Computer Speech and Language (1993) **2**, 169–191

A knowledge-based system for stop consonant identification based on speech spectrogram reading

Lori F. Lamel

LIMSI-CNRS, BP 133, 91403, Orsay, France

Abstract

In order to formalize the information used in spectrogram reading, a knowledge-based system for identifying spoken stop consonants was developed. Speech spectrogram reading involves interpreting the acoustic patterns in the image to determine the spoken utterance. One must selectively attend to many different acoustic cues, interpret their significance in light of other evidence, and make inferences based on information from multiple sources. The evidence, obtained from both spectrogram reading experiments and from teaching spectrogram reading, indicates that the process can be modeled with rules. Formalizing spectrogram reading entails refining the language used to describe acoustic events in the spectrogram, selecting a set of relevant acoustic events that distinguish among phonemes, and developing rules which map these acoustic attributes into phonemes. One way to assess how well the knowledge used by experts has been captured is by embedding the rules in a computer program. A knowledge-based system was selected because the expression and use of knowledge are explicit. The emphasis was in capturing the acoustic descriptions and modeling the reasoning used by human spectrogram readers. In this paper, the knowledge acquisition and knowledge representation, in terms of descriptions and rules, are described. A performance evaluation and error analysis are also provided, and the performance of an HMM-based phone recognizer on the same test data is given for comparison.

1. Introduction

While spectrograms have been used in speech analysis for many years, over the last decade there has been revived interest in the application of spectrogram reading toward continuous speech recognition. The spectrogram displays the energy distribution in the speech signal as a function of both time and frequency. Spectrogram reading involves interpreting the acoustic patterns in the image to determine the spoken utterance. One must selectively attend to many different acoustic cues, interpret their significance in light of other evidence, and make inferences based on information from multiple

The basis of this research was done while the author was at the Massachusetts Institute of Technology, Cambridge, MA, and was supported by DARPA under Contract N00014-82-K-0727, monitored through the Office of Naval Research.

0885–2308/93/020169+23 $08.00/0

sources. In a series of experiments intended to illustrate the richness of phonetic information in the speech signal (Cole, Rudnicky, Zue & Reddy, 1980), Zue demonstrated that high performance phonetic labeling of a spectrogram could be obtained without the use of higher level knowledge sources such as syntax and semantics. The phonetic transcription thus obtained was better than could be achieved by automatic speech recognition phonetic front ends (Klatt, 1977). It appears that the humans' ability to handle partial specification, integrate multiple cues, and properly interpret conflicting information contributes greatly to this high level of performance.

The spectrogram reader applies a variety of constraints to the identification problem including knowledge of the acoustic correlates of speech sounds and their contextual variation, and phonotactic constraints. The reasoning used in spectrogram reading tends to be qualitative in nature. Acoustic events are either present or absent, often extend over both time and frequency, and may occur simultaneously. While it is impossible to know what the expert spectrogram reader is thinking as the spectrogram is interpreted, it appears that much of the knowledge can be expressed as rules (Zue, 1981). However, few compilations of rules or strategies exist (Rothenberg, 1963; Fant, 1968). A knowledge-based system appears to be a natural medium within which to incorporate the knowledge, since it provides a means of understanding how the attributes and rules interact and how the system arrives at its decisions. Recently, several attempts have been made to build automatic speech recognition systems that model spectrogram reading directly (Johannsen, MacAllister, Michalek & Ross, 1983; Johnson, Connolly & Edmonds, 1984; Carbonell, Damestoy, Fohr, Haton & Lonchamp, 1986; Fohr, 1986; Stern, 1986; Stern, Eskénazi & Memmi, 1986; Zue & Lamel, 1986; Lamel, 1988a,b; Fohr, Carbonell & Haton, 1989; Meloni, Betari & Gilles, 1989; O'Kane, Keene, Landy & Atkins, 1989; Tattegrain & Caelen, 1989).

This paper reports on an attempt to formalize the knowledge used in spectrogram reading by incorporating it in a knowledge-based system. Since the emphasis was on formalizing the knowledge, a commercially available expert system shell, ART, was used for the implementation. Spectrogram reading knowledge is encoded in the descriptions of acoustic events visible in the spectrogram, and in the relationships between the acoustic events and phonemes. The acoustic events are described in terms of prototypes, and the relations between phonemes and acoustic events are expressed in a set of rules. The degree to which the knowledge has been formalized can be judged by the performance of the system, the types of errors made by the system, and the reasoning used. The remainder of this paper is as follows. First a brief introduction to spectrograms and spectrogram reading is provided, followed by a definition of the stop identification task. In Sections 3 and 4 the acquisition of knowledge and the knowledge representation are described, followed by the rules and strategy in section 5. The system evaluation and error analysis are provided in Section 6, and the performance is compared to that of an HMM-based phonetic recognizer on the same test data (Gauvain & Lamel, 1992).

2. Spectrograms and spectrogram reading

Since the invention of the sound spectrograph (Koenig, Dunn & Lacey, 1946), spectrograms have been used extensively by researchers in the speech community. In this research, the wide-band spectrogram, produced with a bandwidth of 300 Hz, has been used. Since the wide-band spectrogram is produced with a short time window, it

provides good temporal resolution, enabling accurate location of events in time (such as stop releases or the onset of voicing). In addition, formant frequencies and the spectral energy in noise-like regions are generally easy to resolve. Figure 1 shows an example spectrogram of the phonemic sequence /Ipl/, extracted from continuous speech. The spectrogram is augmented by three parameters: low frequency energy (LFE), total energy (TE) and center-clipped zero crossing rate (ZCR), along with the original waveform display. These parameters are useful to the spectrogram reader in identifying phonemes, particularly in regions where the acoustic energy is weak. Researchers may augment the spectrogram with other parameters. For example, Vaissiere (1983) has found that the fundamental frequency contour aids in interpreting spectrograms of French sentences.

Some humans have learned to interpret the visual acoustic patterns in the spectrogram so as to determine the identity of the spoken phonemes or words, a process known as spectrogram reading. In addition to providing a convenient mechanism for studying acoustic-phonetics (the relationship between phonemes and their acoustic correlates), spectrogram reading provides an opportunity to separate the acoustic characteristics of sounds from other sources of information, such as lexical, syntactic and semantic. It is difficult to assess the role of the different knowledge sources used by listeners interpret-

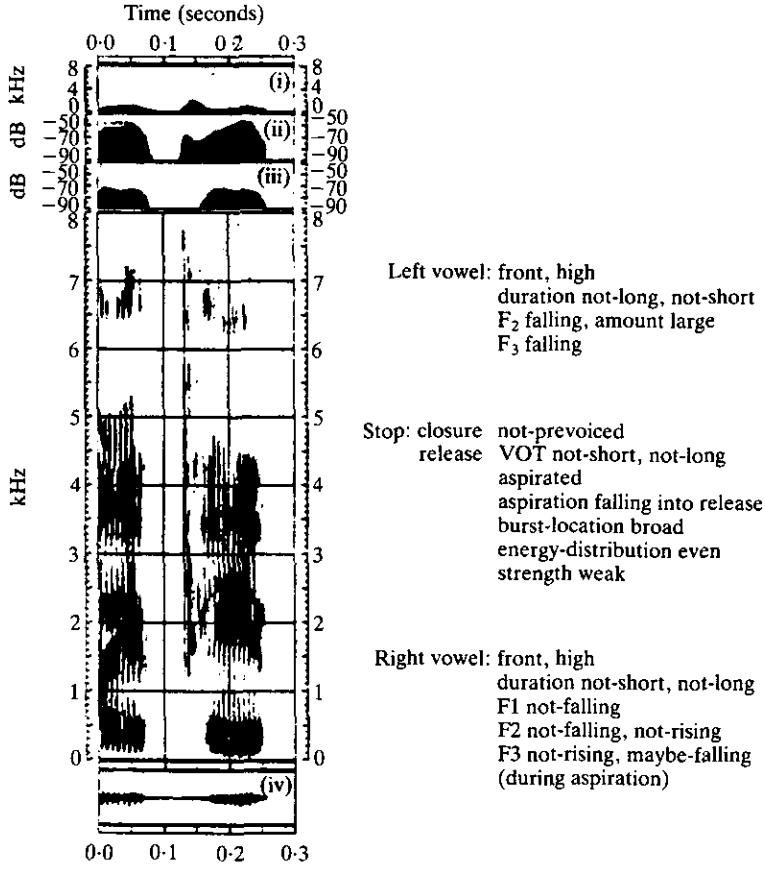

Figure 1. Facts in the dynamic database for the token /Ipl/. (i) Zero crossing rate; (ii) Total energy; (iii) Energy—125 Hz to 750 Hz; (iv) Waveform.

ing continuous speech. Spectrogram readers may be able to rely on their knowledge of the acoustic characteristics of speech sounds, how these characteristics change due to coarticulation, and on phonotactics, the allowable sequences of phonemes in the language. It appears that the spoken phonemes may be labeled in the spectrogram without considering word hypotheses (Cole & Zue, 1980; Lamel, 1988a,c).

Reading spectrograms entails identifying acoustic characteristics in the image and applying a variety of constraints to the identification problem. These constraints include knowledge of the acoustic correlates of speech sounds and their contextual variation, and phonotactic constraints. Some segments may be easily identified by recognition of their canonical characteristics, while others require a partial analysis of the neighboring segments. The skill also requires the ability to integrate multiple cues and to rely on secondary cues when the primary ones are not present.

Protocol analysis of the spectrogram reading process (Cole & Zue, 1980) shows there to be two stages, roughly corresponding to segmentation and labeling. Segmenting the speech involves placing boundaries to mark acoustic change. Experienced spectrogram readers often do not explicitly mark boundaries, but rather implicitly denote them via the labeling. Generally the easy segments, those whose spectral patterns are distinct and relatively context invariant, are labeled first. Then, with successive revisions, incorporating the local context and finer acoustic cues, the remaining segments are labeled. For the reader interested in an example of interpreting a spectrogram, see Zue (1981) or Lamel (1988a).

2.1. Domain knowledge/Existence of expertise

The development of a knowledge-based system requires a lot of domain-specific knowledge and an expert who can solve the problem. While there are still many unresolved questions in the production and perception of speech, a great wealth of knowledge exists. The domain knowledge includes our understanding of articulatory principles, acoustic-phonetics, and phonotactics. For a comprehensive review of the acoustic theory of speech production see Fant (1960) and Flanagan (1972). The evidence, obtained from both spectrogram reading experiments (Cole et al., 1980; Lamel, 1988b) and from teaching spectrogram reading, suggests that there are humans who qualify as experts. The expert must be able to explain his/her reasoning, as the process can only be modeled indirectly from the expert's own descriptions of his/her actions.

2.2 Task definition

The specific task investigated is the identification of stop consonants extracted from continuous speech. The stops occur in a variety of phonetic contexts selected to test the importance of knowledge sources thought to be used in spectrogram reading. A partial segmentation of the speech is provided and the acoustic descriptions are specified by the user or computed from the segmentation. Restricting the information to the segment to be identified and its immediate neighbors greatly reduces the complexity of the problem while retaining much of the contextual influences in American English.

2.3. Selection of a knowledge-based system shell

This research has focused on the acquisition and formalization of the knowledge base, rather than the development of a knowledge-based system, or shell, itself. As a result, existing technology has been used to implement a system for stop identification.

An initial implementation (Zue & Lamel, 1986) of a knowledge-base and a set of rules for stop identification used an available MYCIN-based (Shortliffe, 1976), backward-chaining system. Acoustic measurements were provided semi-automatically to the system and converted to qualitative descriptions. Rules related the qualitative descriptions to phonetic features, which were then mapped to phonemes. Beliefs in the preconditions reflected uncertainty in the acoustic descriptions. Strengths in the rule conclusions reflected how strongly a given acoustic description indicated a phonetic feature. The system set off to determine the identity of the stop, and in the process pursued the subgoals of deducing the voicing and place characteristics of the stop. In each case, the system exhaustively fired all pertinent rules.

The system (SS-1) was evaluated on 400 word-initial, intervocalic stops extracted from the MIT "Ice Cream" Corpus, containing 1000 sentences (10 from each of 50 male and 50 female speakers). Table I compares the system performance to the performance of human spectrogram readers on two sets of 100 stops. The averaged human performance of 2 and 3 spectrogram readers is given, for sets 1 and 2, respectively. The tokens in set 1 were also used to tune the system, which involved setting the thresholds for the mapping functions, and refining the selected acoustic descriptions and the rules. For set 1, the system's performance was comparable to that of the experts. The performance of the system degraded by 4% when it was confronted with new data (set 2), whereas the experts' performance remained high. The degradation of performance from tuning to test data was attributed primarily to the "lack of experience" of the system; it had not learned all the acoustic descriptions and rules used by the experts. The system had comparable performance on another test set of 200 samples (set 3).

If performance in terms of recognition accuracy was the main objective, the SS-1 system may have been acceptable. However, an important objective of this research was to develop a system that models the problem-solving procedures used by human experts, something that the SS-1 system did not do very well. This was partly due to limitations imposed by the structure of the MYCIN-based system. The inferencing of MYCIN did not enable the system to evaluate multiple hypotheses at any given time. In contrast, experts tend to use forward induction, and to simultaneously consider a set of possible

TABLE I. Comparison of human and SS-1 system identification performance

Condition		Number of tokens	Top choice (%)	Top 2 choice (%)
set 1	human (2)	200	90	92
	system	100	88	95
set 2	human (3)	200	92	96
	system	100	84	92
set 3	system	200	83	94

candidates, although they may use goal-directed reasoning to confirm or rule out candidates. Since there is redundancy in the acoustic characteristics for a given phonetic feature, often only a subset of acoustic characteristics are needed to specify it. The goal-directed control structure of MYCIN always exhaustively fired all rules, while experts may quit when they have enough evidence for a feature. Other problems occurred with representing our knowledge in MYCIN's data structure, the "context-tree". The MYCIN system did not allow nodes at the same level of the context-tree to share information, which made it difficult to model coarticulatory effects. As a result, it would have been difficult to increase the capabilities of the system to identify stops in other environments, such as consonant clusters.

Our experience with the SS-1 system indicated the need for a control strategy which better models the reasoning of spectrogram readers. The expert system shell ART, a commercial product developed by Inference Co., was selected because it integrates forward and backward reasoning, allows hypothetical reasoning and has "schemata" data-structures which provide frame-like capabilities. In addition, ART can be augmented to handle confidences in the preconditions and conclusions.

3. Knowledge acquisition

The knowledge incorporated in the implementation was obtained primarily by observing others reading spectrograms, by reading spectrograms myself, and by introspection. Using knowledge about the articulation of speech (and of stop consonants in particular) as a foundation, spectrograms of stop consonants were studied in order to define acoustic correlates of their place of articulation and voicing characteristic. An attempt was also made to determine how the acoustic evidence was weighed and combined in reaching a decision.

Over the extent of this research I was also fortunate to be involved in attending and leading several spectrogram reading groups. Spectrogram reading sessions provide a unique opportunity to gather knowledge. All those attending the session participate in the interpretation of the spectrogram, generally taking turns at identifying one or a few segments. When leading groups of beginning spectrogram readers, we usually try to have them identify easy sounds first (such as strong fricatives, /r/'s and other sounds with easily recognized acoustic correlates), leaving the more difficult interpretation until the end, when more contextual constraints can be applied. As the spectrogram readers gain experience, the spectrogram tends to be read from left-to-right, occasionally skipping over and returning to difficult regions. At his/her turn, each attendee proposes a label or sequence of labels for the region, and provides an explanation for his/her decision. When other spectrogram readers disagree, there can be extensive discussion as to possible consistent interpretations. At the sessions, particular attention was paid to the acoustic attributes used by the spectrogram readers and to the reasons they gave to justify their interpretation. Some of the sessions were tape recorded for further analysis.

Additional knowledge came from spectrogram reading experiments (Lamel, 1988a,b) and from system development. By analyzing the errors made by the expert spectrogram readers, I was able to assess some of the tradeoffs they made. For example, some spectrogram readers consistently favored the information provided by the burst location over that of the formant transitions. Others varied their strategy depending upon which information they felt was more robust in the given case. Each error was discussed with

the spectrogram reader who made it in order to elucidate the reader's reasoning. Implementing the system led to changes and refinements in the rules, particularly in the rule ordering. Rule development is an iterative, interactive process. Typically, a few examples were run through the system and, as a result, rules and rule interactions were modified.

4. Knowledge representation

The representation used for phonetic decoding combines knowledge from the acoustic theory of speech production (Fant, 1960) and distinctive feature theory (Jacobson, Fant & Halle, 1952).

4.1. Static knowledge base

Conceptually there are four levels of representation; phonemes (P), phonetic features (F), qualitative acoustic attributes (QAA) and acoustic measures (M). A block diagram of the representation is given in Fig. 2. Moving from left-to-right in the figure provides a top-down description of the knowledge. The links between the boxes mnemonically reflect their relationships. Phonemes are defined in terms of their phonetic features. Internally, phonemes are also grouped into classes reflecting their manner of articulation, such as stops, vowels and fricatives. Grouping phonemes into classes allows some of the rules of the system to be expressed more succinctly. For example, the features [+ obstruent, − continuant] are associated with the class of stops, and inherited by each member of that class. The phonetic features are related to a set of acoustic attributes, each of which takes on a qualitative value. The qualitative attributes describe acoustic events in the speech signal and the canonical temporal and spectral characteristics of the features. The qualitative attributes are based on our knowledge of the articulation of speech. They are events seen in a spectrogram or are derived from a quantitative acoustic measurement made in the speech signal.

Figure 3 shows a subset of the knowledge used to represent the class of stop

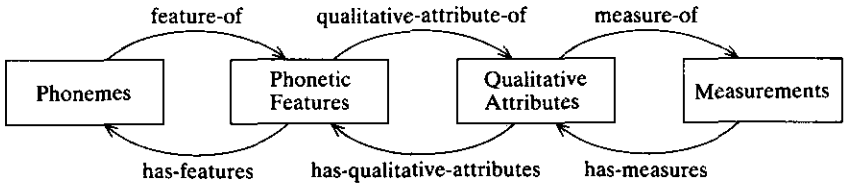

Figure 2. Knowledge representation.

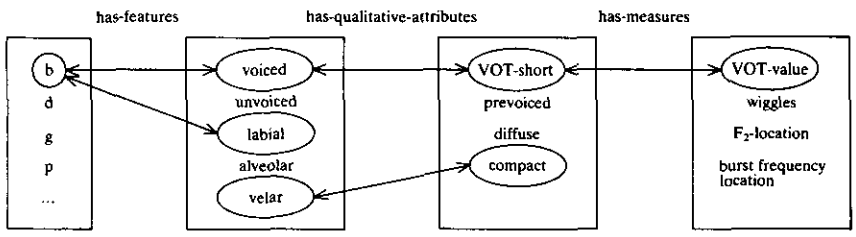

Figure 3. Subset of knowledge used to represent stops.

consonants. The stops, /b,d,g,p,t,k/, are represented by their place of articulation and their voicing characteristic. The voicing characteristic of a stop is determined primarily by the acoustic attributes of voice-onset-time (VOT), prevoicing, and aspiration. The place of articulation decision is based on acoustic attributes describing the frequency and time distribution of the burst, the aspiration (if the stop is aspirated) and the formant transitions into the surrounding vowels. The acoustic attributes take on qualitative values, each of which is associated with an acoustic measure. For example, the VOT is the time measured from the release of the stop to the onset of voicing in the vowel. A VOT of 25 ms would be mapped into VOT-short. Similarly, the distribution of energy across frequency at the stop release may be characterized as compact, diffuse, even, or bimodal. A compact energy distribution has the energy of the release primarily located in a frequency range of 1–2 kHz.

Vowels are also represented in the structure. The place of articulation of the vowel is determined by the tongue height and tongue position, and the position of the lips. The qualitative acoustic attributes associated with vowels describe the locations and movements of the formants. Acoustically, vowels are also described in terms of duration, which may be related to the tense/lax feature. The acoustic measures are the formant frequencies and the duration. For example, a back vowel has a high first formant (F_1) location and a low second formant (F_2) location, and an F_1 of 800 Hz is mapped to a high F_1. Semivowels, nasals and fricatives are represented analogously. Some of the place of articulation attributes for the fricatives and nasals are shared with the stops.

4.2. Dynamic knowledge base

In the preceding section the relationship between objects in the knowledge base were outlined. The knowledge is static in that it defines prototypes that do not change as a function of the rules. The representation of each stop can be thought of as a frame (Minsky, 1975), where the knowledge in the static database defines prototypes and the default values for each "slot" or attribute. A dynamic database of facts is created for each token as it is identified. The token exhibits specific acoustic characteristics which are converted to qualitative acoustic attributes. In turn, these qualitative acoustic attributes are used in phonetic decoding. The acoustic measures and qualitative acoustic attributes are obtained from the utterance or by querying the "user". The responses satisfy the preconditions of rules enabling them to fire, resulting in deductions and further queries for additional information. The framework allows the queries to be replaced with function calls to measure parameters directly or with database requests to retrieve prestored measures and/or attributes.

4.3. Qualitative acoustic attributes

The qualitative acoustic attributes (QAAs) describe the acoustic events visible in the spectrogram. Each segment is represented by a set of QAAs specific for the type of segment. Table II lists some examples of qualitative acoustic attributes used to describe the stop consonants. Each QAA is used to determine the place or the voicing of the stop. The QAAs represent characteristics of the stop release, the closure interval and the formant transitions into the surrounding vowels. The stop in Fig. 1 has the qualitative acoustic attributes listed in the figure.

TABLE II. Examples of qualitative acoustic attributes of stops

Dimension	Region	Attribute
voicing	release	VOT-short
		VOT-long
		aspirated
	closure	prevoiced
place	release	burst-location-HF
		burst-location-MF
		burst-location-LF
		burst-location-bimodal
		energy-distribution-diffuse
		energy-distribution-compact
		energy-distribution-even
		energy-distribution-bimodal
		strength-strong
		strength-weak

QAAs are obtained by querying the user or by mapping acoustic measures. Certain combinations of qualitative acoustic attributes cannot co-occur. For example, it would be meaningless to have a burst-strength that was both weak and strong. To prevent such a situation, the rules that query the user for information take into account the facts already known. For example, if the user responds that burst-strength is strong, then the system will not query to determine if the burst-strength is weak, but instead automatically asserts that the burst-strength is not-weak.

4.4. Probing the knowledge base

Some facilities were developed for probing both the static and dynamic knowledge bases. A *what-is* or *what-has* question returns a table look-up or definitional answer. A *why* or *why-not* question is generally used to justify in a specific example. Examples of the types of queries and the forms of responses are given in Table III.

The system response to the query "*what-is a /p/?*" is that /p/ is a voiceless, labial stop. The response to "*what-is voiced?*" is a list of all the QAAs that specify voiced: short-VOT, prevoiced, and not-aspirated. The answer to the query "*what-has the feature voiced?*" is the set of stops /b,d,g/. The *why-not* query is used to ask the system why a deduction was not made. The system responds with a list of missing and/or contradictory information.

5. Rules and strategy

Plausible strategies for some of the cognitive aspects of spectrogram reading are simulated through the rules. While it cannot be verified that spectrogram readers use these or similar strategies, the strategies "feel right" to the expert. Much of the reasoning is data-driven—the spectrogram reader sees acoustic events in the spectrogram and makes deductions based on them. The spectrogram reader takes into account contextual variation due to coarticulation, is able to combine multiple cues in forming a judgement,

TABLE III. Examples of the types of queries recognized by the system

Question	Object	Answer
what-is	phoneme	feature bundle
	feature	set of acoustic attributes
	acoustic attribute	description (value in context)
what-has	feature(s)	phonemes having feature(s)
	acoustic attribute	features having QAA
why	phoneme	associated deduced features
	feature	associated QAA's
why-not	phoneme	missing features
	feature	missing or contradictory QAA

and often considers multiple hypotheses at once. Goal-directed reasoning may be used to confirm or rule out hypotheses. Spectrogram readers are also able to deal with uncertainty in the acoustic evidence and, to some degree, with acoustic information that may be contradictory.

An attempt has been made to capture most of the above cognitive aspects in the implementation. The implementation integrates data-driven and goal-directed reasoning. The data-driven rules make deductions based on the qualitative acoustic attributes. Goal-directed reasoning is used to query the user (or database) for new information and to confirm or rule out hypotheses. The system models the human capability to simultaneously consider multiple hypotheses by maintaining a ranking of all candidates at all times. The rules may be one-to-one, as linking phonetic features and phonemes, or one-to-many and many-to-one, as in deducing phonetic features from qualitative acoustic attributes. Thus, the rules provide the capability to handle the problems of multiple cues and multiple causes. How spectrogram readers actually combine information from multiple sources and deal with uncertain evidence has not been determined; however, what he/she appears to be doing is modeled. Uncertainty in the acoustic evidence is modeled by allowing users to specify that an acoustic attribute is present, absent, or "maybe" present. Constraining the system to use uncertain acoustic attributes only after using definitive ones provides a mechanism for relaxing constraints under uncertainty. Uncertainty in the deductions is handled by associating a strength with each deduction.

5.1. Rules

Different rule sets cover the relations between levels in the representation. Rules map phonetic features to phonemes, relate qualitative acoustic attributes to phonetic features, and map acoustic measurements to qualitative acoustic attributes.

5.1.1. Definitional rules

A set of "definitional rules" map the phonemes to their phonetic features. The representation of stops according to their place of articulation and their voicing characteristic is shown in Table IV. The rules encode the information in the table explicitly. An example of a definitional rule is:

If the voicing of the stop is *voiced*,
and the place of articulation of the stop is *alveolar*,
then the identity of the stop is */d/*.

While conceptually there are different definitional rules for each stop, they are all implemented with one rule. The following rule explicitly captures the knowledge that a stop can be described in terms of its voicing characteristic and its place of articulation. The rule also combines the beliefs associated with each feature to determine a belief in the identity of the stop.

If the voicing of the unknown-stop is *voicing-value* with *voicing-belief*
and the place of articulation of the unknown-stop is *place-value* with *place-belief*
and there exists a prototype stop with identity *identity*
and with voicing *voicing-value*
and with place of articulation *place-value*
then the identity of the unknown-stop is *identity* with belief(*voicing-belief*,*place-belief*).

5.1.2. Rules relating qualitative acoustic attributes to features

The relationships between the qualitative acoustic attributes and the phonetic features are complicated. The majority of the rules in the implementation deal with these relationships. The rules are all of the form:

If precondition(s)
then conclusion(s) with strength(s).

The preconditions are generally facts that exist in the database. However, the absence of a fact may also be used as a precondition: whenever the fact exists, it serves to inhibit the rule from firing.

A given phonetic feature may be signalled by several qualitative acoustic attributes, resulting in multiple rules to deduce the phonetic feature. For example, both a long-VOT and the presence of aspiration in the stop release are cues for the feature voiceless. The corresponding two rules are:

If the VOT is *long*,
then there is *strong evidence* that
the voicing characteristic is *voiceless*.

If the release is *aspirated*,
then there is *strong evidence* that
the voicing characteristic is *voiceless*.

If, as in the example of Fig. 4(c), the preconditions of both of the rules are satisfied, then the belief that the voicing of the stop is voiceless will be quite strong. Not all the qualitative acoustic attributes for a phonetic feature are always present. For any particular acoustic segment, some or all of the rules may have their preconditions satisfied and those rules will fire.

TABLE IV. Phonetic features of stops

	b	d	g	p	t	k
voiced	+	+	+	−	−	−
labial	+	−	−	+	−	−
alveolar	−	+	−	−	+	−
velar	−	−	+	−	−	+

A given qualitative acoustic attribute may be indicative of different phonetic events, resulting in rules that have multiple deductions with respect to the context. For example, in the absence of contextual information, a burst spectrum that has energy predominantly at high frequencies is likely to indicate an alveolar place of articulation. However, if the stop is in a syllable with a front vowel, the stop is also likely to be velar and may be labial. The contextual influences are directly incorporated into the rules as follows:

If the burst-location is *high-frequency*,
then there is *strong evidence* that
the place of articulation is *alveolar*.

If the burst-location is *high-frequency*,
and the vowel is *front*
then there is *strong evidence* that
the place of articulation is *velar*
and there is *weak evidence* that
the place of articulation is *labial*.

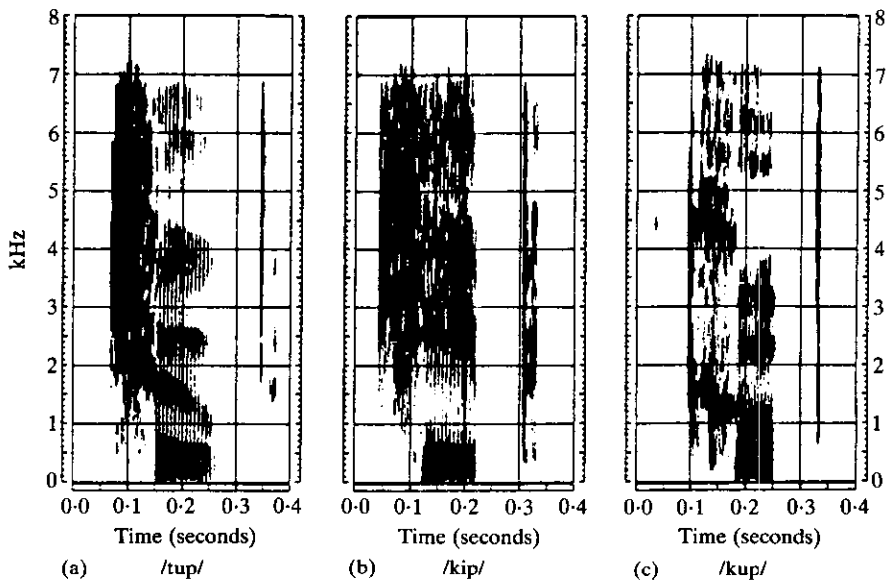

Figure 4. Spectrograms illustrating contextual variation.

Figure 4 illustrates an example requiring the use of such contextual information. The spectral characteristics of the stop release in the spectrograms in (a) and (b) are quite similar: they both have a predominance of high frequency energy. In this example, it would not be easy to determine the identity of either stop only by visual inspection of the release. The spectral characteristics of the release are consistent with both a /t/ and a front-/k/. However, knowledge that the following vowel in (a) is an /u/ indicates that the stop is a /t/. The spectral characteristics of a back, rounded /k/ in the syllable /ku/ are quite different, as can be seen in the spectrogram in Fig. 4(c).

The presence or absence of acoustic evidence may be important. For example, if a stop is syllable-initial and has a VOT value that is medium (maybe-short and maybe-long), then the presence of aspiration indicates that it is voiceless, and the absence of aspiration is indicative of voiced. Two rules that use aspiration to deduce the voicing characteristic are:

If the stop is syllable-initial
and the VOT is *maybe-long* and *maybe-short*,
and the release is *aspirated*,
then there is *strong evidence* that
the voicing characteristic is *voiceless*.

If the stop is syllable-initial
and the VOT is *maybe-long* and *maybe-short*,
and the release is *not-aspirated*,
then there is *medium evidence* that
the voicing characteristic is *voiced*.

Note that the presence of aspiration is a stronger indicator of voiceless than the lack of aspiration is of voiced. In other cases, the presence of an acoustic attribute may indicate a feature, but the absence of the acoustic attribute does not provide negative evidence for that feature. One such acoustic attribute is a "double burst". When a double burst is observed it is a strong indicator of a velar place of articulation. However, since a double burst is not that common, the system must have some evidence that the place of articulation is velar before attempting to find a double burst. The double-burst rule is:

If the place of articulation is *velar* with *belief*
and the release has a *double burst*
then there is *strong evidence* that
the place of articulation is *velar*.

The value of the voicing characteristic and of the place of articulation are deduced independently. While it is possible to deduce phonemes directly instead of features, deducing phonetic features adds another level of representation and generalization. This allows commonality in the rules for place or voicing to extend to different manner classes. For example, vowels and nasals are both shorter preceding voiceless consonants than voiced consonants in the same syllable. This phonological effect can be captured in one rule, instead of individually for each phoneme. The formant motion between a vowel and a consonant depends primarily on the place of articulation of the consonant, and not on its identity. Thus, for example, the qualitative acoustic attribute of falling formants can be associated with the feature labial, covering multiple phonemes.

Phonotactic constraints are implemented in rules which account for the phonetic context. For example, if the stop is preceded by a fricative, the system will attempt to determine whether the fricative is an /s/ or a /z/. If the fricative is a /z/, the system asserts that the stop is syllable-initial. If the fricative is an /s/, the system must determine whether or not the /s/ and the stop form a cluster. If the stop is in a cluster with the /s/, then the stop is voiceless. If the stop is not in a cluster with the /s/, then there is a syllable boundary between the fricative and the stop, and the syllable-initial rules to determine the voicing of the stop may be applied.

The context is specified in the preconditions of the rules to ensure that they fire only under the appropriate conditions. When the stop is preceded by a fricative, the formant motion in the vowel to the left of the fricative is not used, since the formant transitions should always indicate the alveolar place of articulation of the fricative. This is implemented by preconditions in the vowel formant rules which specify that the right context cannot be a fricative. For a stop preceded by a homorganic nasal, the formant motion in the vowel preceding the nasal is used to infer the place of articulation of the nasal, which is the same as that of the stop.

5.1.3. Mapping rules

The mapping rules convert acoustic measurements into qualitative attributes. The mapping rules are implemented as backward-chaining rules, and therefore do not fire unless they are called upon to produce a result needed by another rule. The mappings are schematically illustrated in Fig. 5. The rules which map from numerical quantities into qualitative acoustic attributes are of the form:

If the measured-value is $< a$
then the attribute has the qualitative-value *short*
else if the measured-value is $> b$
  then the attribute has the qualitative-value *long*
otherwise the attribute has the qualitative-values
  maybe-short and *maybe-long*.

The mapping rules typically divide the range into disjoint regions, where measures falling between regions are associated with both labels. The mapping ranges were hand-selected by looking at histograms of the measure on a set of training samples. However, these could be statistically trained if enough data were analysed.

5.2. Control strategy

Spectrogram readers appear to extract acoustic attributes in the spectrogram and to propose a set of features consistent with the attributes. The candidate set is refined by looking for additional acoustic evidence to confirm or rule out some of the possibilities. The control strategy attempts to model the behavior of spectrogram readers. The order in which the rules fire is controlled by priorities associated with the rules and by the use of preconditions so as to have the behavior of the system appear more "intelligent".

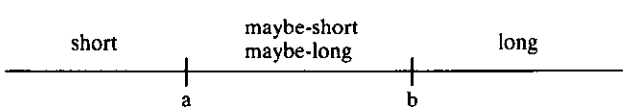

Figure 5. Example of mapping ranges for numerical quantities.

The system uses general information before using more specific information. This is implemented by associating higher priorities with the more general rules. Preconditions are also used to prevent the system from asking for detailed information too early. An example was shown in the double-burst rule, where there had to be some belief that the place of articulation was velar before the rule could be applied.

The system maintains a ranking of all candidates. Each time a new fact is asserted into the database, the ranking rules reorder the candidates for place and voicing. In this way, the system simultaneously considers multiple hypotheses at the same time. The list of ordered candidates enables the system to pursue the most likely candidate first. The rules are ordered so as to seem logical to the user. For example, if labial is the top candidate, the system tries to determine if the stop has a weak release, as a weak release provides confirming evidence for labial. If the top candidate is alveolar, and the second candidate is labial, the system will attempt to find out if the release is strong, as a strong release favors alveolar over labial. However, if the top two candidates are alveolar and velar, rules using the strength of the release are postponed, since the release strength is not a good attribute to distinguish between them.

Without specific "termination" rules, the system exhaustively fires all rules until there are no more left. However, the system may be run in a mode where, when it has enough evidence for a feature, it does not exhaustively pursue all the alternatives. This behavior is implemented by rules which attempt to confirm the top candidate and rule out the closest competitor when the belief in the top candidate is large enough. If the belief in the top candidate (and the distance between the top two candidates) increases, then the system confirms the top candidate and no longer attempts to determine the value of the feature.

5.3. Combining evidence

The rules provide evidence that a given feature has a particular value. Since there are multiple rules which deduce the same feature, some way is needed to combine the evidence from the different rules. Combining evidence is an unsolved problem in expert systems research. There have been different approaches to the problem, including probabilistic, such as Bayesian (Duda, Hart & Nilsson, 1976), fuzzy logic (Zadeh, Fu, Tanaka & Shimura, 1975), and more ad hoc formulations (Keeney & Raiffa, 1976; Shortliffe, 1976).

The goal in building a knowledge-based system is to use domain knowledge to solve the problem. By using rules that are based on our knowledge of the articulation of speech and our experience in spectrogram reading, the hope is that reasonable performance can be obtained with a small amount of training data. Some properties that a reasonable scoring scheme for this application should have are:

- The scoring must be able to handle both positive and negative evidence.
- The combining of evidence should be order independent.
- The combining of evidence should be monotonic. Positive evidence can never decrease the belief in something and negative evidence can never increase it.
- Since the rules assert conclusions with strengths, the combining should also preserve the relative strengths of the conclusions. A weak conclusion cannot increase the belief more than a strong one can. The converse is also true. In addition, a strong positive conclusion and a weak negative conclusion cannot combine to reduce the belief in something.

Two simple scoring schemes satisfying the above properties have been investigated. The first assigned numerical values to weak, medium, strong and certain evidence, and summed the evidence. Positive evidence was added, and negative evidence subtracted. The numbers used were:

with-certainty = 1·0
strong-evidence = 0·8
medium-evidence = 0·5
weak-evidence = 0·2

The second scheme counted the number of reasons of each strength and ranked the candidates according to lexicographic ordering (Keeney & Raiffa, 1976).

6. System evaluation

The performance of the system was compared to human performance on two sets of tokens covering a variety of phonetic contexts. The first set contained tokens that were *heard correctly* by all listeners *and* were *read correctly* by all spectrogram readers (AC). The second set contained tokens that were *misheard or misread* by at least one subject (SE). The tokens were selected from a larger set of tokens which had been used to evaluate the abilities of human listeners and human spectrogram readers to identify stop consonants (Lamel, 1988a), so as to include roughly equal numbers of each stop. The tokens were extracted from continuously spoken sentences taken from the DARPA TIMIT Acoustic-Phonetic Corpus (Garofolo, Lamel, Fisher, Fiscus, Pallett & Dahlgren, 1992) and the MIT Ice Cream Corpus. The test tokens come from 139 different speakers (78 male, 61 female), with approximately half of the tokens from each sex. The test tokens cover the five tasks shown in Table V. Each speech token contains the stop (or cluster) of interest, and the preceding and following vowels in their entirety.

TABLE V. Definition of phonetic contexts for the 5 recognition tasks

	Description	Grammar
Task 1	syllable-initial singleton stops	V S V
Task 2	syllable-initial stops preceded by /s/ or /z/	V F S V
Task 3	syllable-initial stops in semivowel clusters and affricates	V S SV V or V A V
Task 4	non-initial singleton stops	V S V
Task 5	non-initial homorganic nasal-stop clusters	V N S V

V: vowel; S: stop {b,d,g,p,t,k}; F: fricative {s,z}; SV: semivowel {l,r,w}; A: affricate {čJ}; N: nasal {m,n,ng}.

6.1. Results

System performance for each of the five tasks is given in Table VI for the two subsets of tokens. Both the top choice[1] and top-two choice accuracies are provided. The system performance was about the same for syllable-initial singleton stops (task 1) and syllable-initial stops preceded by /s/ or /z/ (task 2). The system identified singleton stops better when they occurred in syllable-initial position (task 1) than in non-syllable-initial position (task 4). Performance was better for non-initial stops in nasal clusters (task 5) than for singleton non-initial stops (task 4).

As expected, the performance on the AC subset (87%) was better than on the tokens that were misidentified by humans (71%). The system failed to propose the correct candidate for 2% of the AC tokens and 12% of the SE tokens. The system performance varied across context in a manner similar to that observed for human spectrogram readers (Lamel, 1988a). The system performance on multiple tasks is seen to be comparable to or better than that of the original Mycin-based SS-1 system on syllable-initial stops.

6.1.1. AC tokens

Even though listeners and spectrogram readers were able to identify the tokens in this set, the system made errors on 12 of the 94 tokens. The second candidate was correct in 9 of the 12 errors. In half of the errors, the system's top choice was listed as an alternate candidate by a spectrogram reader. Averaged across the tasks, 75% of the errors were in place of articulation and 17% were in voicing.

Most of the errors are reasonable, even though there may have been acoustic evidence for the correct answer. A few examples of tokens on which the system made errors are shown in Fig. 6. The errors made on the two left tokens are more reasonable than the errors for the two tokens on the right. The leftmost token was called /d/, with /b/ as a

TABLE VI. Knowledge-based system evaluation on five tasks. Task 1: syllable-initial singleton stops, Task 2: syllable-initial stops preceded by /s/ or /z/, Task 3: syllable-initial stops in clusters with /l,r,w/ and affricates, Task 4: non-syllable-initial singleton stops, Task 5: non-syllable-initial stops in homorganic nasal clusters

	Percent correct: top/top 2			
	N	AC	N	SE
Task 1	24	88/96	27	82/93
Task 2	26	89/96	11	64/73
Task 3	14	100/100	17	82/94
Task 4	18	67/89	19	58/68
Task 5	12	100/100	6	83/100
Overall	94	87/96	80	71/85

[1] When there was a tie for the top choice, the system was credited with having identified the stop correctly. Ties occurred on only 6 of the 174 tokens.

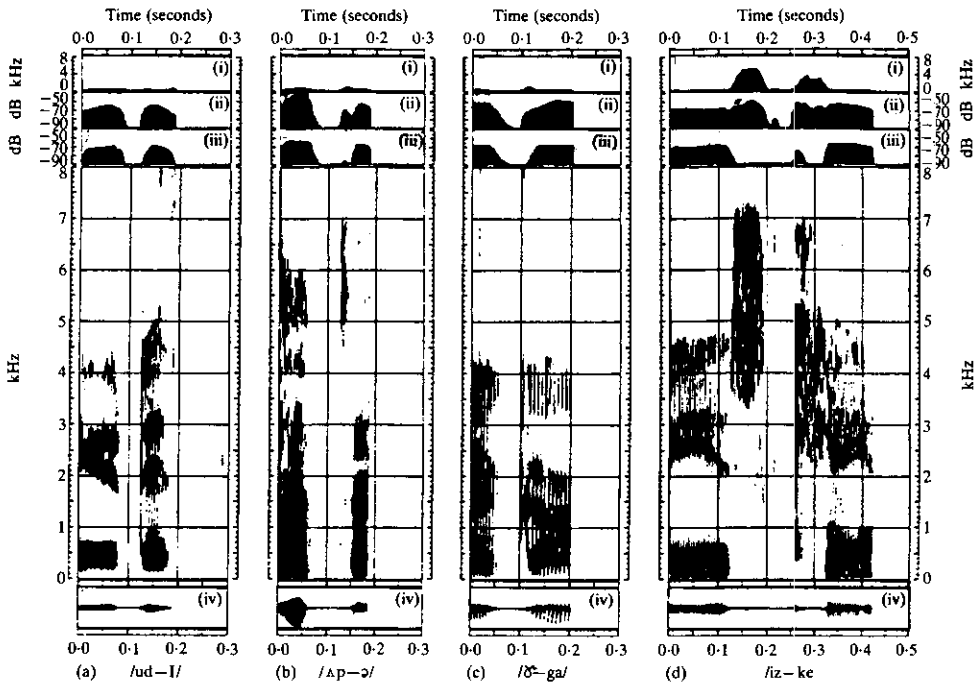

Figure 6. Examples of system errors on AC tokens. (i) Zero crossing rate; (ii) Total energy; (iii) Energy—123 Hz to 750 Hz; (iv) Waveform.

second candidate. The burst release is located primarily at mid frequencies (2–4 kHz), a lower frequency location than is expected for an alveolar stop between front vowels. However, the preceding vowel is a fronted-/u/, which would probably not be fronted if the stop was labial. However, the system did not have the information that the preceding vowel is fronted. The formant transitions are also slightly better for alveolar than for labial.

The middle two spectrograms both have conflicting evidence between the burst characteristics and the formant transitions. The formant transitions in the vowels surrounding the /p/ in (b) favor a labial place of articulation, but the high frequency concentration of energy in the release supports alveolar. While the weak release of /g/ into (b) suggests labial, the formant transitions from the stop into the next vowel /a/ are incompatible with labial. In this case, the system favored labial quite strongly, even though the spectrogram readers ruled labial out. While the rightmost spectrogram was called a /t/, /k/ was a close second choice. The distinguishing acoustic cue is subtle, but there is more energy in the release around 3 kHz than there is at higher frequencies, supporting velar.

6.1.2. SE tokens

The tokens in the SE subset had an error by at least one spectrogram reader or one listener. The system made errors on 23 of the 80 tokens. The same error was made by a listener in 14 of these cases. A spectrogram reader made the same error as the system on 11 of the tokens, and supplied the system's top choice as an alternate candidate for another 5. Only in 2 instances did the system propose an answer that did not agree with any of the human subjects (listeners or readers).

6.1.3. Performance with termination

The system was also evaluated using termination rules. If there was sufficient evidence for a phonetic feature, the system stopped attempting to determine that feature. Performance using the termination rules was essentially unchanged. The termination rules applied 53% of the time for voicing, 48% of the time for place, and 27% of the time for both voicing and place. The system terminated early more often on AC tokens than on SE tokens.

6.2. Comparison with an HMM-based phone recognizer

In order to compare the performance of this knowledge-based approach to statistically-based approaches, an HMM-based phonetic recognizer (Gauvain & Lamel, 1992) was evaluated on the same set of test tokens. The phone recognizer was trained on the TIMIT corpus using the newly defined training and test partitions as specified in Garofolo et al., 1992.[2] The 16 kHz speech signal was Mel frequency bandpass filtered and 15 cepstral coefficients were computed every 10 ms. There are 52 context-independent phones models[3], each of which is a 3-state left-to-right HMM with Gaussian mixture observation densities. Duration is modeled with a gamma distribution per phone model. Maximum likelihood estimators were used for the HMM parameters and moment estimators for the gamma distributions. In addition, the HMM recognizer was provided with a task grammar (see Table V) to use during recognition, which supplied the same information which had been given to the knowledge-based system and the human subjects.

The results of the evaluation are given in Table VII for the AC and SE tokens. Only stop identification errors were counted—errors in vowel or fricative identification were ignored. Comparing Tables VI and VII, it can be seen that overall the knowledge-based system outperforms the HMM-based system (83% vs. 73·5%). The HMM-based system performs better on task 4 for the AC tokens, and on task 2 for the SE tokens. For the AC tokens, 53% of the errors are in voicing alone, 18% in place of articulation, and 29% in both. An even larger percentage of the errors (75%) are in voicing for the SE tokens. The overall voicing accuracy is 85% for the AC tokens and 72% for the SE tokens, compared to 97% (AC tokens) and 84% (SE tokens) for the knowledge-based system. That the HMM-based system makes more errors on voicing suggests that duration, which is a primary cue for voicing, is not particularly well-modeled. This contrasts with the knowledge-based system which had most of the errors in place of articulation. The HMM-based system correctly identifies place of articulation for 91% of the tokens compared to 89% for the knowledge-based system.

6.3. Sensitivity due to scoring

The same test set of 100 syllable-initial stop consonants used to evaluate SS-1 system (see Section 2.3) was also used to assess the importance of the strengths associated with the

[2] For reference, our 52 CI model set had an overall phone accuracy of 60.1% (65.1% with 39 folded phones) when evaluated on the TIMIT core test set (Garofolo *et al.*, 1992). This performance is superior to that reported by Lee and Hon (1989), using a smaller set of 48 CI phones. Both Lee and Hon (1989) and Robinson and Fallside (1991) have reported higher phone accuracies using context-dependent models, on a different training/testing subdivision of TIMIT.

[3] The same experiment was run using sets of 446 and 1669 context-dependent phone models, but on average no difference in performance was observed.

TABLE VII. HMM-based system evaluation on five tasks. Task 1: syllable-initial singleton stops, Task 2: syllable-initial stops preceded by /s/ or /z/, Task 3: syllable-initial stops in clusters with /l,r,w/ and affricates, Task 4: non-syllable-initial singleton stops, Task 5: non-syllable-initial stops in homorganic nasal clusters

	Percent correct: top			
	N	AC	N	SE
Task 1	24	79	27	63
Task 2	26	81	11	82
Task 3	14	86	17	71
Task 4	18	78	19	53
Task 5	12	92	6	67
Overall	94	82	80	64

rule conclusions. Two experiments were conducted. In the first experiment the rule strengths were all set to 1·0, eliminating the distinction between weak and strong evidence. The resulting error rate of 27%, shown as "count" in Fig. 7, is almost double that of the baseline system, "add", which has a 15% error rate. (For comparison, two human spectrogram readers had a 90% top choice identification accuracy for the same tokens.) In a second experiment, a random number in the range [0, 1·0] for positive evidence and [−1·0, 0] for negative evidence, was generated and summed. This experiment was conducted 10 times. Shown as "random", the mean error rate was 30% with a standard deviation of 4%. Both of these experiments indicate that the rule strengths associated with the rule conclusions are important and that not all the evidence should be treated equally.

To evaluate the dependency of the performance on the selection of the numerical values, experiments were conducted in which a random number, in the range [−a,a], was added to the score at each update. The system was evaluated 10 times each for a = 0·1, 0·2, 0·3, and 0·4. The results are shown in Figure 7. The difference in the mean error rate from the baseline of 15% is insignificant at the 0·05 level for all cases. For a = 0·3, 0·4 the difference is significant at the 0·01 level. These experiments indicate that the system is relatively insensitive to the numerical values assigned to the strengths.

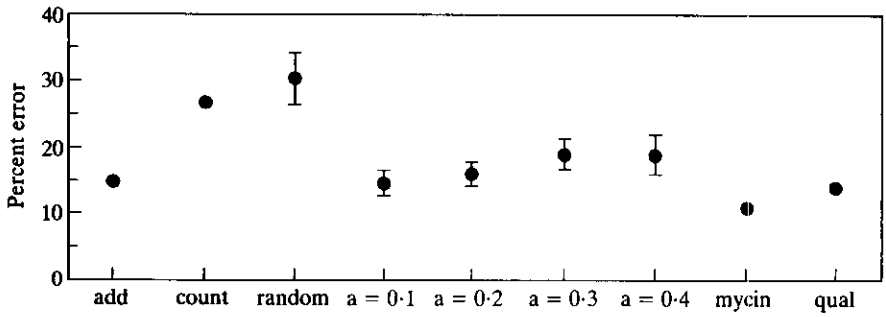

Figure 7. Comparison of scoring strategies

The remaining two data points in Figure 7 show the performance of the system using two other scoring strategies. The point labeled "mycin" used the EMYCIN-combine function (Shortliffe, 1976). The point labeled "qual" used the lexicographic scoring. That the system performance is comparable to the "add" method illustrates the robustness of the system to changes in scoring methods.

7. Discussion

The most difficult part of the system implementation, and by far the most time-consuming, was controlling the system to have acceptable behavior. Because of the way backward chaining is integrated in ART, different rules had to be written for almost every query in the system, with each having a priority to ensure that it fired appropriately.

In order to simulate the concept "use information that is known to be true, before using evidence that might be true", the rules had to be duplicated, assigning a lower priority to rules based on uncertain evidence. It would have been better if a "meta-rule" could have been written that said to use certain evidence before using uncertain evidence. Duplicating rules requires extra care when the rules are modified, as the modifications may have to be made in multiple copies of the rule.

It is difficult to model the human ability to selectively pay attention to acoustic evidence, particularly when there is contradictory evidence. By design, the system is relatively conservative and rarely ignores evidence, even in the presence of conflicting evidence. At times human experts will say things like "the formants are better for alveolar than velar, but I like the release so much better as velar, that I'm willing to ignore the formant transitions". The system is reluctant to use evidence as strongly as a spectrogram reader will, as the conditions under which readers do so are not well understood.

A related issue is the identification of the acoustic attributes. It may be relatively easy to develop algorithms to locate some of the attributes, such as location of the energy in the release, and the strength of the release. Other attributes might be quite difficult. For example, the formant transitions between stops and vowels may occur over a short time interval, such as the 30 ms before and after the stop. Humans are often able to determine the formant motion, while the problem for formant tracking is still unsolved despite many efforts.

8. Summary

Knowledge obtained from spectrogram reading was incorporated in a rule-based system for stop identification. The emphasis was on capturing the acoustic descriptions and modeling the reasoning thought to be used by human spectrogram readers. Because there is ambiguity in relating acoustic events to the underlying phonemic representation, multiple descriptions and rules were used. The reasoning of the system "feels" acceptable to a spectrogram reader. The system simultaneously considers multiple hypotheses and maintains a ranking of hypotheses for each feature independently. Evaluation on a set of tokens from a variety of contexts indicated that the errors made by the system were often reasonable and in agreement with spectrogram readers and listeners. The system performance was higher than the performance of an HMM-based system trained on TIMIT and tested on the same test tokens. The system was shown to

be relatively insensitive to changes in scoring strategies, and it's level of performance indicates that knowledge formalization has been somewhat successful. However, the ability of human spectrogram readers and listeners surpasses that of the knowledge-based system, indicating the need for additional knowledge. There appears to be much more happening in our visual system and in our thought processes than we actually express, even when asked to explain our reasoning.

References

Carbonell, N., Damestoy, J. P., Fohr, D., Haton, J. P. & Lonchamp, F. (1986). APHODEX, design and implementation of an acoustic-phonetic decoding expert system. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing ICASSP-86*. Tokyo, Japan, pp. 1201–1204.

Cole, R. A., Rudnicky, A. I., Zue, V. W. & Reddy, D. R. (1980). Speech as patterns on paper. In *Perception and Production of Fluent Speech*, (Cole, R. A., ed.), pp. 3–50. Lawrence Erlbaum, Hillsdale, New Jersey.

Cole, R. A. & Zue, V. W. (1980). Speech as eyes see it. In *Attention and Performance VIII*, (Nickerson, R. S. ed.), pp. 475–494. Lawrence Erlbaum, Hillsdale, New Jersey.

Duda, R. O., Hart, P. E. & Nilsson, N. J. (1976). Subjective Bayesian methods for rule-based inference systems. In *Proceedings of National Computer Conference AFIPS*, **45**, pp. 1075–1082.

Fant, G. (1960). *Acoustic Theory of Speech Production*, Mouton, The Hague.

Fant, G. (1968). Analysis and synthesis of speech process, *Manual of Phonetics*, (Malmberg, B. ed.), pp. 173–277. North Holland, Amsterdam.

Flanagan, J. L. (1972). *Speech Analysis Synthesis and Perception*. Springer-Verlag, Berlin.

Fohr, D. (1986). *APHODEX: Un système expert en décodage acousticophonétique de la parole continue*, Thèse de Doctorat d'Université en informatique, Université de Nancy I, France.

Fohr, D., Carbonell, N. & Haton, J. P. (1989). Phonetic decoding of continuous speech with the APHODEX expert system. *Proceedings of Eurospeech-89*. Paris, France, **2**, pp. 609–612.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. & Dahlgren, N. L. (1992). The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM (printed documentation for NIST Speech Disc 1-1.1), NTIS order number PB91-100354.

Gauvain, J. L. & Lamel, L. F. (1992). Speaker-independent phone recognition using BREF. *Proceedings of the DARPA Speeach and Natural Language Workshop*, Arden House, Feb. 1992. New York, pp. 344–349.

Jacobson, R., Fant, C. G. M. & Halle, M. (1952). Preliminaries to Speech Analysis, MIT Acoustics Laboratory, Technical Report No. 13.

Johannsen, J., MacAllister, J., Michalek, T. & Ross, S. (1983). A speech spectrogram expert. *Proceedings of the IEEE ICASSP-83*, Boston, MA, pp. 746–749.

Johnson, S. R., Connolly, J. H. & Edmonds, E. A. (1984). Spectrogram Analysis: A Knowledge-Based Approach to Automatic Speech Recognition, Leicester Polytechnic, Human Computer Interface Research Unit, Report No. 1.

Keeney, R. L. & Raiffa, H. (1976). *Decisions with Multiple Variables: Preferences and Value Tradeoffs*. John Wiley & Sons, New York.

Klatt, D. H. (1977). Review of the ARPA speech understanding project. *Journal of the Acoustical Society of America*, vol. 62, **6**, 1345–1366.

Koenig, W., Dunn, H. K. & Lacey, L. Y. (1946). The sound spectrograph. *Journal of the Acoustical Society of America*, vol. 18, **1**, 19–49.

Lamel, L. F. (1988a). *Formalizing Knowledge Used in Spectrogram Reading: Acoustic and perceptual evidence from stops*, PhD. Thesis. Massachusetts Institute of Technology, Cambridge, Massachusetts.

Lamel, L. F. (1988b). A knowledge-based system for stop consonant identification. *Proceedings of IEEE International Computer Science Conference '88: Artificial Intelligence, Theory and Applications ICSC-88*, Hong Kong, December, 1988, pp. 558–566.

Lamel, L. F. (1988c). Spectrogram readers' identification of stop consonants. *Proceedings of the Second Australian International Conference on Speech Science and Technology SST-88*, Sydney, Australia, November 1988, pp. 92–97.

Lee, K.-F. & Hon, H.-W. (1989). Speaker-independent phone recognition using hidden Markov models. *Proceedings of the IEEE Transactions ASSP*, vol. 37, **11**, pp. 1641–1648.

Meloni, H., Betari, A. & Gilles, P. (1989). A knowledge-based system for speaker-independent recognition of letters. In *Proceedings of Eurospeech-89*, Paris, France, **2**, pp. 625–628.

Minsky, M. (1975). A framework for representing knowledge. *The Psychology of Computer Vision* (Winston, P. H., ed.), pp. 211–277. McGraw-Hill, New York.

O'Kane, M., Keene, P., Landy, D. & Atkins, S. (1989). Generalising from single-speaker recognition in a feature-based recogniser. In *Proceedings of Eurospeech-89*, Paris, France, **2**, pp. 409–412.

Robinson, T. & Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, **5**, 259–274.

Rothenberg, M. (1963). *Programmed Learning Problem Set, to teach the interpretation of a class of speech spectrograms.* Ann Arbor Publishers, Ann Arbor, Michigan.

Shortliffe, E. H. (1976). *Computer Based Medical Consultations: MYCIN.* American Elsevier, New York.

Stern, P.-E. (1986). *Un Système Expert en Lecture de Spectrogrammes*, Thèse de Docteur Ingénieur, Université de Paris 11, Centre D'Orsay, France.

Stern, P.-E., Eskénazi, M. & Memmi, D. (1986). An expert system for speech spectrogram reading. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing ICASSP-86*, Tokyo, Japan, pp. 1193–1196.

Tattegrain, H. & Caelen, J. (1989). Phonetic unit localization in a multi-expert recognition system. In *Proceedings of Eurospeech-89*, Paris, France, **1**, pp. 256–259.

Vaissiere, J. (1983). *Speech Recognition: A tutorial.* Course given in Cambridge, U.K., July, 1983.

Zadeh, L. A., Fu, K.-S., Tanaka, K. & Shimura, M. eds. (1975). *Fuzzy Sets and their Applications to Cognitive and Decision Processes.* Academic Press, New York.

Zue, V. W. (1981). *Acoustic-Phonetic Knowledge Representation: Implications from Spectrogram Reading Experiments.* Presented at the 1981 NATO Advanced Summer Institute on Automatic Speech Analysis and Recognition, Bonas, France.

Zue, V. W. & Lamel, L. F. (1986). An expert spectrogram reader: a knowledge-based approach to speech recognition. In *Proceedings of IEEE Conference on Acoustics, Speech, and Signal Processing, ICASSP-86*, Tokyo, Japan, pp. 23.2.1–4.