l/. In "often make" [Qf=m melk], the syllabic nasal [=m] possibly starts with labio-dental articulation following the /f/ and becomes bilabial under the influence of the following /m/.

To conclude this discussion of levels of segmental labelling, figure 4 gives the example utterance "pin prick" showing the relationship between the citation phonemic, broad phonetic, narrow phonetic, and acoustic-phonetic levels of labelling.

3.5. Prosodic Labelling

Definition: Labels that define aspects of an utterance that extend beyond the bounds of a single segment, such as the relative stress level of a syllable, the intonational pattern of one or more syllables, the rhythm of an utterance or parts of an utterance due to stress placement and pausing.

Discussion: Prosodic labelling cannot be considered one single level of labelling, since it can be represented at all four levels discussed above. Prosodic labels can be defined as a separate tier at each level. There are several physical parameters which can capture aspects of prosodic structure (Fx/F0; durational relations, intensity relations, stretches of silence); acoustic phonetic phenomena can be located (rising F0, falling F0, F0 peak, pause); phonetic events can be identified (rise, fall, fall-rise, stress); functional categories, equivalent to the phonemic level of segmental labelling can be defined (tone unit, nuclear tone, sentence accent, pitch accent, juncture); at this level, of course, the categories depend totally on the prosodic theory subscribed to by the labeller.

As yet, no conventions have been agreed and applied, either in the UK or internationally, for the prosodic labelling of speech recordings. However, the need for prosodic labelling is generally accepted, and urgent consideration of the following points is required:

1. What aspects of prosodic structure can be specified as potential illuminators of the complex relationship between the speech signal and higher-level (syntactic, semantic, and pragmatic) aspects of speech and language? It is clear that the answers to this will vary considerably with theoretical persuasion.

2. Can the properties of the speech signal involved in prosodic structuring be characterised for labelling purposes in a way which will facilitate cross-language comparison while still allowing language-specific analysis at the functional level?

3. How can these aspects be defined in terms of the levels described for segmental labelling? The complex relationships can best be quantitatively modelled if the formalisms used for prosodic labelling are as "theoretically transparent" as possible, i.e. with as precise a definition of label criteria as possible.

4. Conclusions

The above discussion of labelling issues within a proposed multi-level, multi-tiered approach is intended as a first step towards a theoretical framework for labelling. For the first time, the speech sciences are able to address theoretical and applicational questions using large speech corpora, and for the first time, the descriptive adequacy of many speech and language theories can be compared directly in quantitative terms. However, to achieve this aim, and to exploit the present speech database intitiatives to the full, an open, flexible, yet stringently defined set of labelling criteria needs to be developed and applied.

5. References

Autesserre, D., Pérennou G., Rossi M. (1989) "Methodology for the transcription and labelling of a speech corpus", J. Int. Phon. Assoc.

Browman, C.P. and Goldstein, L.M. (1986) "Towards an articulatory phonology", Phonology Yearbook, 3, 219-252.

Fourcin, A.J., Harland, G., Barry, W., Hazan, V. (1989) Speech Input and Output Assessment. Multilingual Methods and Standards, (Ellis Horwood Ltd, Chichester).

Hieronymus J., Alexander H., Bennett C., Choen I., Davies D., Dalby J., Laver J., Barry W., Fourcin A., Wells J., (1990) "Proposed speech segmentation criteria for the SCRIBE project", SCRIBE-Project Report

Jespersen, O. (1920²) Lehrbuch der Phonetik, (Teubner, Leipzig).

Kelly, J. and Local, J.K. (1989) Doing Phonology. (Manchester University Press, Manchester).

Lehiste, I. (1970) Suprasegmentals, (MIT Press, Cambridge).

Williams B., Dalby J., (1987) Segmentation Criteria for EUSIP Data Base. CSTR-Internal Report, Edinburgh

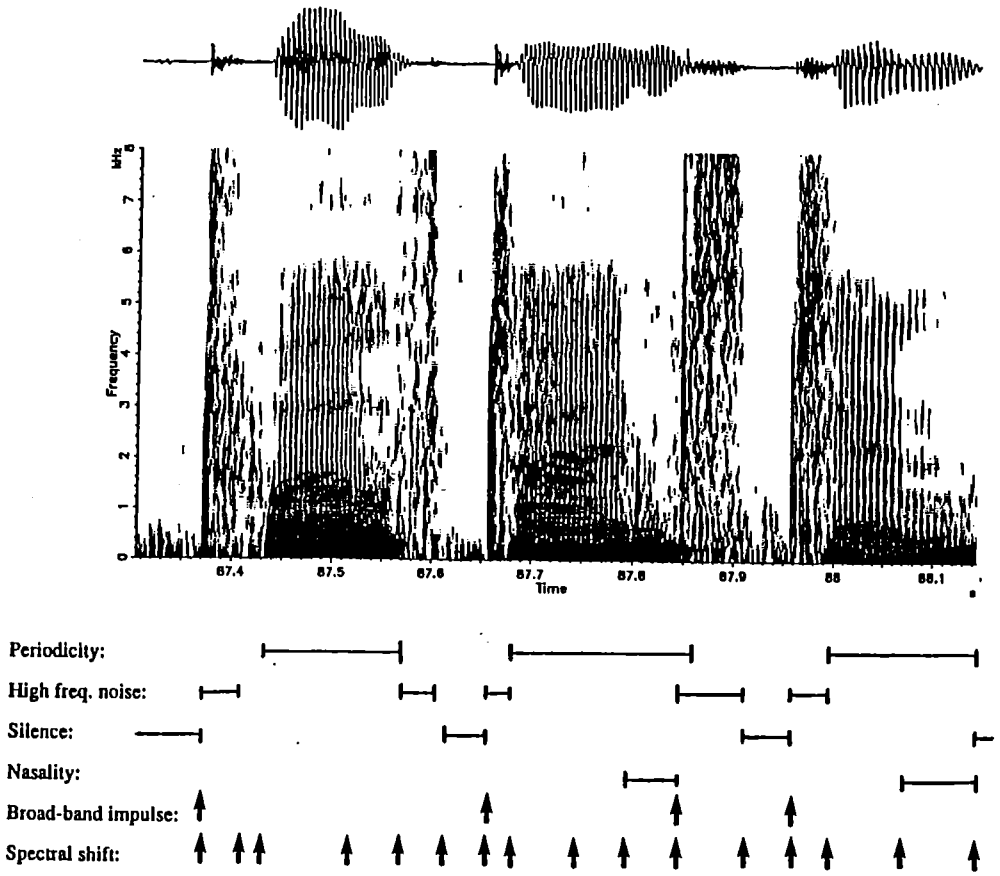

Figure 1 Acoustic event labelling of the utterance "Twelve times ten"

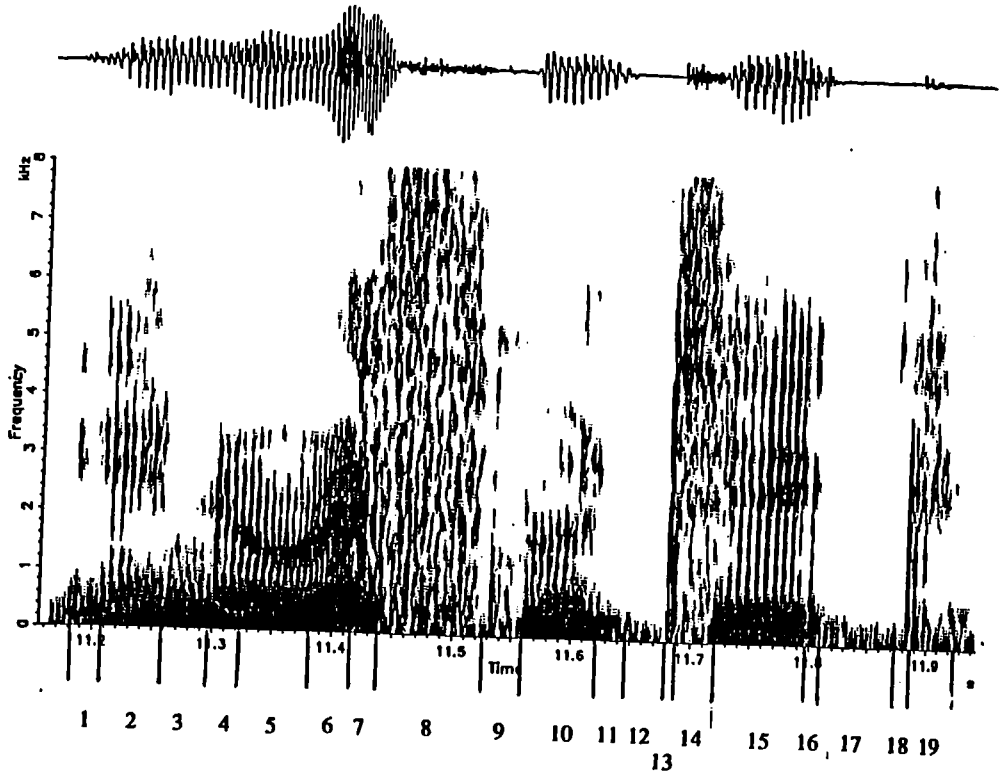

Figure 2 Acoustic phonetic labelling of the utterance "in arithmetic"

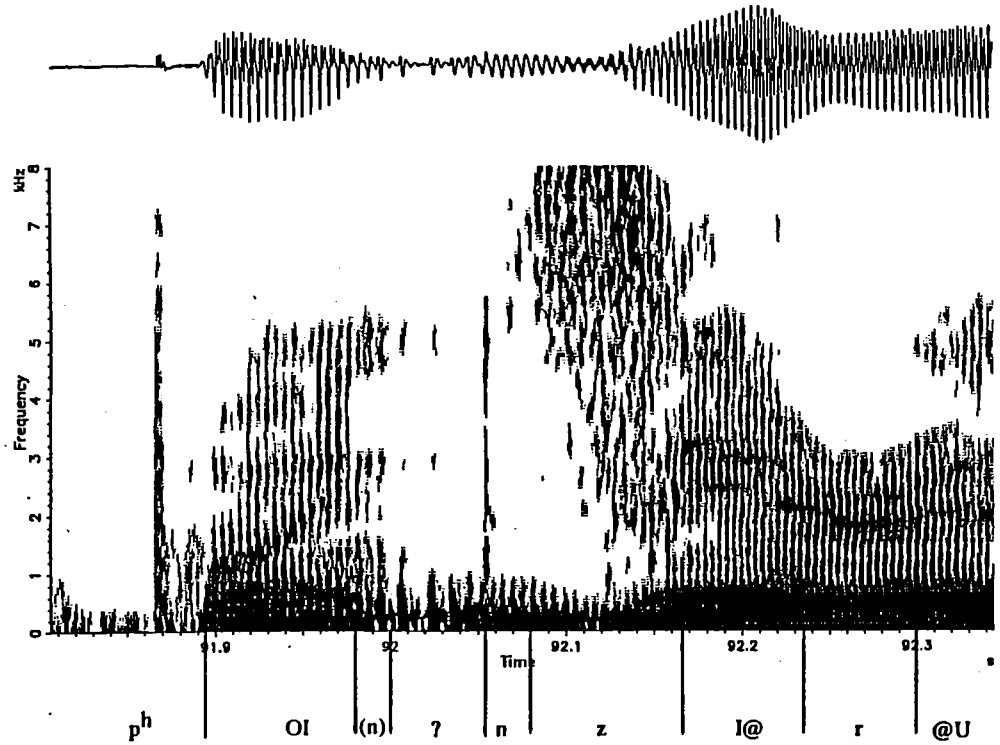

Figure 3 Speech pressure waveform and spectrogram of the utterance "point zero", showing the de-sequencing of acoustic events associated with the phonemic structure

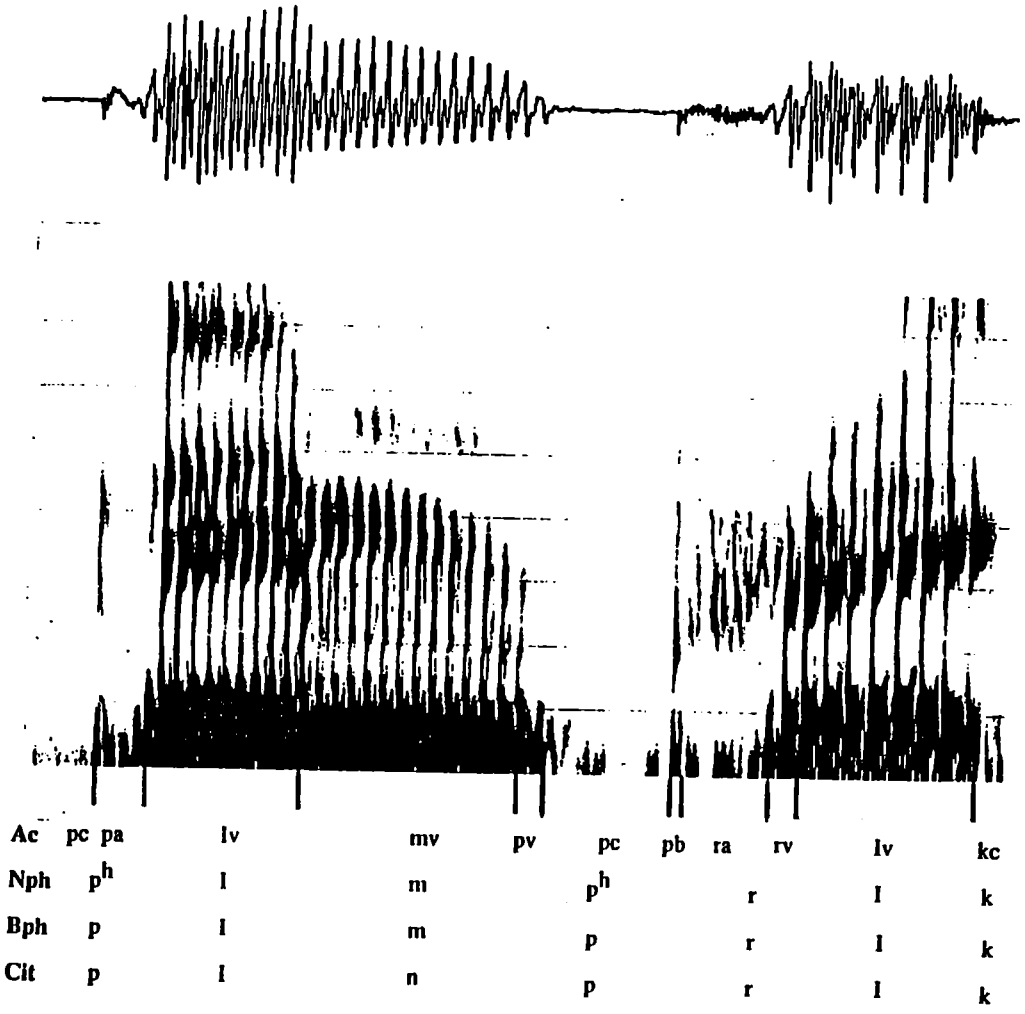

Figure 4 The utterance "pin prick" illustrating four levels of labelling: acoustic phonetic (Ac), narrow phonetic (Nph), broad phonetic (Bph), and citation phonemic (Cit)