POSITIONS, PROBABILITIES, AND LEVELS OF CATEGORISATION

Mary E. Beckman[1] & Janet Pierrehumbert[2]
[1]Ohio State University and Macquarie Centre for Cognitive Sciences; [2]Northwestern University

ABSTRACT: In looking back on the successes of speech science and technology in the last quarter of the 20th century, three lessons stand out. The first is that prosody and other positional information cannot be separated from the specification of phonetic contrast. Children learn the sounds of their native languages in context, and machine systems for synthesis and recognition can be improved by taking position into account. The second is that frequency matters. Many recent studies suggest that humans (like modern speech recognition systems) use probabilities to interpret the speech signal, and there seems to be no level of phonological knowledge that can be encapsulated away from probabilistic models of speech processing. Finally, the old two-level models of discrete phonological categories feeding into continuous phonetic dimensions needs to be replaced by a more accurate understanding of how categorical behaviour emerges in the course of language acquisition. Just as a speech recognition system must be trained on a database that provides a broad enough coverage of phonemes and their contexts, the human infant must be exposed to enough lawful variability in the input in order for the native language categories to emerge.

BRIDGES

The theme of this eighth Conference on Speech Science and Technology is stated eloquently in the sub-title of Wednesday's plenary session on "The future of Australian speech research: Building bridges between speech science and technology". This is also the theme of the conference series as a whole. The SST conferences were intended from the beginning to build bridges between these two apparently different approaches to our common object of study. Each SST has been a forum where researchers from all sub-areas of speech science and speech technology can come together to present their own work and learn new methods, results, and insights from each other. Such "bridging" conferences are now a regular occurrence, of course, with EUROSPEECH and ICSLP alternating to provide an annual international venue for researchers in both speech science and speech technology. It is important to remember that this was not always the case. The first SST conference was held as recently as 1986. That was a year before the first EUROSPEECH and a full four years before the first ICSLP. Why is the history of such conferences so short?

One possible reason is that there was not such a clear division between the speech science and speech technology communities three decades ago. Before the 1970s, the commercially important questions were in the area of coding, not synthesis or recognition. The communications industry needed to know how narrow bandwidth could be and still provide intelligible transmission over a telephone line. Companies vied to find materials that could be used to increase bandwidth cheaply, and the proprietary secrets were not ones that involved modelling speech per se. As a consequence perhaps, there was no clear demarcation between commercial and academic speech research. The speech synthesis systems of the 1970s were rule-based programs that built directly on the acoustic theory of speech production (Fant, 1960) and on the Generative Phonology model of the relationship between speech production and phonological contrast (Jakobson, Fant, & Halle, 1952/1967; Chomsky & Halle, 1968). In these early days of development, such rule-based systems were not commercially viable, and software was exchanged freely among laboratories. The stimuli used in a large majority of the studies on speech perception published in the *Journal of the Acoustical Society of America* in the 1970s, for example, were specified in terms of their OVE IIId or Klatt synthesiser parameter values. By 1986, however, Klatt's final rule set for synthesising American English had long since been sold to a commercial entity and had became proprietary code.

Another possible reason for the recent origins of conferences such as the SST is that speech recognition is different from speech synthesis. Whereas the Generative Phonology model of speech production that dominated speech science in the 1970s and 1980s was easily implemented in rule-based synthesis systems such as MITalk (Allen, Hunnicutt & Klatt., 1987), the analogous approach to modelling speech perception was not as successful in speech recognition systems. When the HARPY system achieved much higher recognition rates than any of the more traditional features-and-rules-based approaches in a defence department competition in the mid 1970s, it was easy to conclude that speech science is not relevant for speech recognition technology. To bridge the

science-technology gap, then, it became necessary to build a bridge in the opposite direction, incorporating ideas from recognition into models of perceptions (see, e.g., Pisoni et al., 1985). We will elaborate on this point below, but first we would like to point out that the SST conference organizers in 1986 were among the earliest to recognize and to try to bridge the gap between these two approaches to speech perception/recognition research. We would also like to point out several other ways in which the theme of bridges is particularly salient at this SST.

This year, the conference returns to the city of Canberra, where the series began. Canberra, of course, is a capital city designed for the whole of Australia, deliberately situated as a bridge between the competing urban centres of Victoria and New South Wales. This is the first SST to be held after the Berlin meeting of the Acoustical Society of America, and that was the first ASA meeting to be held outside of North America. This is also the first SST to be held after the European Speech Communication Association changed its name to the International Speech Communication Association. (We note the foresight of the original SST organizers to include the word "International" in the full name of the conference series, well before the first *International* Conference on Spoken Language Processing was held in Kobe.) As far as we are aware, this SST is also the first international speech conference to have had a completely electronic paper submission and review. And it is the first SST to have an international representation so proportionally large, with nearly 40% of the papers by non-Australians. A reasonable conclusion from all of these firsts is that modern information technology is helping us bridge the tyranny of distances that once made for distinct speech research communities in Europe and North America, in Australia and Japan. Finally, we note that this year is also the bridge between the 20th century and the 21st. These circumstances invite us to look both backwards and forwards. We want to look back at the achievements of the first century of modern speech science, and in particular at the achievements of the last few decades, to try to pick out the most important lessons and themes to carry us forward into the new millennium.

THREE THEMES

When we do this, three themes stand out. The first is that position matters. A /t/ that is not before a vowel, as in the word "act" spoken in isolation, is not the same sound as a /t/ before a high front vowel, as in the last syllable of "Canberra, ACT". Moreover, it is not enough to specify the segmental context, since the /t/ in "city" also is different from the /t/ in "ACT" even though it also precedes a high front vowel, albeit an unstressed rather than a stressed one. The fact that position matters is, of course, something that was noticed by the pioneers of modern speech science in the first decades of the 20th century. A careful listener can hear positional effects such as the differences among the English /t/ sounds in "act", in "ACT", and in "city", and a well-trained listener in the early 1900s had adequate tools to record these differences more accurately than does English orthography. These tools were a standardised International Phonetic Alphabet and heuristics such as the distinction between a "narrow" and "broad transcription" record of an utterance. Of course, by the middle of the century, the development of the sound spectrograph and early attempts at concatenative speech synthesis made it is abundantly clear that these tools were not adequate. A discrete symbolic record cannot capture the full extent of positional variation that is produced by a fluent adult speaker. In response to this fact, the more successful automatic speech recognition systems use extremely detailed phonetic representations that are trained on large samples of speech. Many of the biggest debates in speech science in the last few decades have been triggered by the emerging realisation of the extent to which position matters, and of the need for better mathematical models of the principles governing the variation. And speech researchers are only now beginning to come to terms with the implications that this has for our most basic assumptions about the nature of language and about how babies go about learning a language.

The second theme that stands out is that frequency also matters. The /t/ in the English word "act" can be released in a careful production in isolation, but in fluent connected speech, it probably will not be clearly released. However, this need not pose a problem for the child learning English, because in any word-final stop cluster in English, the probability is also extremely high that the second stop will be an alveolar. The young child who hears the word "act" for the first time in a sentence such as "Let's act in the play together." and who correctly parses the silent interval as being too long for just a /k/ should be able to rely on knowledge of these probabilities to recognise and learn the new word correctly. Adopting probabilistic models was key to the advances in speech recognition technology in the 1970s, and probabilistic models are becoming increasingly central to speech synthesis technology as well. In this paper, we will describe some recent findings that support the idea that frequency

matters for humans as much as it does for machines. Also, since position matters and frequency matters, a related point will be that positional frequency matters.

A third theme that will be important for bridging the gap between speech science and technology is that level of representation matters. The fact that the second of two stops in an English word-final cluster is almost certainly an alveolar is a generalisation over the frequencies with which different consonants occur in different positions in all of the words of the language. The right level of representation here is the inventory of exactly three distinct categories of stop place in English. Is this a position where each of the three places of articulation is equally likely? Could the stop be a suffix, separated from the preceding /æk/ by a morpheme boundary? In contrast, the fact that the /t/ in "act" is probably not going to be released into the same clear sharp burst as the /t/ in "ACT" is a generalisation which requires two levels of representation. It is based on the frequencies with which different spectral patterns occur when realising the stop's place of articulation in tokens of the word "act". Thus one level is the phonemic/lexical level, or something like it, and the other is level at which we characterise the spectral pattern itself. Is there a sharp spectral discontinuity as pent-up air behind the alveolar closure is vented through the suddenly opened passage between the tongue blade and the alveolar ridge? If there is, what frequencies dominate in the burst spectrum? How likely are these frequencies to dominate, if the stop is an alveolar? In modelling how frequency is used in human speech processing, it is important to keep an open mind about what is the most appropriate level (or levels) of representation for each frequency-related phenomenon. How to determine the appropriate levels of representation for any given phonological phenomenon is one of the most hotly questions in speech science today. To see why, consider the relationship between a plausible theory of how knowledge of a word is represented in a speaker's mind and a viable model of how that speaker processes a particular instance of that word.

KNOWLEDGE AND PROCESSING

Linguists have known since early in the 20th century that a plausible theory of word knowledge must include at least two levels of representation to account for duality of patterning. In all human languages, there are patterns that have to do with how meanings are organised, and there are patterns that have to do with how sounds are organised independent of their meanings. More recently, linguists have begun to appreciate that each of these levels is itself more complex.

As an example of complexity on the meaning side, consider the distribution of verbs. Verbs are distinguished semantically by being words that name events. They are distinguished syntactically by how they are distributed relative to other word classes such as nouns. In English, for example, the probability is low that a verb such as "sleep" will occur before its subject noun. So "The baby's sleeping." is a commonly attested sentence of English, and "Colourless green ideas sleep furiously." is a possible sentence of English, but "Slept the baby tranquilly through the night." is not a grammatical sentence of English, as any native speaker can tell you. Part of learning a language means internalising general principles that describe the word classes and their well-formed combinations. Linguists observe how speakers of a particular language process sentences under imaginatively contrived circumstances to try to discover what principles the speakers have internalised about the language's word classes and syntactic structures. More generally, the goal of linguistic theory is discover what sets of such principles may constitute a possible human language.

Modelling the sound patterns of a language is no different. Several levels of representation are necessary to capture the speaker's knowledge of phonology. Speech categories (such as the phoneme /b/) must be characterised both by how they realised in the acoustic stream and by how they are distributed relative to each other. The infant who is acquiring English learns how to produce a labial stop, how to hear a labial stop, and where a labial stop can occur. The child eventually comes to know implicitly that /brɪk/ ("brick") is a word of English, and that /blɪk/ is a possible word, but that /bnɪk/ is not a possible word of the language. Learning a language means internalising general principles that describe the phonological entities and well-formed combinations of entities in that language. These general principles are revealed in well-formedness judgements and other productive behaviours that can be elicited by the speech scientist. Speech scientists observe how speakers of a particular language process speech in imaginatively contrived circumstances to try to discover what principles the speakers have internalised about the sound patterns of the language. More generally, the goal of a scientific theory of language sound structure is to discover sets of such principles may constitute a possible human phonology.

Note that this characterisation of the relationship between knowledge and behaviour is somewhat different from the approach traditionally taken in Generative Phonology. The traditional distinction between speakers' implicit knowledge and their processing skills -- i.e. between a theoretically interesting mental component of "competence" and the less interesting details of "mere performance" -- does not invite imaginatively controlled observational techniques. It also assumes a rigid differentiation between two levels of description -- a "phonological" level of language-specific categories and strings, and a "merely phonetic" level of universal mechanical processes for generating a continuous acoustic signal from the input string. By contrast, the claim here is that the relationship between knowledge and processing is symbiotic. Knowledge feeds on processing. Processing feeds on knowledge. They must be closely tied because one is the synoptic and the other the dynamic description of the same mental system. Babies acquire knowledge of the words of their native language by gradually becoming more and more adult-like in the ways that they process the acoustic patterns that they hear around them every day. They learn to pronounce the words of their native languages by paying attention to frequently recurring patterns and trying to reproduce them in their own vocal tracts.

The Generative Phonology model led to several synthesis-by-rule systems in the early 1980s that produced extremely intelligible segments. However, the model did not work as well for speech recognition. The advantage of recasting the traditional distinction between competence and performance as a distinction between the synoptic representation of a linguistic pattern in the long-term memory store and the dynamic representation of the pattern for use in immediate processing is that it gives us a viable account of human learning. A more accurate account of how humans learn to process sound and how this relates to what they know about the sound patterns of their language can only help in refining the ways that we train speech recognition systems. In this paper, we will review a small portion of the experimental literature from the last decade which supports our claim that knowledge and processing are closely linked, and that abstract, categorical knowledge of a language's phonology emerges only as a result of frequent and sufficiently varied experience with processing sound patterns at many different levels of representation. A good place to begin is by reviewing what infant studies in the last two decades have taught us about how babies begin to learn the words of their native language.

WHAT DO BABIES KNOW ABOUT WORDS?

One of the first problems that babies face in this task of learning the words of the ambient language is figuring out how to match up chunks of sound with chunks of meaning. In order to acquire word meanings, babies must be able to spot potential words. How do they do this? The evidence from acquisition studies suggests that infants are sensitive to the frequency with which different patterns occur, and pay attend to these frequencies to pick out memorable parts of the speech stream. The frequencies involve three types of pattern.

First, words in major grammatical classes often show regularities in their prosodic patterns. Patterns that are more frequent in one language may be less frequent in another language, but infants behave from a very early age as if they know the more frequent patterns. In French, for example, a content word in connected speech frequently will be demarcated by a longer "accented" final syllable (Wenk & Wioland, 1982; Fletcher, 1991), and Cristophe et al. (1994) show that French infants react differently to disyllabic sequences that contain or do not contain a word boundary (e.g., /mati/ extracted from "panaorama typique" versus /mati/ extracted from "mathematician"). In English running speech, on the other hand, a majority of the content words are likely to begin with stressed syllables (Cutler & Carter, 1987). The first sentence of this paragraph, for example, contains eight content words beginning with a stressed syllable containing a strong vowel and only two words -- "prosodic" and "grammatical" -- that begin with a syllable containing a reduced vowel.

Jusczyk, Houston, & Newsome (in press) show that by the age of 7.5 months, English-acquiring infants are sensitive to this regularity. When they are familiarised to words beginning with a strong syllable, such as "hamlet" and "kingdom", they listen longer to passages containing many occurrences of these words, and are not foiled by passages containing the embedded strings "ham" and "king". Moreover, the familiarisation works in both directions. After listening to passages containing many occurrences of the words "hamlet" and "kingdom", the infants listen longer to list presentations of many tokens of the words "hamlet" and "kingdom" than to list presentations of "ham" and "king". By contrast, when familiarised to words beginning with a weak syllable, such as "guitar" and "surprise", they listen just as long to passages containing the embedded monosyllables "tar" and "prize" as to

passages containing the target bisyllabic words. Moreover, if they are familiarised to passages containing "guitar" and "surprise", but these words are always followed in the passages by "is" and "in", respectively, they listen longer to list presentations of the nonsense words /tartz/ and /praIzIn/ than to lists of the real embedded words "tar" and "prize", as if parsing the entire trochaic sequence as a word. In other words, by 7.5 months, English-learning infants have learned to listen for recurring patterns that start at strong syllables, but they need to hear lawful variation in the following syllables to know where the pattern stops.

Second, phonemes and phoneme sequences often do not occur with equal frequency in all positions, and these phonotactic constraints are different for different languages, even when the languages have comparable phoneme inventories. In English, for example, /d/ occurs at the ends of many monosyllabic words that young children are likely to know ("food", "bad", etc.), whereas /kn/ does not occur in any English words in initial position and in only a few words in medial position (e.g., "acne"). In Dutch, by contrast, /kn/ is a fairly common word-initial cluster, occurring in words that are cognate with English "knee", "knife", etc., whereas /d/ does not occur word-finally except in clear speech "spelling pronunciations" of words such as "hemd" ('skirt'). Infants become sensitised to such phonotactic frequencies relatively early, although not as early as they become sensitised to the prosodic regularities (see, e.g., Friederici & Wessels, 1993; Jusczyk, Luce, & Charles-Luce, 1994).

Third, the varying acoustic values that can cue a particular phoneme occur with different frequencies in different positions. For example, the word-final /t/ in forms such as "cat", "act", or "night" often is not released with a strong burst, and voice onset time values frequently are rather short, even when a vowel or sonorant consonant follows, as in "Put the cat out." or "These are the night rates." By contrast, a stressed-syllable-initial /t/ in forms such as "tack" or "ACT" tends to show longer voice onset times. The difference in typical VOT values for the two positions is quite audible to the trained ear, leading to the traditional description of English /t/ as having categorically distinct unaspirated versus aspirated "allophones" in the two positions. When the following segment is an /r/, as in "train" or "nitrates", VOT values are longer still. They are frequently so long that traditional descriptions specify the /r/ also as having a categorically different "devoiced" allophone in this context. If such allophonic variation in the distribution of fine-grained phonetic values is associated with different word positions, as in "night rates" versus "nitrates", it should help native speakers (and machine recognition systems) segment the speech stream (see, e.g., Church, 1987a; 1987b).

Jusczyk, Hohne, & Bauman (1999) used the targets "night rates" and "nitrates" as foils for each other to show that by 10.5 months, English-learning infants are sensitised to such allophonic differences in the distribution of phonetic values across position. If familiarised to "night rates" in a repetitive list presentation of the phrase, they listen longer to passages containing that phrase than to passages containing the word "nitrates", and vice versa. Conversely, if they hear passages containing many occurrences of "night rates" or "nitrates", they listen longer to repetitive presentations of the target item than to the phonemically identical foil. This effect was not found with 9-month-old infants. Moreover, when the younger infants were familiarised to the word "night" in a list presentation, they did not differentiate between a passage containing the phrase "night rates" and a passage containing the word "nitrates", although they did listen longer to a passage containing the word "night" if it was followed by a variety of other words ("night cap", "night games") and not just by "rates". This is just like results for the "(gui)tar is"/"(sur)prise in" passages. The 9-month-old infant can extract the word "night" if it is presented in other contexts, and not just in the phrase "night rates". The older infant, by contrast, probably could use the allophonic differences to identify the word edge in "night rates" (although Jusczyk and colleagues have not yet completed the companion experiments to show this).

A striking fact about these infant studies is that they all describe the infant's phonological knowledge before or just at the beginning of the first-word stage, when a few recognisably meaningful word productions are just emerging from the infant's babbling, and a year or more before the child has enough words that minimal pairs must be differentiated precisely from each other (see review of stages in Bates & Goodman, 1999). Summarising these infant studies, then, we can say that even before they have any words *qua* words -- that is, even before they can associate any well-rehearsed patterns of vocal production with fixed set of meanings and syntactic functions -- infants are beginning to process the speech stream as if they implicitly "know" what is likely to be a word of the ambient language. Already by 7.5 months, English-learning infants are sensitive to the more frequent strong-weak stress pattern of bisyllabic words. By 9.5 months they are sensitive to differences in frequency for different phoneme sequences. And by 10.5 months they are sensitive to allophonic differences in

the distribution of VOT values and other release parameters that will help them use "juncture" cues to pick out word edges. Evidence from adult studies shows that an adult's explicit phonological knowledge also is sensitive to the same three types of distributional differences, to which we now turn.

FREQUENCY AND ACCEPTABILITY

Considering first the regularities in prosodic pattern, recall that the infant studies show that English-learning babies listen longer to lists of bisyllabic words that have the more frequent strong-weak pattern than to words that have a weak-strong pattern (e.g., Jusczyk, Cutler, & Redanz, 1993). The adult studies show that the more frequent pattern also makes a new word sound like a "better" word for adults. Vitevitch et al. (1997) asked native speakers of English to rate bisyllabic nonsense words on a 10-point scale of acceptability from "GOOD ENGLISH WORD" to "BAD ENGLISH WORD", and found consistently better mean acceptability ratings for forms with first syllable stress. Adults were also significantly faster at repeating the nonsense words that had initial stress. The grammaticality judgements and the processing effect both reflected the prosodic pattern frequency in the same way.

Several of our own studies of English consonant clusters further suggest that adult grammars are quite sensitive to frequency-based generalisations about what is a likely or an unlikely sequence. For example, Pierrehumbert (1994a) used a large on-line dictionary to tabulate all of the attested word-medial English clusters that contain at least three consonants. She then compared the attested clusters to the predictions of two models: a traditional context-free Generative Phonology, as in Figure 1, and a stochastic grammar that modelled acceptability in terms of frequency. (In this tabulation, "attested" cluster were ones that occurred morpheme-internally in two or more reasonably familiar words, where "reasonably familiar" excluded words such as "anschluss" and "pozzuolana", but not "pancreas", "extirpate", and "palfrey", and "morpheme-internally" excluded words such as "softbodied", but not words such as "complex" and "obtuse".)

PrWd --> Syl$^+$	(a prosodic word consists of 1 or more syllables)
Syl --> Onset Rhyme	(a syllable consists of an onset and a rhyme)
Onset --> {C, CC, CCC}	(an onset consists of 1-3 consonants)
Rhyme --> V (C (C))	(a rhyme consists of a vowel and optional coda of 1-2 consonants)

Figure 1. A context-free grammar for generating English words.

The rules for generating possible words in Figure 1 predict that the acceptable medial clusters in English should be the cross-product of the acceptable final clusters in rhymes and the acceptable initial clusters in onsets. Thus, given the existence of /l/ as a coda consonant in words such as "fall" and "eel" and the existence of /skr/ as an onset cluster in words such as "scrabble" and "screed", we should find words containing the medial cluster /lskr/. A frequency-based model of the grammar, by contrast, predicts that /lskr/ will be found only if the frequencies of final /l/ and initial /skr/ are high enough to make the joint probability large relative to the number of words in the language.