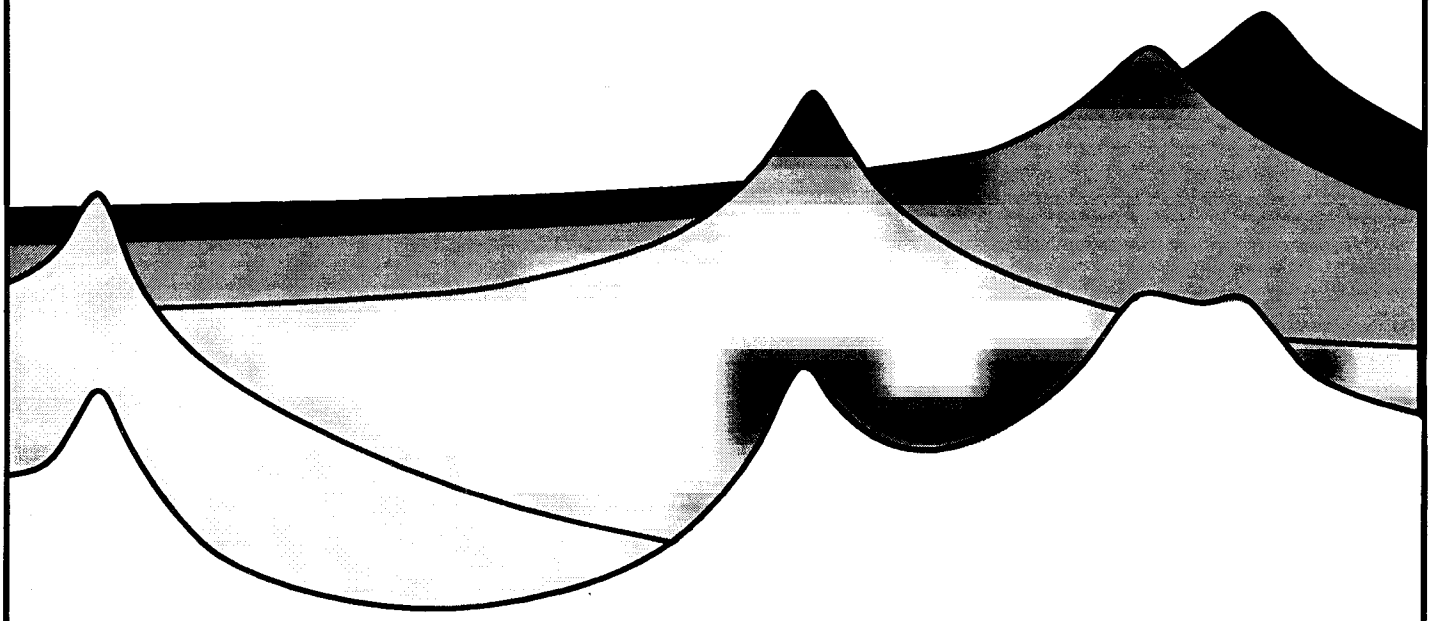


UCLA Working Papers in Phonetics

Number 88

September 1994

RH94

A Manual for Phonetic Transcription: Segmentation and Labeling of Words in Spontaneous Speech

Pat Keating, Peggy MacEachern, Aaron Shryock, Sylvia Dominguez

1. Introduction

This manual describes a set of conventions for the segmental phonetic transcription of the kind of speech that is found in spontaneous conversation. These conventions were developed for a project in which particular target words were segmented and labeled from the conversational utterances in the Switchboard corpus. The Switchboard corpus, collected by Texas Instruments and now available from the Linguistic Data Consortium at the University of Pennsylvania, comprises 3 million words of telephone conversations by 550 speakers. It has been completely transcribed in standard orthography by TI, but has not been phonetically transcribed. In our project we marked begin and end timepoints of selected words and provided a segmental phonetic transcription for each. We did not time-align the individual segments within the words, but the guidelines contained here should be sufficient for that task as well. However, due to the nature of this project, the transcription conventions we developed have little to say about the treatment of pauses, hesitations, or disfluencies.

The starting point for the transcription system we developed was the one associated with the TIMIT corpus of read speech. The symbol set presented here will be called the UCLABET. It is an extension of the TIMITBET, the symbol set used by transcribers at MIT for the TIMIT corpus, which is itself an extension of the ARPABET. The UCLABET adds to the TIMITBET new symbols that allow a narrower transcription, in particular, additional places of articulation for stop and fricative consonants, and three diacritics. Throughout this manual we note the correspondence between our use of the UCLABET, and what is seen in the TIMIT database transcriptions.

In terms of the phonetic categories added to the symbol set, we follow the IPA system of Place by Manner distinctions for consonants, along with various diacritics. However, some arbitrary mapping between IPA and ASCII symbols is required for maximally useful machine-based transcriptions. We chose a TIMIT style of symbol set over two other, more extensive, symbols sets, namely PhonASCII (Allen 1988) and Worldbet (Hieronymous ms., n.d.) because of what we judged to be easier learning and keyboarding of the individual symbols. Our experience is that it is preferable for symbols to consist largely of lower-case letters, not numbers or non-alphabet characters, and for symbols to be formed of these characters in systematic ways. We believe that the ease of use of such a symbol set offsets the price it exacts, namely, symbols of variable length that must be separated by spaces. In any event, our symbol set could be automatically translated into one of these others.

A symbol set is only the starting point for phonetic transcription. A set of conventions for the use of those symbols is also necessary, most notably because the transcription imposes the two idealizations of **discrete segments** in a **linear order** on the speech signal. It is not always clear, especially in fluent speech, how the signal is to be segmented into phoneme-sized units, no matter how narrowly defined they may be. Other transcription difficulties arise from the limited number of symbols (in any finite symbol set), and from differences in the phonemic systems of the various speakers and transcribers. Therefore any phonetic transcription project must incorporate guidelines for transcribers. In general, the literature is of little practical help in this regard. For example, the TIMIT database is accompanied by only sketchy explanations of transcription practice

(Seneff and Zue 1988, Garofolo et al. 1993). PhonASCII (Allen 1988), being a symbol set rather than a transcription method, offers no explicit conventions.

Most prior transcription projects have apparently not yielded publicly-available materials. Even textbooks are notably vague on the actual practice of anything but the broadest transcription, and in any event they rely entirely on the listening ability of the transcriber. One naturally looks for conventions, like the TIMIT ones, that refer to the acoustic signal as well as to the trained listener's percept. We found most useful the conventions for computer-corrected transcription by Henton and Bladon (1987), and a draft of the OGI conventions for transcription of telephone speech by the Spoken Language group at the Oregon Graduate Institute (Metzler and Nathman 1993), kindly made available by Ron Cole. It is for this reason that we consider it worthwhile to circulate our own conventions: that there be something in the public domain that at least aims at completeness and which can serve as a starting point for debate and development.

Our transcription was carried out using the Labeler facility in Entropics' *xwaves*. The screen display consisted of a time-aligned waveform, wideband spectrogram and orthographic transcription for the portion of the signal containing the word to be transcribed. Examples of this display will be seen throughout this manual (except as otherwise noted). The phonetic transcription was typed into an additional window, in which prior transcriptions of the same token (e.g. by other transcribers) could also be displayed.

2. Consonants

2.1. Presence vs. absence of a consonant. Though we did not perform word-internal temporal segmentation, we tried not to transcribe any segments which had no plausible segmentation in the target word. Sometimes a consonant is heard but (especially with flaps or other weak sonorants) no clear interval that could be uniquely associated with the consonant is seen in the signal. If there is a brief local amplitude drop in the signal which could be assigned to the consonant, the consonant is transcribed. If there is no amplitude drop in the signal and/or no interval in which that consonant seems to predominate, then no consonant is transcribed. An example of this latter kind is that in the word "mavericks", it seems that the /v/ and /r/ often overlap and it can be impossible to distinguish any piece of the signal that might be attributed to the /v/ alone. It would be possible to devise a form of segmental transcription in which nominally-adjacent segments are allowed to overlap, but we have not done so.

Occasionally a fricative, e.g. [s], is clearly heard, but only a gap is seen in the acoustic displays. In these cases the fricative is transcribed.

2.2. Stop closure symbols
2.2.1. TIMIT *pcl,tcl,kcl,bcl,dcl,gcl*. We follow the TIMIT practice of dividing stop consonants into separately-transcribed acoustic closure and release portions. The closure symbols are formed from the basic stop character followed by "cl". For example, the full transcription for a labial stop would be *pcl p:* a labial closure followed by a labial release. A stop is expected to have a closure interval unless it is the second stop in a cluster (section 2.2.3.4 below). The onset of stop closure is associated with a sudden sharp drop in amplitude and loss of energy in formants above F1. Voicing offset is not a criterion for onset of closure.

A closure may contain some noise and still be transcribed as a stop: most signals are somewhat noisy, so only noise beyond the background level can be taken as a clear indication of loss of stop closure. As a result, some instances of weak frication in a stop

consonant will be ignored. Figure 1 ("limited") shows a token with two instances of *dcl* which are clearly defined even though they do contain some weak energy. Figure 2 ("but") shows a token in which stop closures *bcl* and *tcl* are transcribed against a high level of background noise. However, if frication noise or sonorant energy during a phonemic stop constriction is strong enough that a true fricative or an approximant is heard and seen, then it is generally transcribed as a fricative. The symbols available for spirantized or sonorized stops are *ph, bh* for the bilabials, *tfr, dfr* (these are non-sibilant alveolar fricatives) or *dx* (flap) for the alveolars and *x, gh* for the velars. For more discussion of flaps, see below. Figure 3 ("baby-sitter") shows a pronuciation typical for this word, in which the first /b/ has a closure interval and release (*bcl b*) but the second /b/ is a voiced sonorant *bh*. One other context in which stops appear as fricatives is next to another fricative. In this case the two fricative segments must be distinct in order to be labeled separately. Figure 4 ("chips") shows a /ps/ sequence in which the /p/ has no stop closure and yet is clearly distinct from the /s/; it is transcribed as *ph*. Figure 5 (also "chips") shows the initial affricate phoneme with a spirantized closure, distinct from the fricative part of the affricate. If this interval were a stop, it would be transcribed as *tcl*, but because it is not a stop yet does not sound sibilant (that is, does not sound like [s]) it is transcribed as *tfr*. If a stop is spirantized or sonorized for some but not all of its constriction, then it is transcribed as a stop. For example, in a /ks/ sequence, the /k/ often exhibits some noise during part of its closure, but it is not always fricated enough to be heard as *x*, so it would be transcribed *kcl* despite the noise. At the same time, such weak frication may have as a consequence that the stop has no burst, so the whole sequence would be *kcl s*, and therefore distinct from *kcl k s*.

In a few instances a speaker had a clearly nasal closure for /b/, i.e. a prenasalized stop. This was not due to the context but appeared to be a consistent characteristic of the speaker. We transcribe this as *m b*, without *bcl*. Figure 6 ("bear") shows one such instance.

These stop closure symbols are also used for phonemic fricatives which are produced with a clear stop closure interval. In practice this means *tcl, dcl* as the other fricatives are covered by the extra stop symbols given immediately below. A stopped fricative will usually have a release, which can be transcribed with the fricative symbol (e.g. *tcl s*). The same transcription would be used for a more extended frication interval after a stop closure without a clear burst. Note that stop symbols are not used for short bits of silence adjacent to fricatives. The silent interval must actually sound like a stop; otherwise, the silence is not noted. If there is doubt about whether a fricative is stopped, the fricative is transcribed.

This discussion applies to a so-called epenthetic stop adjacent to a fricative. If it sounds like a stop followed by a fricative, it is transcribed as such, with or without a burst depending on the signal. Figure 7 ("since", displayed using CSL, not xwaves) shows an epenthetic stop with a closure *tcl* and release burst *t*. Sometimes there may be an ambiguity between an epenthetic stop and a single pitch period of glottalization at the end of the previous segment. In the case of glottalization, the frequency band of the pulse should match that of the previous segment. The frequency band of a burst for an epenthetic stop, however, will not match the frequency of the previous segment. Figure 8 shows another token of "since" which clearly contains glottalization (and nasalization) of the vowel, not an epenthetic stop.

Figure 1. Examples of stop closures.

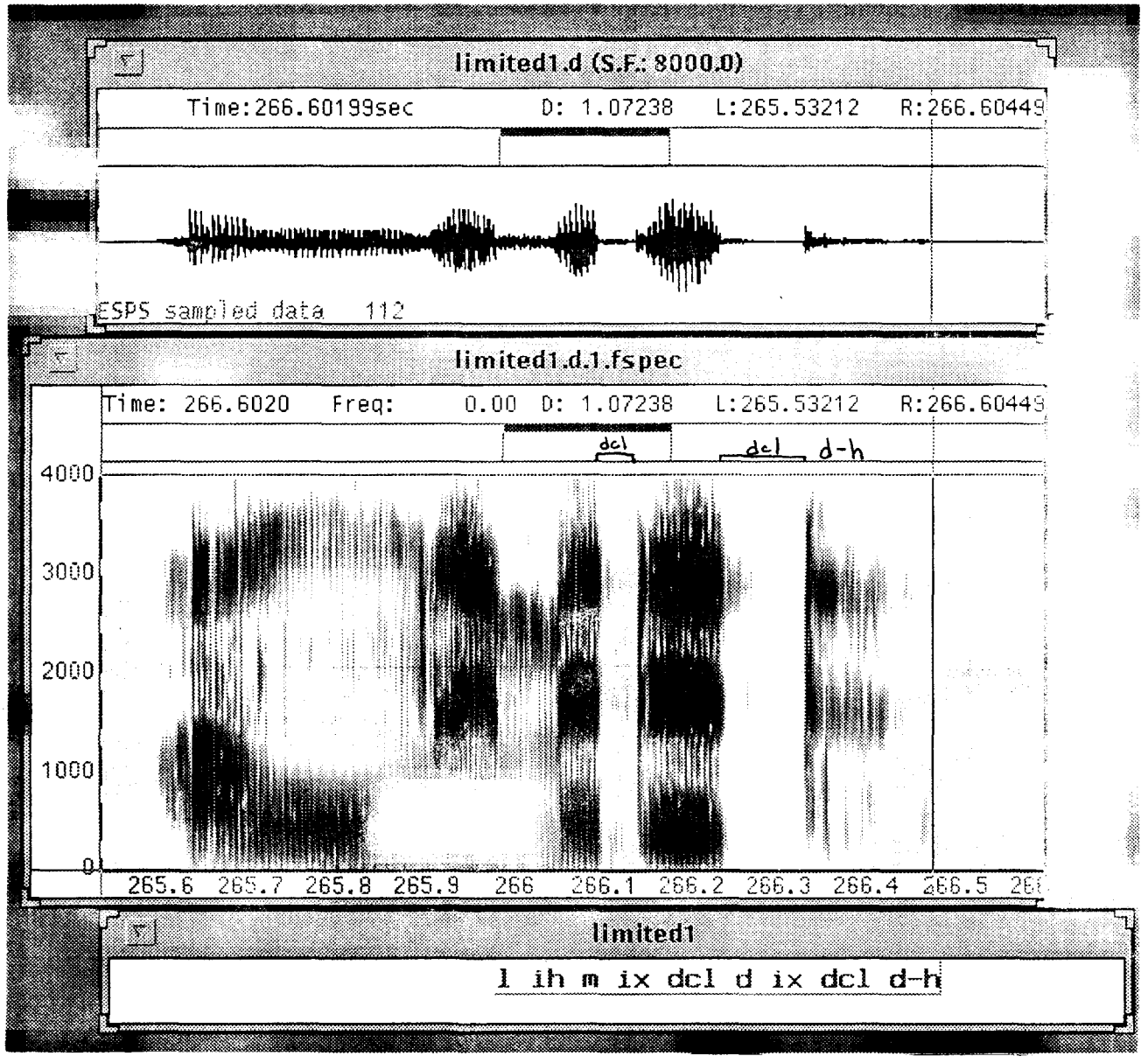


Figure 2. Examples of stop closures.

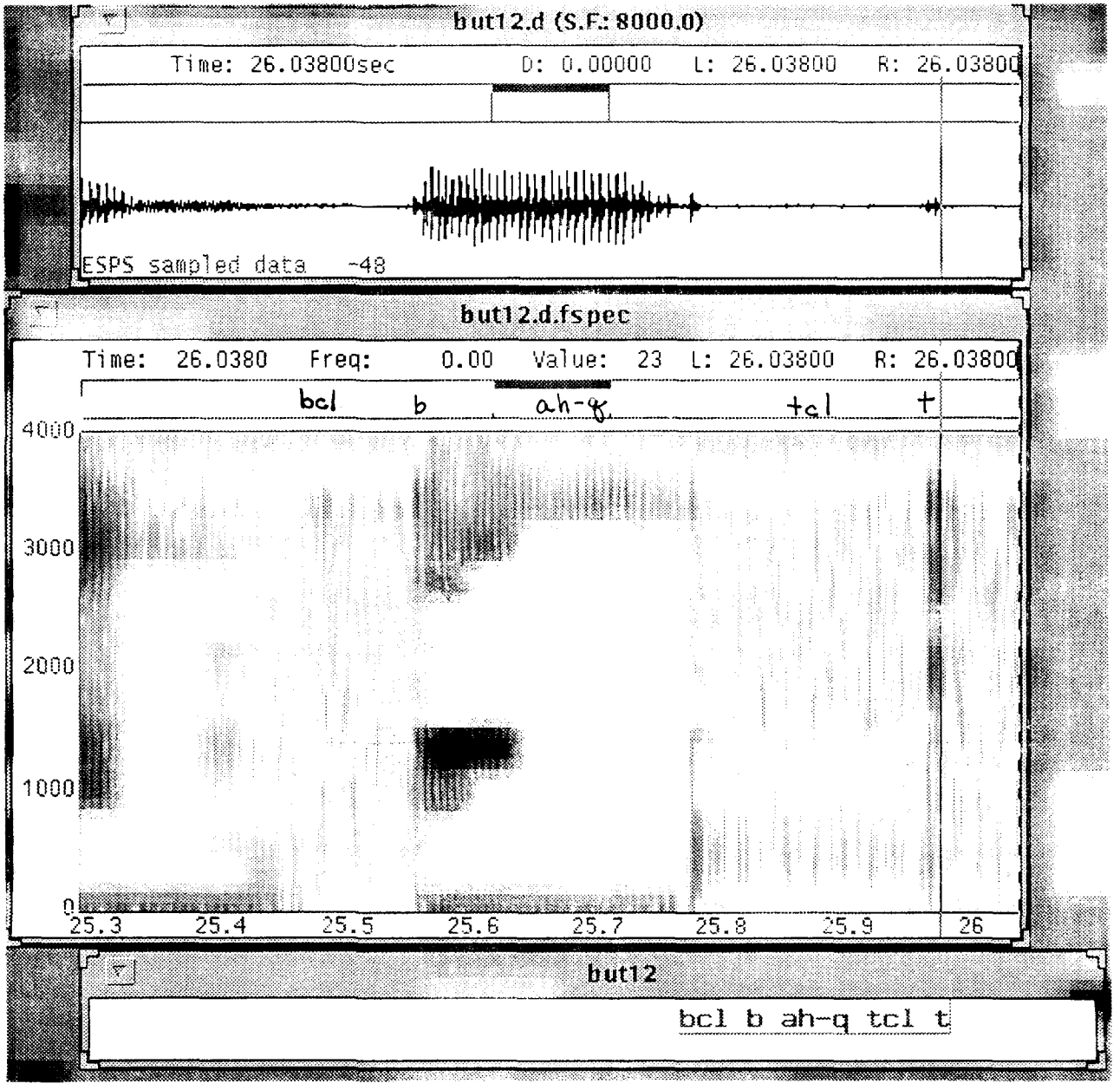


Figure 3. Example of *bh*.

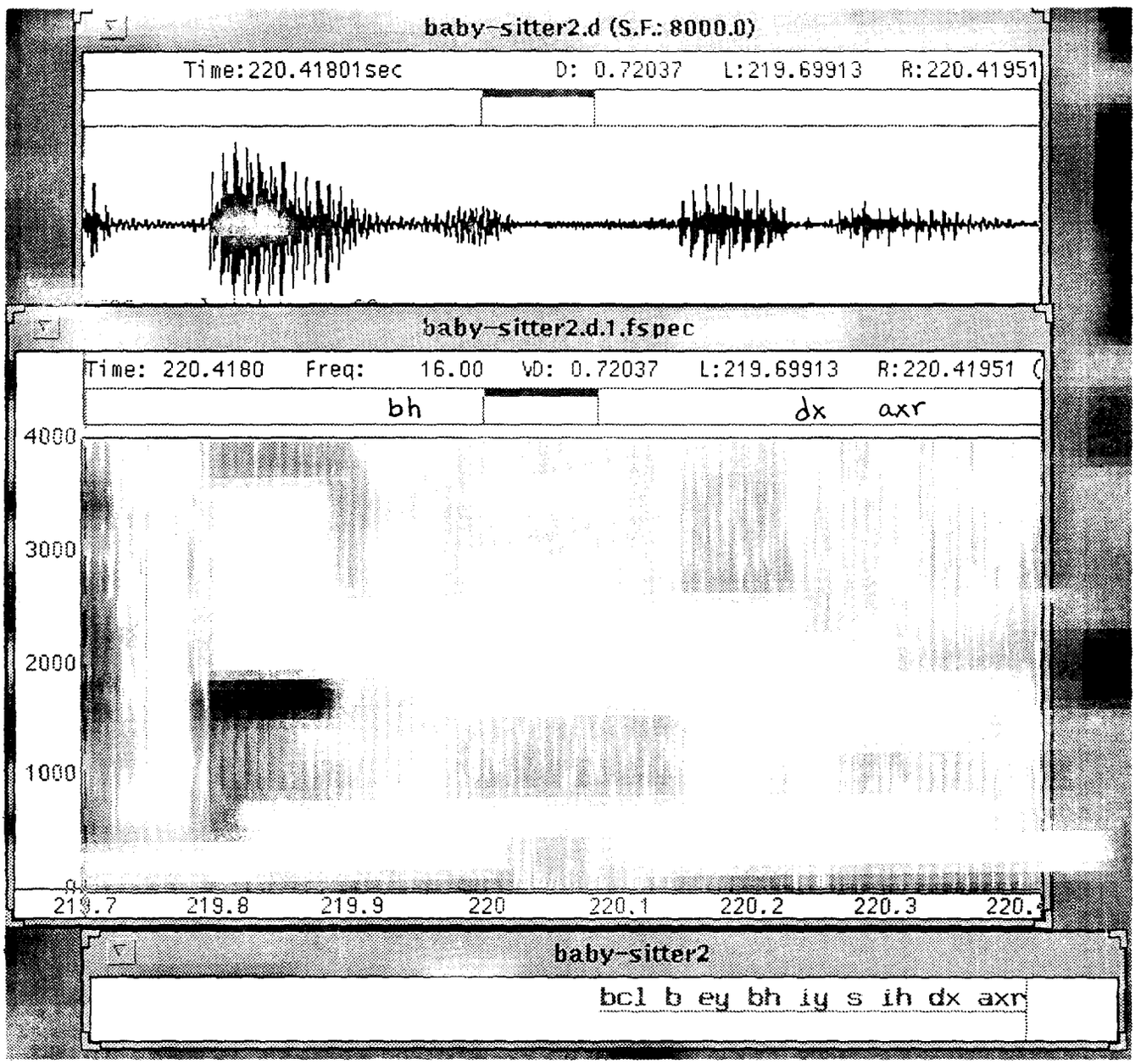


Figure 4. Example of *ph*.

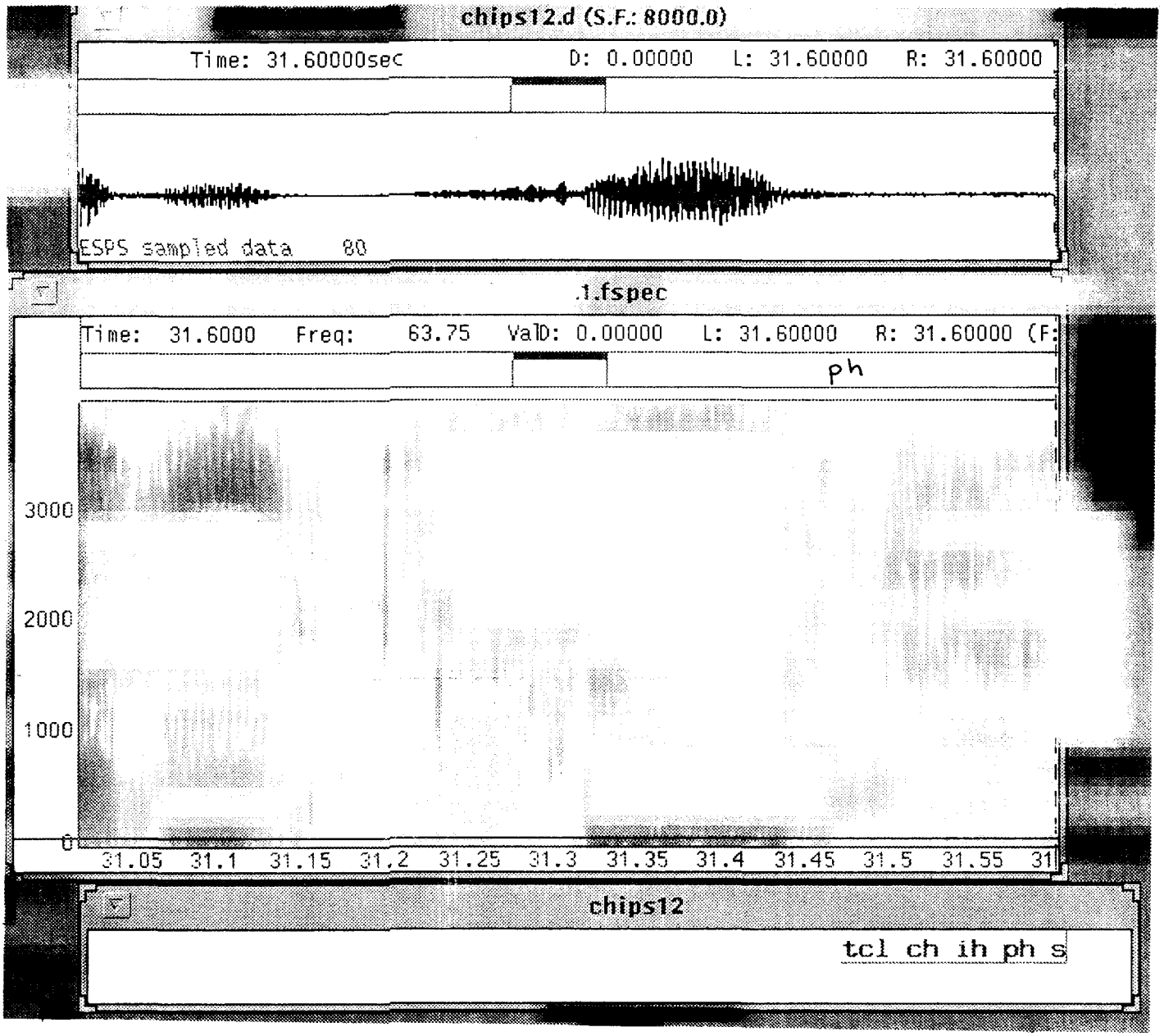


Figure 5. Example of *tfr*.

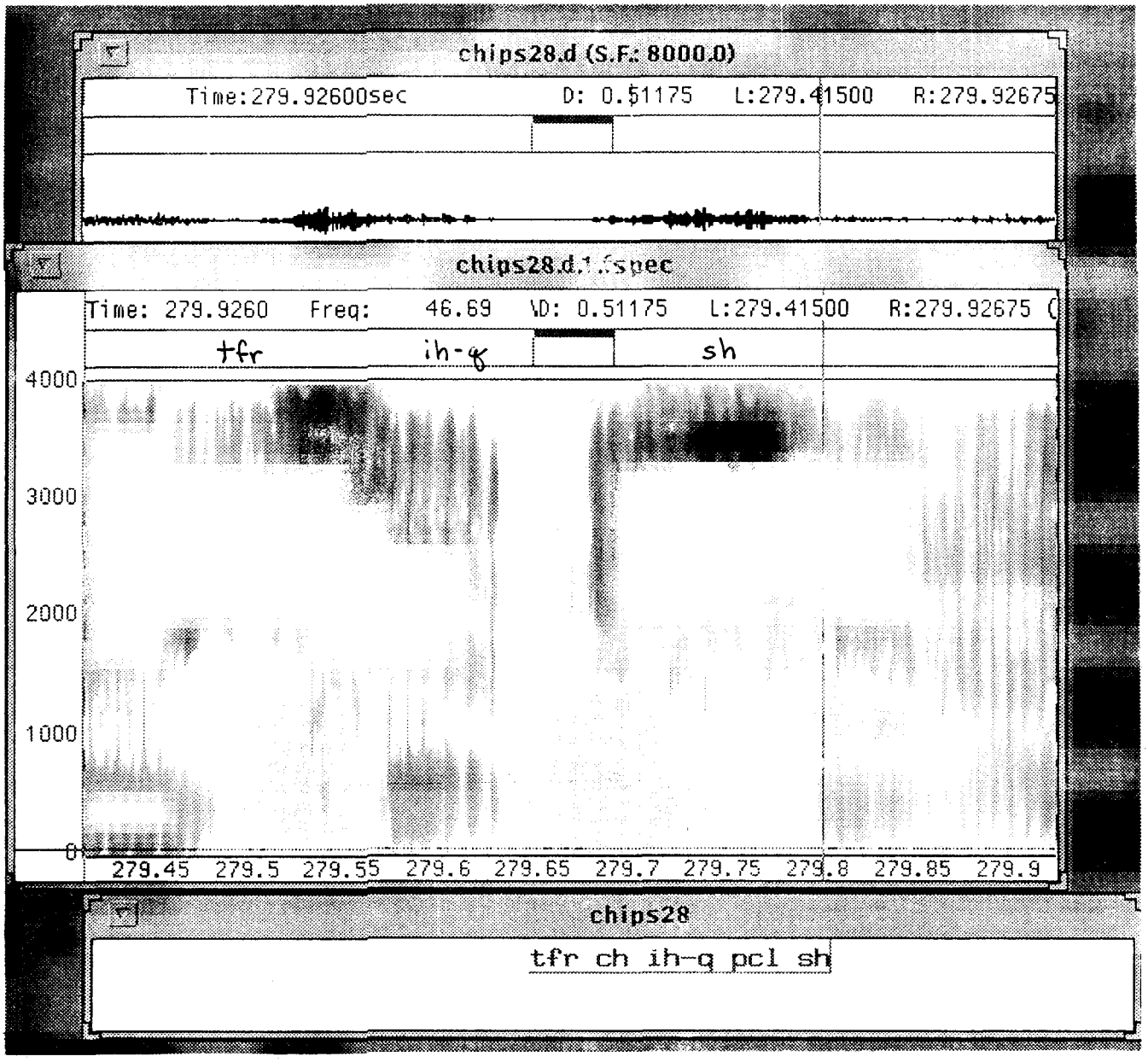


Figure 6. Example of prenasalized stop.

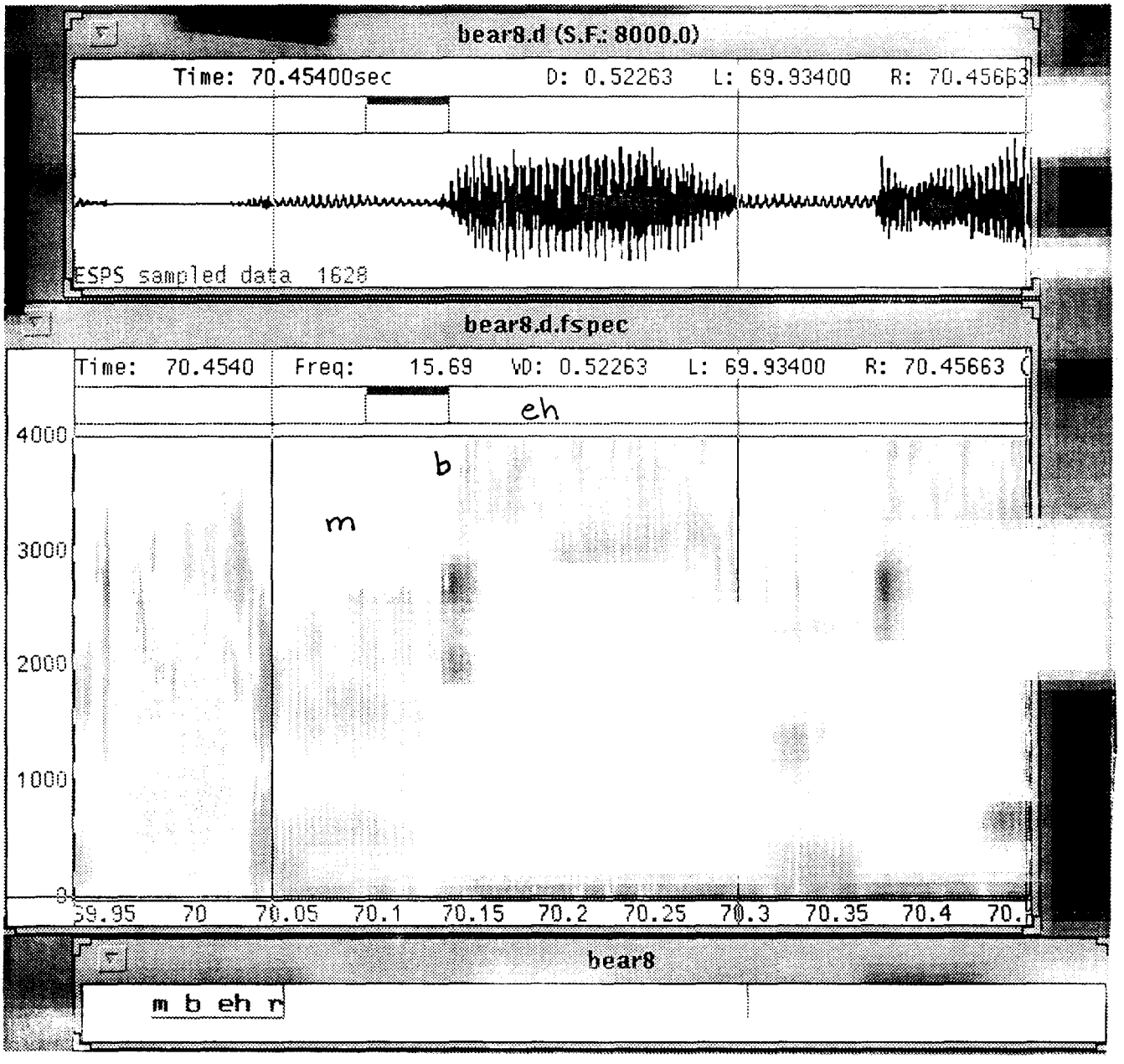


Figure 7. Example of epenthetic stop.

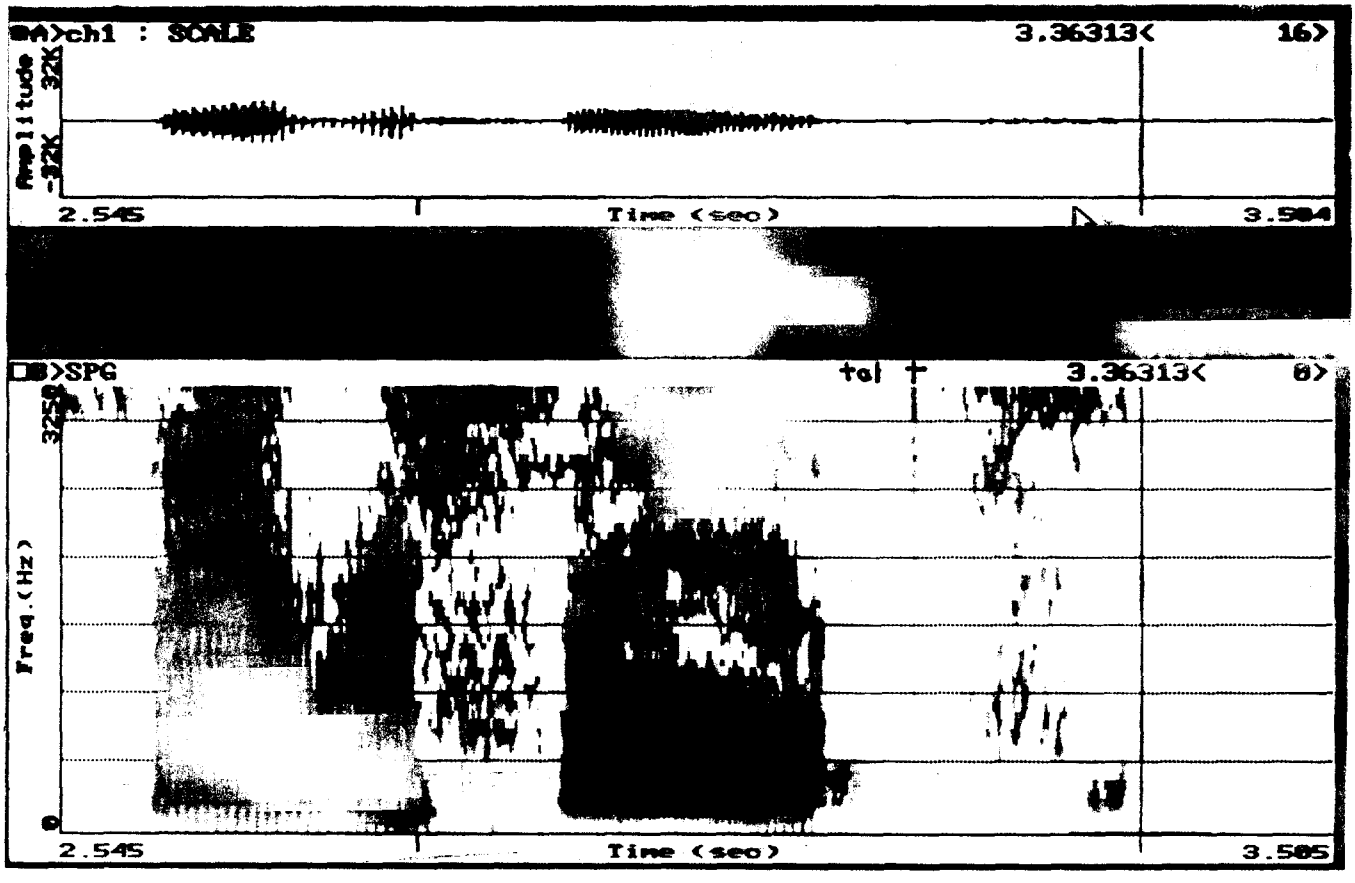
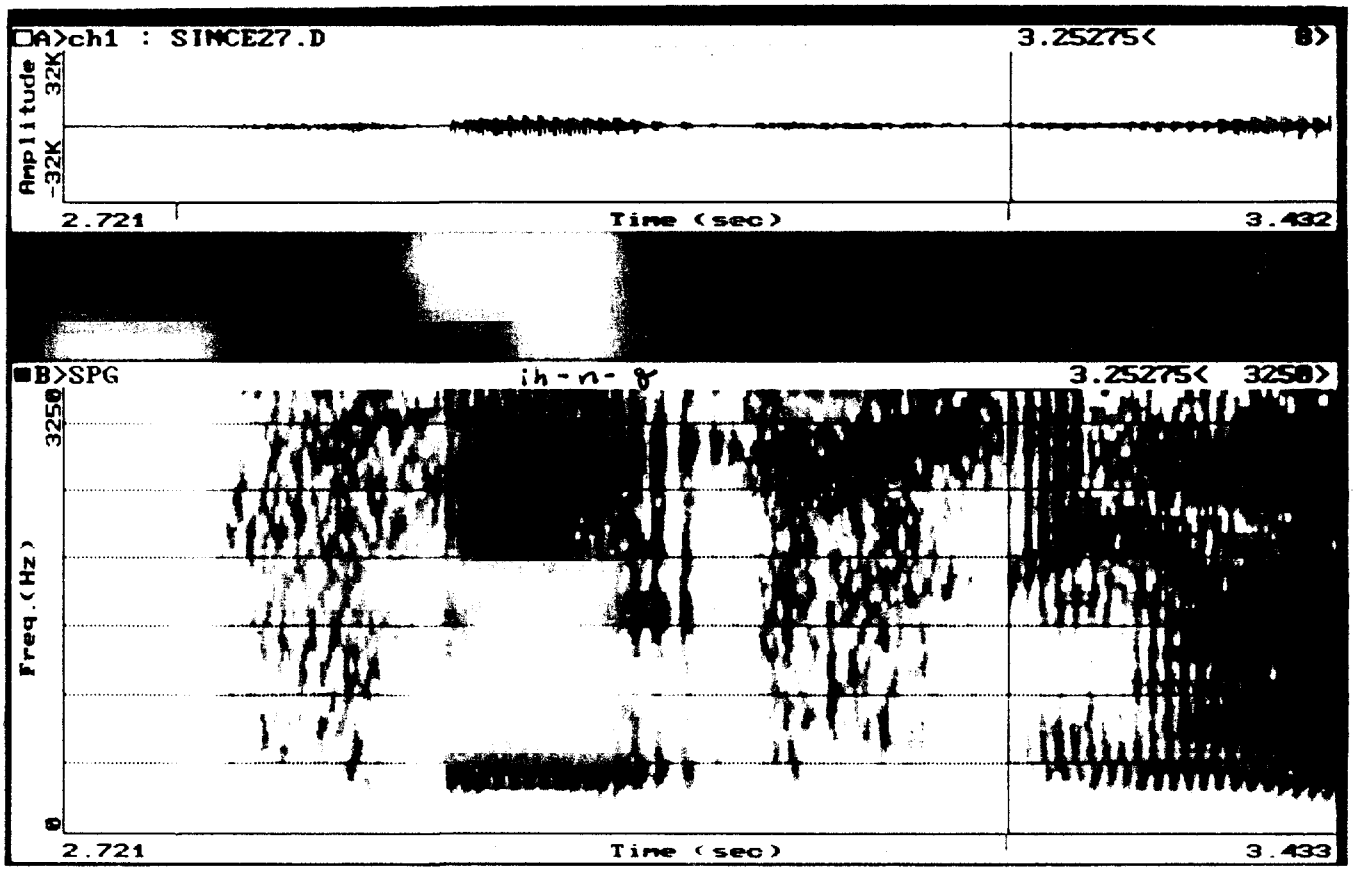


Figure 8. Example of glottalization.



2.2.2. Additional stop closure symbols *fcl, vcl, thcl, dhcl.* Sometimes labiodental and dental fricative phonemes are pronounced as stops or affricates, with stop closures. TIMITBET provides no symbols for stops at these places of articulation, so we have added symbols for the closure intervals. The releases of these stop closures are transcribed with the corresponding fricative symbols *f, v, th, dh.*. For example, a phonemic /v/ pronounced as either a stop or an affricate would be transcribed *vcl v.* No distinction is made between stops and affricates in this transcription.

It should be noted that these dental stop symbols have been used only for stopped variants of the dental fricatives. They have not been used to transcribe dental variants of phonemic alveolar stops. We would expect contextual assimilations in phrases such as "about the", but the /t/ is then in a cluster and so is typically treated as unreleased. Judging whether the closure itself is assimilated in place is too difficult a decision, and so such cases are transcribed simply as *tcl dh.*

TIMIT correspondences: None of these symbols are in the TIMITBET. It seems that *fcl* and *vcl* should be mapped into TIMIT *f,v.* thcl corresponds variably to *th, tcl,* and *epi;* presumably *dhcl* similarly. The apparent inconsistency of TIMIT practice in dealing with stopped fricatives is one reason we have added these symbols.

2.2.3. Other aspects of transcribing stop closures.
2.2.3.1. Adjacent to pause: Voiceless closure adjacent to a silent pause cannot be accurately segmented because either its beginning or its ending point is not known. With TIMIT no principled basis is given for transcribing such a voiceless closure. As a result, in TIMIT voiceless stops after pause are consistently transcribed without any closure, in the way that second stops in stops clusters are transcribed without closures. On the other hand, in TIMIT voiceless stops before pause are treated differently. If their closure were not transcribed, then they would not be recorded in the transcription at all, since closure and release are the only stop elements transcribed -- formant transitions, no matter how audible, are not recorded. In TIMIT, such closure intervals are transcribed so that the stops are recorded, but their segmented duration in the signal seems to be random. Our approach to this difficulty is the following. We consistently transcribe closures for stops adjacent to pause, giving them the arbitrary duration of 200 msec. This number was chosen to be longer than any observed delimited closure, so that these closures would stand out in any distribution of measured closure durations in a corpus. (This criterion is meant to apply generally. In our actual transcription project, no durations were assigned to individual segments in words. The 200-msec closures do however determine the location of word boundaries.) Note that this provision applies only to silent pauses. If a breath is heard as part of a pause, the closure will not extend (forward or backward) into the breath, and the closure will therefore have a duration less than 200 msec.

2.2.3.2. Voicing decision: Whether there is acoustic voicing during the closure is not the most important thing in determining the closure voicing label. A voiceless phoneme can have some voicing during closure and a voiced phoneme can have none. We determine voicing by listening to the segment before the stop, including the transitions, or to the stop release and following segment if there is no preceding segment. In ambiguous cases, the phonemic voicing of the segment is preferred in labeling.

If both a closure and release are present for a **single** phonemic stop segment, they must agree in transcribed voicing. For example, with a final stop, if the closure sounds voiced but the release is clearly voiceless, then this is a case of partial devoicing, which we do not transcribe (just as when a fricative is partially devoiced). Therefore in this case the whole stop is transcribed as voiced. Only if the whole stop, in its context, sounds voiceless would it be transcribed as such.

2.2.3.3. Place decision: Place of articulation is determined by the formant transitions into the stop. Assimilation of place is transcribed if it is both seen and heard. As with voicing, a closure and its release must agree in their transcribed place.

2.2.3.4. Clusters: If the first of two stops is unreleased and the closure interval is otherwise indivisible, then only one closure and one release are transcribed. Generally the first phonemic stop's place and voicing are used for the closure, and the second phonemic stop's place and voicing are used for the release. This convention is followed both within words and across word boundaries. Across word boundaries this convention can be confusing, as part of the stop is not recorded within a given word. For a word-initial stop, there is in fact little ambiguity. If the stop has no closure, then it must be part of a cluster with a stop in the preceding word (except as described in the next section). That is because, as described in 2.2.3.1, even a closure in a silent pause is transcribed, and as described in 2.2.1, a spirantized stop is transcribed as a fricative. On the other hand, a word-final stop with no release may or may not be followed by another stop.

2.2.3.5. After nasals: The oral closure of a stop may or may not be present after a nasal, i.e. it is possible to have "*n d*" instead of "*n dcl d*" -- no non-nasal closure visible before the release. The acoustic convention here is that a complete vertical band of "white space" above the baseline is enough to count as a closure, or a clear sharp drop in amplitude after the nasal before the release -- so usually some closure will be seen. Figure 9 ("conditions") is an example in which the nasalization was judged to extend up to the stop release. Similarly, prenasalized voiced stops as in Figure 6.

TIMIT correspondences: To the extent determinable from the TIMIT documentation, use of these closure symbols follows TIMIT except in three respects: (1) we standardize silent closures, not delimited by a breath, to 200 msec, and use these postpausally as well as prepausally; (2) because we transcribe closures for stopped fricatives, it is possible that some instances of *tcl, dcl* used by us for stopping of alveolar fricatives would not appear in TIMIT transcriptions; (3) because we do not use *epi* we use stop closure symbols where TIMIT would sometimes use *epi* .

2.3. Stop release symbols.
2.3.1. Oral stop releases.
2.3.1.1. TIMIT *p,t,k,b,d,g*. Transcription of a stop release is associated with an acoustic burst, a sudden sharp increase in amplitude. The beginning of a release is the left edge of this burst (on the waveform) and the end of a release into a sonorant is the Voice Onset, as seen at higher formants. Before a voiceless segment or silence, the end of release is the end of the full-frequency range of noise seen in the release. Sometimes there is no clear release. The default transcription is none, that is, positive evidence of a release is required for one to be transcribed.

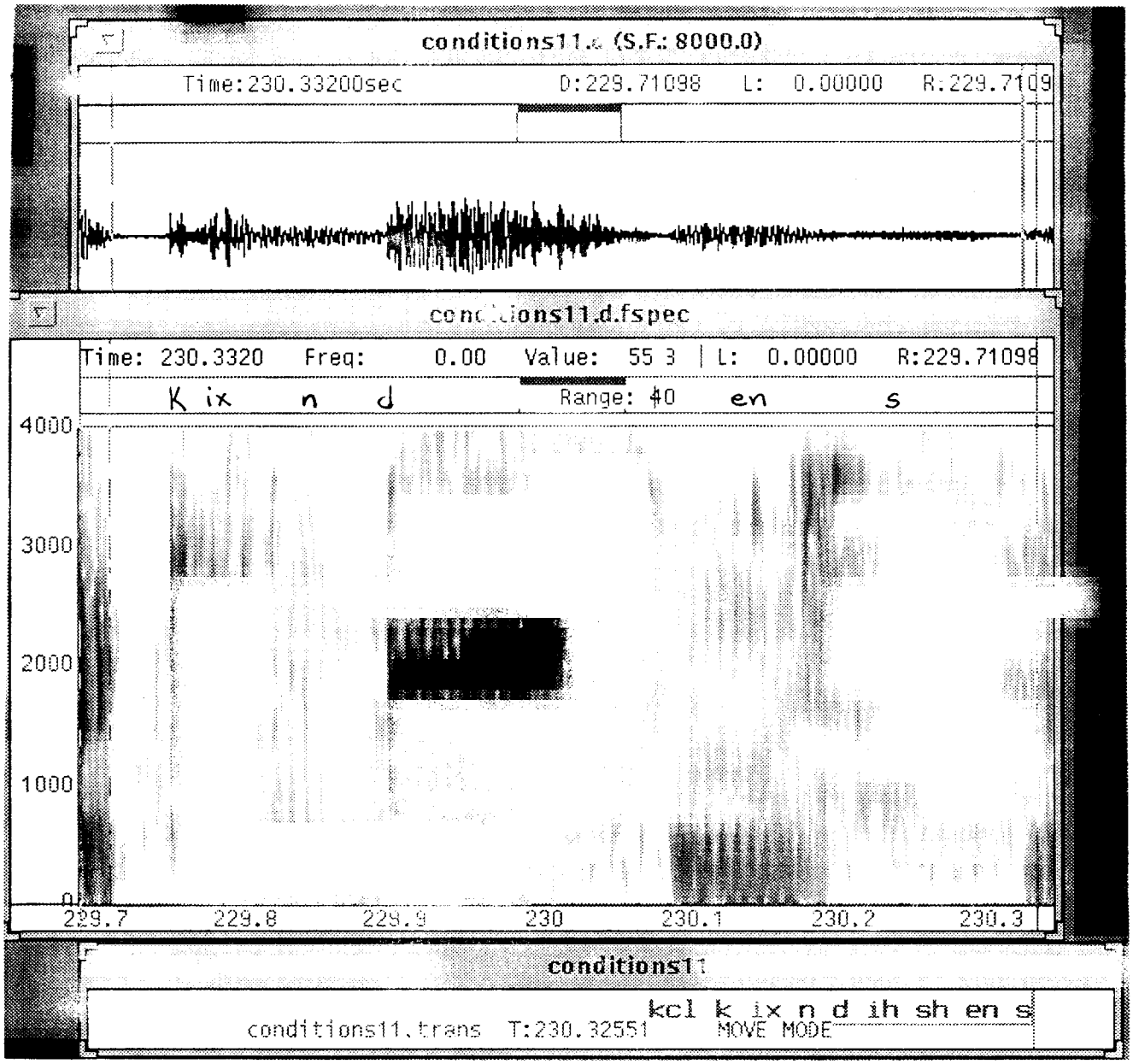
For affrication of stop releases, see section 2.8 on affricates below.

TIMIT correspondences : Seems the same.

2.3.1.2. Aspiration. Transcription of stop consonant aspiration (-h) is described in 4.2. Of relevance here is that a final stop (voiced or voiceless) with a long voiceless release is coded as aspirated, e.g. *p-h* or *b-h*. Figure 1 ("limited") shows an example of a final /d/ transcribed in this way.

2.3.1.3. Other stop releases. As described above in section 2.2.2, the fricative symbols *f,v,th,dh* are also used to transcribe releases of labio-dental and dental stops.

Figure 9. Example of nasal plus oral stop sequence.



2.3.2. Glottal stop. Our treatment of glottal stop and laryngealization differs from that in TIMIT. Parallel to our treatment of other stops, "glottal stop" includes glottal stop closure *qcl* and glottal stop release *q*. We try to distinguish such a full glottal stop (a more pronounced closure/release of pressure) from mere laryngealization (creaky voicing) of a voiced segment. With laryngealization, discussed in section 4.3 below, the formant structure of the oral articulation is preserved. In practice, full glottal stop is expected to be less common than laryngealization. Glottal stop can be found (1) phrase-initially (glottal attack); (2) when substituted for an oral voiceless consonant (e.g. /t/), as verified by sound and by lack of formant transitions. Figure 10 shows a token with "about" followed by "the", in which neither a /t/ nor a /ð/ is heard. There is simply a glottal closure without audible release (against a background of noise).

For consistency with the treatment of oral stops, *qcl* adjacent to a pause is given a duration of 200 msec.

TIMIT correspondences: Our *q* probably should be merged with our *qcl* for TIMIT correspondence. These together correspond to (one use of) TIMIT *q*. *qcl* with 200 msec duration has no TIMIT correspondence (was not segmented in TIMIT).

2.4. Flaps (taps)
2.4.1. Oral flap *dx*. A flap is prototypically a very short voiced closure with no release burst. However, flaps can be of medium duration, and they can also have bursts, but they should not sound like stops. The closure of flaps may contain frication; in practice, flap is transcribed where a fricative would be more accurate, but where the brief duration of the sound leads the transcriber to hear a flap. Figure 3 ("baby-sitter") shows an instance of such a flap. Flaps are not determined by duration per se, but by sound, except for the following guideline: if the closure is less than 20 msec, then we use the flap symbol even if there is a burst (except after a nasal, where a very short *dcl* could be found), but if there is no burst, then decide by sound. Figure 1 ("limited") contains a phonemic /d/ in a flapping context transcribed as a stop, not a flap.

TIMIT correspondences: Seems the same.

2.4.2. Nasal flap *nx*. These are hard to distinguish from short /n/. One criterion in listening is to segment out the nasal plus a following vowel: if the nasal sounds like a plausible syllable onset for the following vowel, then /n/ is preferred.

TIMIT correspondences: Seems the same.

2.5. Nasals *m,n,ng*. Place assimilation of nasals is transcribed. Figure 11 ("explain") shows a final /n/ transcribed as [m]; it occurs before the word "more" (the transcribed [m] thus belongs to both words). Sometimes there is no interval in the signal that corresponds to a nasal stop: there is a nasalized vowel followed by an oral stop. The TIMIT convention is to arbitrarily mark the last one or two pitch periods of the vowel as the nasal consonant in these cases; instead we use a nasal diacritic (see below) followed by the oral stop.

TIMIT correspondences: Seems the same, except for TIMIT's use of *n* as a nasal diacritic as noted above. These very short *n* 's which result from the diacritic use of /n/ in TIMIT do not appear in our transcriptions.

Figure 10. Example of *qcl*.

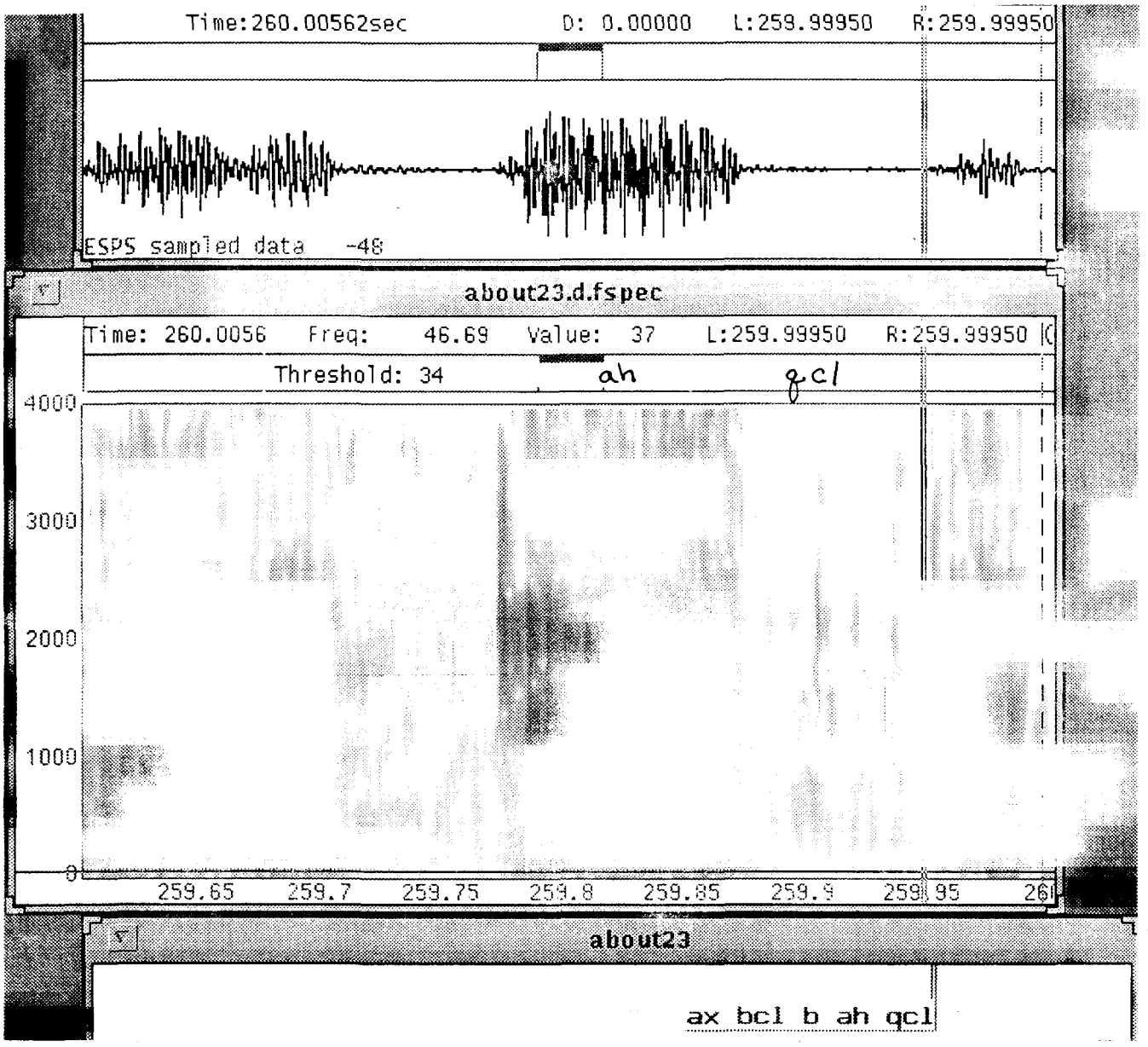
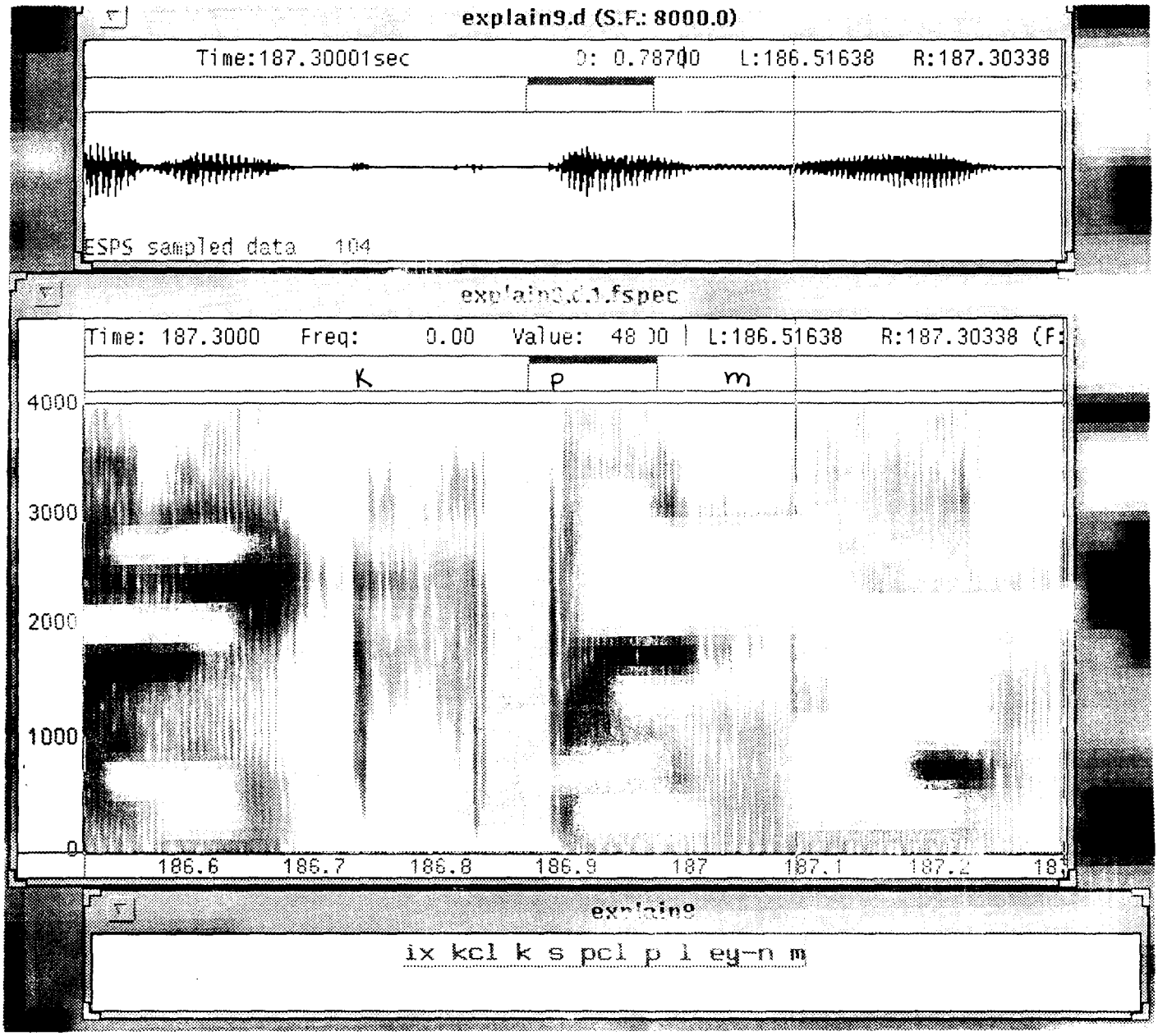


Figure 11. Example of place assimilation.



2.6. Syllabic sonorants *em,en,eng,el,er,axr*. Syllabic sonorants are used when /m n ŋ l r/ appear as syllabic nuclei, without a supporting /ə/ or other vowel that could be segmented out. All of these except *er* are expected only when stressless. As the default, transcription of a vowel is preferred, except for *er* and *axr* where the syllabic consonant is the default. Figure 3 ("baby-sitter") contains an example of syllabic stressless *axr*. Figure 9 ("conditions") contains an example of stressless syllabic *en*; compare the first syllable, with a vowel-nasal sequence, to the last syllable, with a syllabic nasal. Figure 12 ("well") shows a syllabic /l/ *el* in a monosyllabic function word.

We differ from TIMIT in our treatment of coda /r/. Unlike TIMIT, we use *r* for codas as well as onsets and reserve *axr* for syllabic-r. As a result, we do sometimes have to make difficult decisions about syllabicity when /r/ follows a high vowel or glide. Since we have to make difficult decisions about syllabic vs. onset /r/ in any case, it seems preferable to make these decisions more generally and keep *axr* consistent with the other syllabic sonorants.

Such a decision about syllabicity is needed for /r/ in words like "mavericks" or "mystery". To justify the syllabic variant, we look for amplitude dips before and after the sonorant which set it off as a nucleus, and look for the sonorant to have locally high rather than low amplitude. For example, with /r/, if the part of the signal having the lowest F3 value is very low in amplitude, this suggests the /r/ is an onset consonant, whereas if that part has a higher amplitude, this suggests it is a nucleus. In listening to the word, we segment the portion in question and try to judge that part alone as being one or two syllables. We must stress that these are very difficult judgments, however. Figures 13 and 14 show two tokens of "mystery", the first transcribed as two syllables (onset *r)*, and the second transcribed as three syllables (syllabic *axr*). The onset-*r* shows low amplitude throughout the low-F3 region, while the syllabic-*axr* shows a higher energy portion before the minimum, especially in F1.

No additional onset consonant is transcribed after a syllabic consonant, e.g. there is no *r* after *axr* in "mystery". This convention is from Henton & Bladon (1987). However, it would be possible and perhaps desirable to develop the criteria for when a syllabic consonant extends into onset position, so that each token could be transcribed accordingly.

TIMIT correspondences: Mostly the same as their description (but we hope we are more consistent in following it), except that we use *r* for nonsyllabic postvocalic /r/ where TIMIT uses *axr* .

2.7. Fricatives *ph,bh,f,v,th,dh,tfr,dfr,s,z,sh,zh,x,gh*. Fricative symbols are used for phonemic fricatives produced as such, and also for phonemic stops produced as fricatives or the corresponding approximants. They are also used for the releases of affricates.

Fricative symbols *ph,bh,tfr,dfr,x,gh* are used for spirantized stops, as described in section 2.2.1 above. Use of stop closure symbols for stopped fricatives is also described in section 2.2.2 above.

Devoicing of fricatives is transcribed only when it clearly dominates the percept. It must be judged in the context of the preceding vowel (e.g. something that in isolation sounds like /s/ might still sound like /z/ after a long vowel). A small amount of voicing at the onset of the fricative, or a relatively short fricative after a long vowel, will generally make a fricative sound voiced. Figure 9 ("conditions (that)") shows a devoiced /z/; note its s-like duration.

Figure 12. Example of *el.*

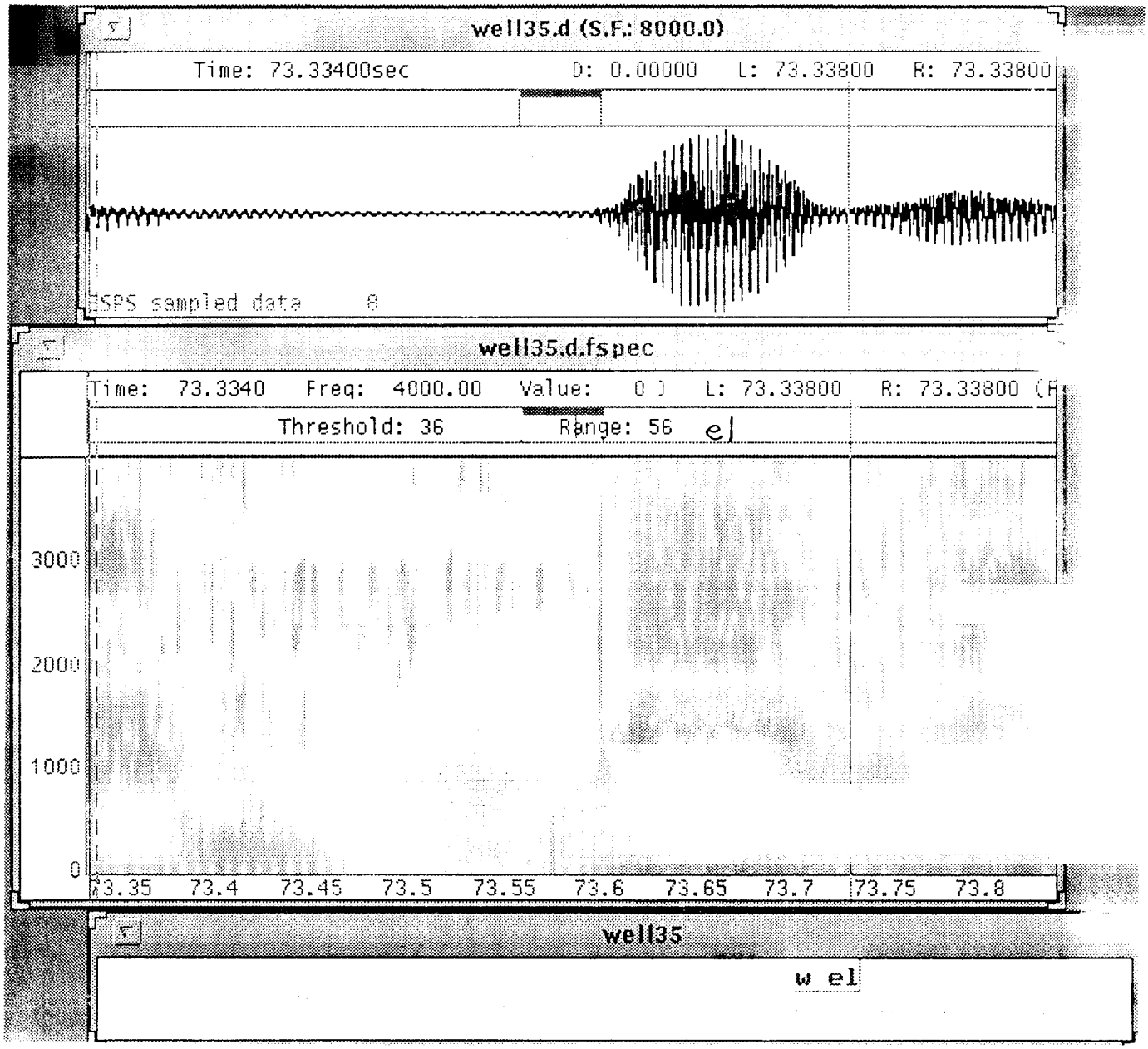


Figure 13. Example of onset *r*.

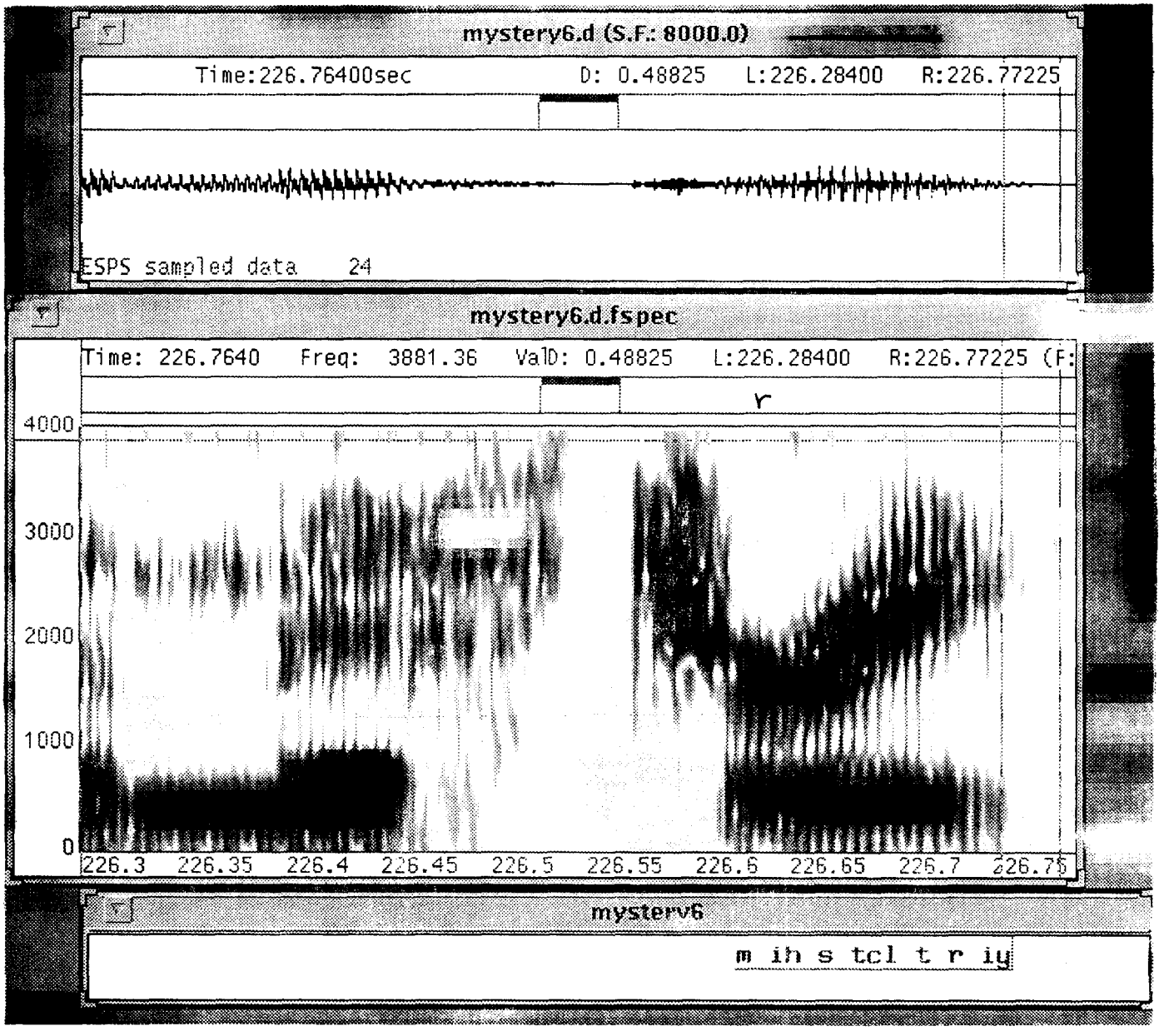
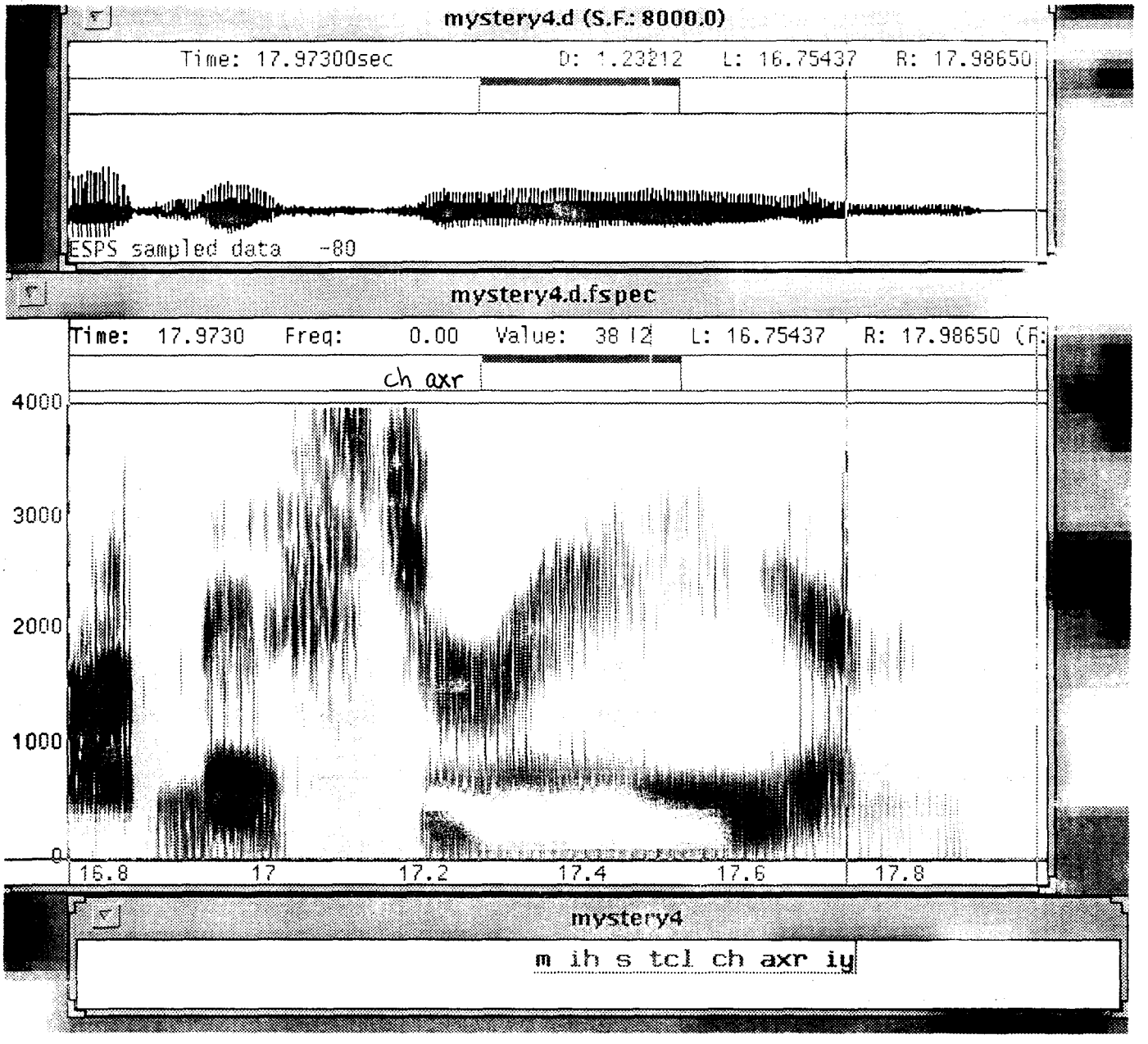


Figure 14. Example of syllabic *axr*.



Place of articulation assimilation is transcribed if it is complete. Partial assimilation is not transcribed, e.g. /s/ before /ʃ/ with lower frequency noise than is usual in /s/, but still distinct from /ʃ/, is transcribed *s*. Figure 5 ("chips") shows a token in which the frequency-lowering is deemed extreme enough to yield a place of articulation change; whether it is an assimilation (due to the previous affricate, or to the following glide *y*) or not need not be decided by the transcriber.

TIMIT correspondences : We have added new fricative symbols *ph,bh,x,gh*. In TIMIT these would presumably appear as *pcl, bcl, kcl, gcl.*

2.8. Affricates. These have *tcl* and *dcl* as their closures and *ch* and *jh* as their releases and are therefore distinguished from most stop-fricative sequences, which would be *tcl t sh* or *dcl d zh.*

Affrication of /t,d/ before approximants (e.g. *dcl jh r* for /dr/ as in "driver's") is transcribed only if the affricate percept is clear and neutralizing, with substantial noise after the release (not just a labialized or retracted stop release). See Figure 14.

TIMIT correspondences: Seems the same.

2.9. Liquids *r, l.* As discussed in section 2.6 above, our use of *r* differs in one respect from TIMIT. As in TIMIT, syllable-initial or intervocalic /ɹ/ is *r,* but we use *r* for coda /ɹ/, whereas TIMIT uses *axr.* That is, we use *r* for all non-syllabic /ɹ/, just as we use *l* for all non-syllabic /l/.

A very dark /l/ might sound like [w]. The criterion for labeling is that if F3 is high, it is *l*, while if F3 falls, it is *w*. Figure 15 ("probably") shows an example of this: note the fall of F3 between the flanking vowels.

2.10. Glides *y,w*. These are not used for off-glides of diphthongs, even if prolonged, since unit diphthong symbols are available. They are also not used as onsets for syllables which follow a diphthong; the diphthong is segmented to include any such material.

In a reduced syllable like /pju/ in "reputation", the glide might be missing altogether, or it might be manifested only as some influence on the surrounding segments. If there is no interval in the signal that might be segmented as the glide, then it is not transcribed. Figure 16 ("reputation") shows a token in which the /ju/ sequence appears as a steady fronted vowel *ux* .

As mentioned in 2.9, *w* can be transcribed for a very dark /l/.

We do not systematically distinguish phonemic voiceless-/w/ (IPA /ʍ/) for those speakers who preserve this as a phoneme distinct from /w/. If a /w/ appears completely voiceless, for any reason, it would be transcribed *w-h.*

TIMIT correspondences: Seems the same.

2.11. Voiced and voiceless /h/. This distinction is determined by voicing in the acoustic signal. Voiced *hv* is quite common.

TIMIT correspondences: Seems the same.

Figure 15. Example of *w* for /l/.

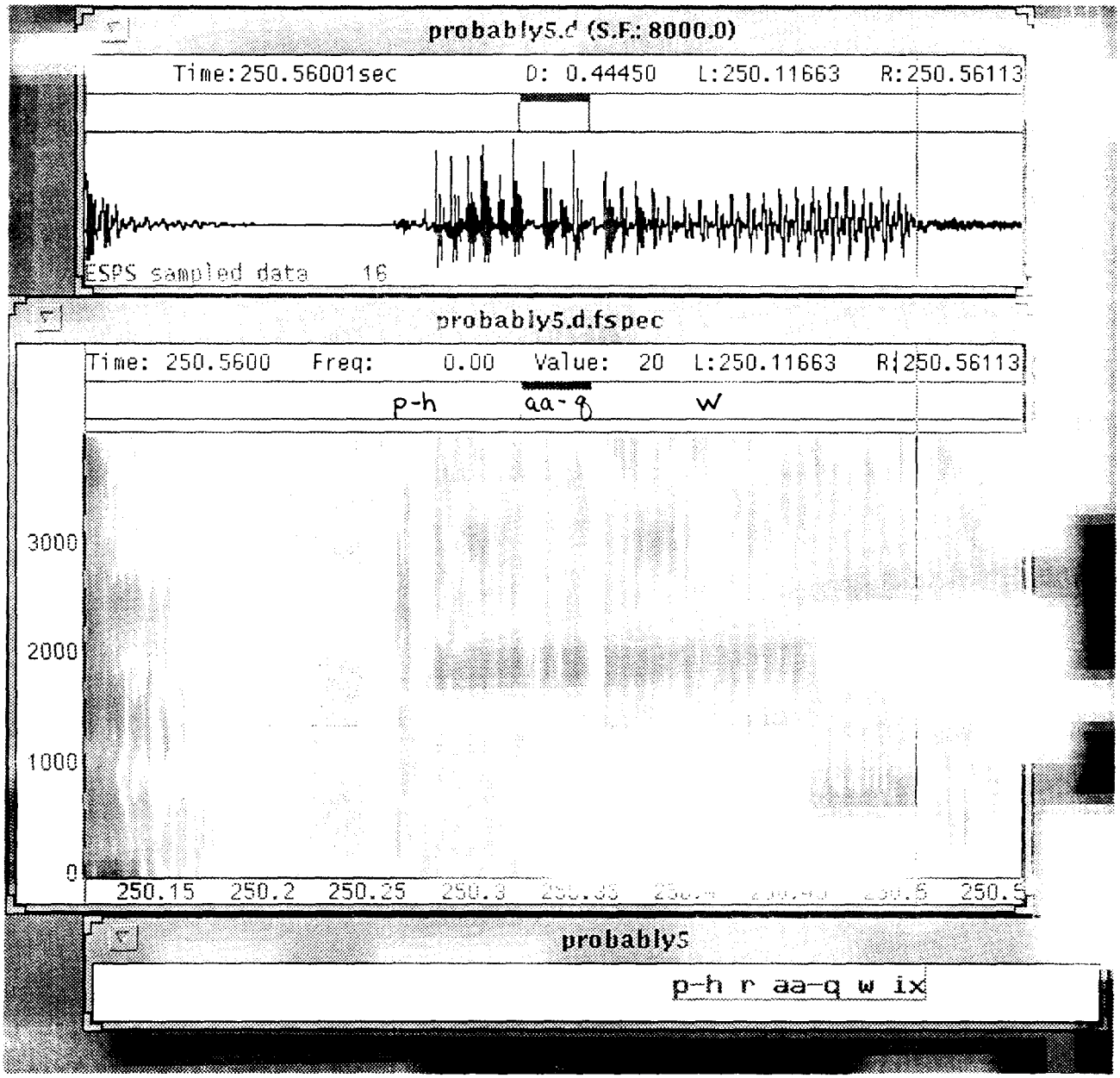
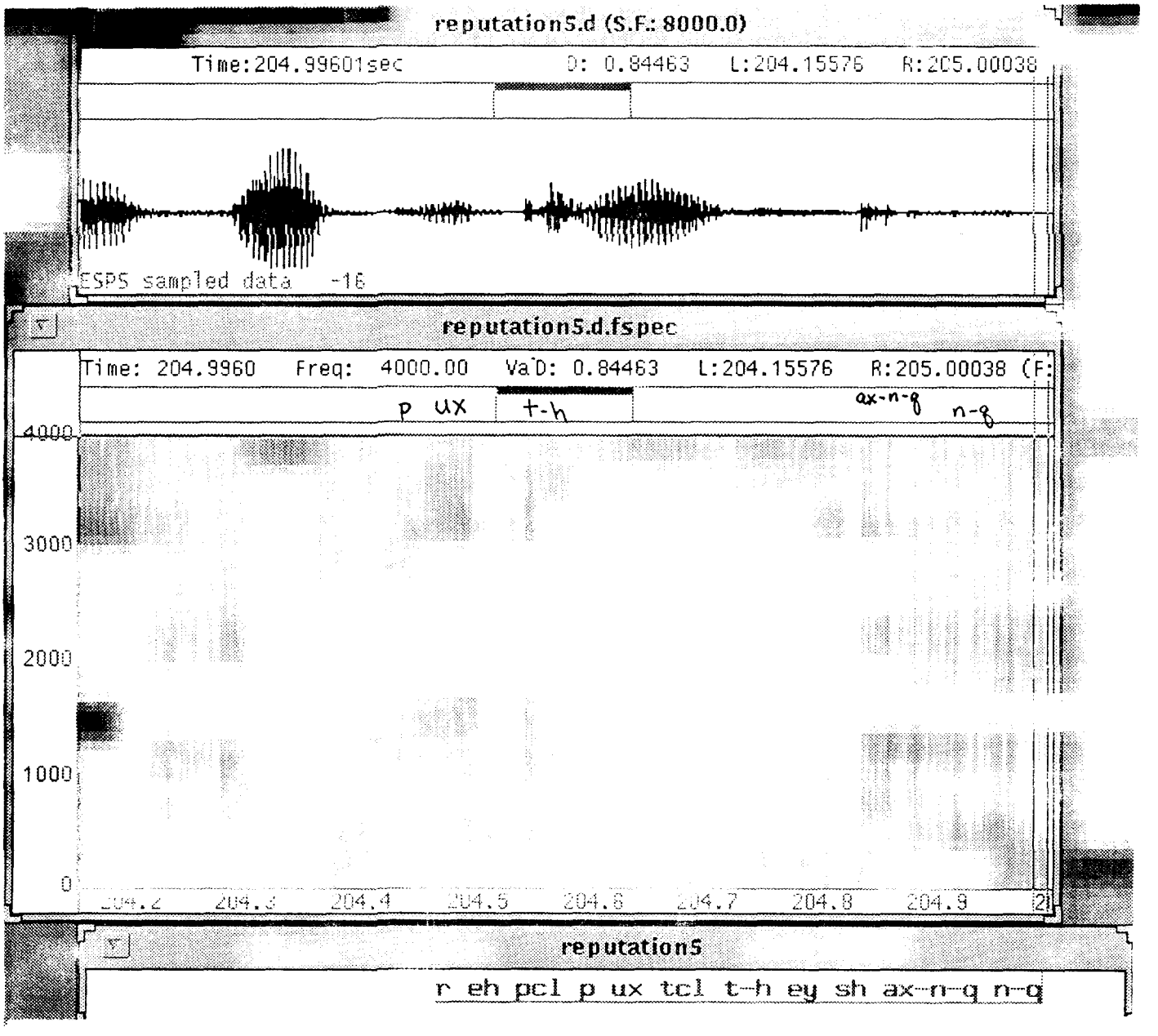


Figure 16. Example of glide-vowel coalescence.



3. Vowels

We use the TIMITBET symbols for vowels. Each symbol, whether monophthong or diphthong, contains two characters. In general, tense vowels end in *y*, lax vowels end in *h*, and centralized vowels end in *x*, except for the low vowels.

The strong preference is for one vowel segment&label per phonemic vowel. If vowel quality changes over the course of a phonemic unit, it is one segment and the dominant quality determines the label.

TIMIT correspondences: Seems the same.

3.1. Diphthongs. These are units, not segmented. No additional glides are transcribed as onsets after these, even if they are quite extreme (e.g. no *y* after *ay* or *ey* even if it sounds like [j] rather than [ɪ]). Similarly, if a vowel before a glide (or any other consonant) sounds diphthongal by anticipation, that gliding is not transcribed.

Monophthongization of diphthongs is rarely transcribed. The formants must look quite steady in the spectrogram (except of course for consonantal transitions at the margins). Most variation that sounds like monophthongization is more accurately greatly-reduced diphthongal movement, and that is something we make no effort to transcribe. Figure 10 ("about") shows one of the few diphthongs transcribed as a monophthong.

TIMIT correspondences : Seems the same.

3.2. *ax, ix, axr.* Following TIMIT, one of these is used for any short and/or reduced vowel, and they are the default set if the vowel is phonologically stressless. In practice, the full/reduced distinction is a very difficult one. We look for vowels which have 6 or fewer pitch periods, except that next to an approximant the longer transitions must be factored out. Following TIMIT, the distinction between *axr* and the others depends on F3, with *axr* having a low F3, while the distinction between *ax* and *ix* depends on the relative frequency of F2 at its strongest point. If the F2 is closer to F3, we use *ix*; if closer to F1, we use *ax*. Note that /i/ and /o/ also occur as stressless vowels in English, but these will be transcribed only if the quality is clear. One difficult distinction is *ix* vs. *iy* for short vowels; when in doubt, we prefer the label corresponding to the phoneme. Another difficult decision arises with vowels that are not short and may have some degree of stress, which sound between *ih* and *ix* (e.g. final syllable in "limited").

TIMIT correspondences: Basically the same, but we are not satisfied with critera for when to use a reduced vowel symbol over a full one.

3.3. *uw* vs. *ux*. If F2 is closer to F3, we use *ux*; if closer to F1, we use *uw*. Fronted and backed versions of *ow* and *uh* are not distinguished.

TIMIT correspondences: Seems the same.

3.4. r-colored vowels. The two vowels *er* and *axr* are to be distinguished by stress (*er* is the regular vowel, *axr* is reduced).

It should be recognized that vowels before /r/ have different qualities than the same vowels elsewhere. We have chosen not to transcribe contextual r-coloring with a diacritic or other symbols. For vowels before /r/, we follow the TIMIT convention which is to prefer the lax vowel symbols over the tense ones: *ih* and *eh* (over *iy* and *ey*), as in Figure 6

("bear"). We use iy, *ey* only if there is a distinct rising movement in F2 and F3, distinct from consonant transitions, before the fall into /r/.

TIMIT correspondences: Seems the same as what they say.

3.5. Voiceless vowels. TIMIT has only one voiceless vowel, ax-h. Because we have a general *-h* diacritic, described below, we can indicate any vowel quality as voiceless. In practice, this means that we distinguish *ax-h* from *ix-h;* these two are distinguished according to F2, as for the voiced counterparts. Figure 17 ("reputation") contains 2 voiceless vowels.

TIMIT transcribes a voiceless vowel even for vowels with one or two pitch periods, but we use it only for completely voiceless vowels.

As also described in 4.2 below, if a voiceless consonant is followed by a voiceless vowel, such that there is a choice between transcribing aspiration on the consonant and devoicing of the vowel, we transcribe the voiceless vowel but not aspiration on the consonant. (This is an arbitrary solution to a logical difficulty in segmental transcription.)

TIMIT correspondences: We are more conservative than TIMIT in what counts as voiceless, and in what counts as *ax*. Our *ix-h* should be folded into our *ax-h* for correspondence with TIMIT.

4. Diacritics

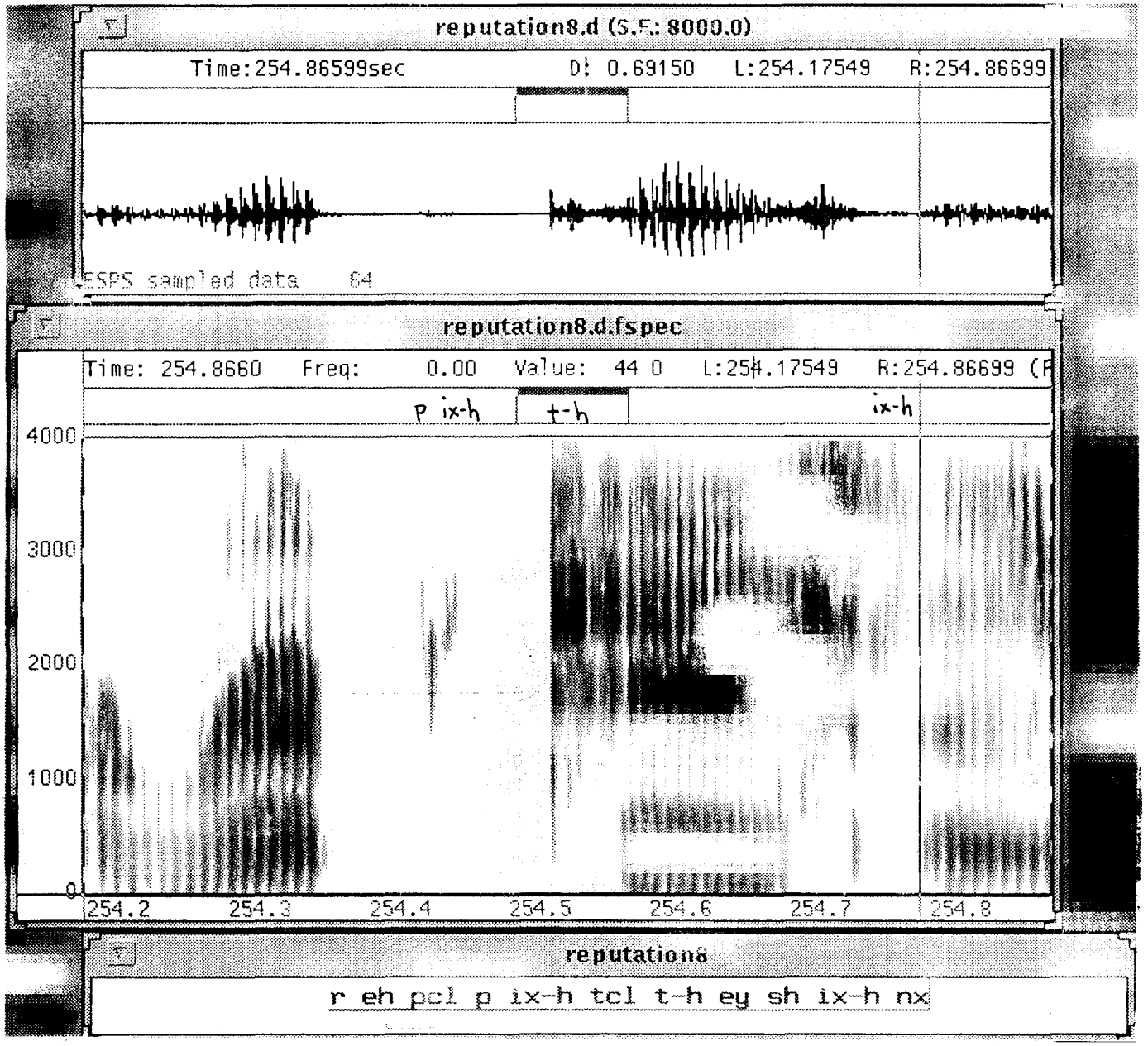
A major difference between our symbol set and the TIMITBET is that we provide three diacritics which can be freely added to other segments. Multiple diacritics on a single segment are listed in alphabetical order (for example Figure 16). No attempt is made to render any temporal ordering of the features involved.

4.1. *-n.* This is a nasalization diacritic applied to an entire vowel to indicate audible nasalization over any part of the vowel. Most instances involve nasalization of a vowel near a nasal consonant, but a few instances are of vowels which are heavily nasalized seemingly independent of the context. No effort is made to segment a nasal vs. non-nasal portion of a vowel, and the criterion for use of *-n* is auditory rather than acoustic: we listen to the vowel segment out of context, especially without the couple of pitch periods adjacent to any contextual nasal consonant, and if it sounds nasal, we use the diacritic. Various figures above show vowels adjacent to nasals with and without *-n*.

Nasalized vowels often show a quality shift compared to their non-nasalized counterparts. We generally do not transcribe such differences. The use of *-n* indicates the possibility of a quality difference, as well as nasalization.

4.2. *-h.* This is for sounds made with the vocal cords audibly spread apart, that is, two kinds of sounds: devoiced versions of voiced phonemes, and aspirated versions of voiceless stops. With respect to devoiced allophones of voiced phonemes, one use is for reduced vowels which are entirely voiceless--*ax-h, axr-h,* or *ix-h*, using the same F2/F3 frequency criterion as for voiced versions. Another use is for utterance-final devoicing of any sonorant, such as nasals. Partial devoicing of voiced phonemes is not generally transcribed.

Figure 17. Two examples of voiceless vowels.



With stops as base symbols *-h* indicates aspiration, e.g. *p-h* means an aspirated /p/. The rule of thumb used in transcribing aspiration is that labials and alveolars must have a VOT of 21 or more msec, and velars must have a VOT of 35 or more msec, always measured from the beginning of the **last** release burst. Figures 15, 16, 17 contain aspirated voiceless stops while Figures 2, 7, 9, 11 contain unaspirated voiceless stops. Figure 1 illustrates that this diacritic can be used even with a "voiced" stop as a base, e.g. *d-h*, because the stop has a voiceless release and clear aspiration beyond the required VOT value. In principle the affricate release *ch* could be aspirated as well.

In a consonant cluster beginning with a voiceless consonant, as in Figure 15, the "aspiration" is seen in effect as partial devoicing of the following consonant. We label this simply as *-h* on the stop, and do not also mark the following consonant as voiceless. This is for consistency with aspiration before vowels -- we never indicate that part of the following vowel is voiceless. Thus *-h* is interpreted as some devoicing of whatever segment follows the aspirated stop. The exception to this general rule is if the following segment is completely voiceless: in that case, the complete voicelessness takes precedence in labeling and we label for example the vowel as voiceless, but no aspiration on the consonant. For example, if the initial stop in "conditions" were aspirated, we could have either *k-h ax* or *k ax-h*, depending on whether the vowel is partially or completely voiceless after the /k/. Thus devoicing is labeled only once in a given sequence.

4.3. *-q*. This is a laryngealization (or "glottalization" or "creaky voice") diacritic for voiced sonorants. It is used for a segment which is partially or completely laryngealized; the entire segment is given the diacritic no matter how much or how little of the segment is laryngealized, and the diacritic follows the base symbol regardless of where in the segment the laryngealization occurs. Examples include vowels in hiatus (surprisingly rare); sonorants before glottalized coda consonants (Figures 2, 5); sonorants adjacent to *q* itself; intonational creak (Figures 15,16). However, if every segment in a speech sample is laryngealized, i.e. that is simply the speaker's usual voice quality, then it is not transcribed.

The difference between *-q* and the glottal stop (*qcl q*) is that with laryngealization, a sonorant maintains its particular F-pattern; that is, along with the constricted glottis there is a clearly identifiable primary supraglottal articulation.

TIMIT correspondences: TIMIT has no such general diacritics. (1) Our vowels with *-n* should be mapped into the vowels without diacritics, and at the same time the last pitch period (or 10 msec) of each such vowel should be segmented out and labeled *n,* to correspond to TIMIT. (2) Our i*x-h* should be mapped into *ax-h*, and any of our vowels with *-h* that follow the same vowel symbol without *-h* should be merged into that symbol. *-h* on any consonant symbol should be omitted. (3) Vowel symbols with *-q* should be mapped to TIMIT *q*. Other symbols with *-q* should omit the *-q* to correspond to TIMIT.

5. Segmentation/time-alignment

No word-internal temporal segmentation was performed, but a time alignment for the target word as a whole was provided. The segmental transcription and time alignment include only the target word. In general this segmentation was minimal, allotting to the target word only material that unambiguously belonged to that word. No segments from adjacent words and no adjacent pauses are included in the transcription of the target word or are otherwise transcribed, except that shared segments are presumed to belong in both words' transcriptions (5.3 below). Thus, if a segment is seen in a transcription, it was part of the pronunciation of that token. For example, if "bear" is transcribed as beginning with [mb], that means that there was no neighboring /m/ but that instead the speaker produced a

prenasalized /b/. Other conventions for segmentation at word boundaries include the following:

5.1. Boundary between two vowels. If there is an abrupt frequency or amplitude change suggesting an acoustic boundary, that is used as the boundary. If there is no clear boundary, then any transition between two vowels is not included in our target word.

5.2. Boundary between two consonants. Between two stops, the general TIMIT convention for stop clusters is followed: If the first of two stops is unreleased and the closure interval is otherwise indivisible, and if the transitions into the closure are consistent with the first stop (as judged from listening), then the entire closure is assigned to the first stop and thus the first word; the release is assigned to the second stop and thus the second word. If the transitions into the closure are consistent with the second stop, then the closure is assigned to the second stop and thus the second word (the first stop is treated as deleted). The default option is to give the whole closure to the first stop in this fashion. Distinct positive evidence can however be the basis for dividing the closure between the two stops: a release of the first stop in mid-closure, or a change in voicing consistent with the phonemic sequence.

Between consonants, especially approximants, where transitions are seen, the transition is divided at an acoustic boundary given by an abrupt change in amplitude or frequency.

5.3. Boundary between a shared consonant. Following TIMIT, if two identical consonants are adjacent across a word boundary and are realized as a single long or short consonant, then it is given one phonetic symbol but assigned to both of the words. That is, the first word ends at the end of this segment, and the second word begins at the beginning of this segment. Because we transcribe only single words we do not have to deal with the technical problem of notating that a consonant belongs to two words at once. (It is also possible to have two separate but identical consonants across a word boundary, if some boundary between them can be located, e.g. if there is an amplitude drop or glottal stop between them.)

5.4. Boundary between consonant and vowel. Between a consonant and a vowel there can be extensive formant transitions. In our segmentation these transitions always go into the vowel, even when they are very long (as when the consonant is *w, y, r* or *l)*. The consonant is therefore segmented as just a steady state, and the vowel is segmented to include all formant transitions as well as any steady-state portion it may have.

Note that diphthongs are transcribed as unit segments. That is, off-glides are not distinguished from peaks, and so no segmentation between these portions needs to be determined in our transcriptions.

6. Silences

epi. We chose not to use this TIMIT symbol at all, as its use in TIMIT transcriptions was not entirely clear to us.

pau. We transcribed only words that had been orthographically transcribed as complete, so that there were no word-internal pauses.

TIMIT correspondences: Some instances of TIMIT *epi* correspond to some of our use of stop closures in our transcriptions, in a way that cannot be automatically converted in either direction.

ACKNOWLEDGEMENT

This work was supported by a contract from the Linguistic Data Consortium at the University of Pennsylvania.

REFERENCES

Allen, G.D. (1988). The PhonASCII system. *Journal of the IPA 18*(1): 9-25.

Henton, C. and Bladon, A. (1987). Developing computerized transcription exercises for American English. *Journal of the IPA 17*(2): 72-82.

Hieronymous, J.L. (n.d.). ASCII Phonetic Symbols for the World's Languages: Worldbet. Ms., AT&T Bell Laboratories.

Garofolo, J.; L. Lamel; W. Fisher; J. Fiscus; D. Pallett; N. Dahlgren (1993). DARPA TIMIT. Distributed with TIMIT on CD-ROM, second (full) release, 1990. Section 5.2, Notes on Checking the Phonetic Transcriptions, by L. Lamel.

Metzler, T. and T. Nathman (1993). Labeling conventions. Ms., Center for Spoken Language Understanding, Oregon Graduate Institute.

Seneff, S. and V. Zue (1988). Transcription and alignment of the TIMIT Database, Distributed with TIMIT on CD-ROM, first (prototype) release.