ARGUMENTS FOR A NONSEGMENTAL VIEW OF SPEECH PERCEPTION

Sarah Hawkins, Department of Linguistics, University of Cambridge, U.K.

ABSTRACT

Systematic acoustic variation reflects vocal tract dynamics; it provides the acoustic coherence that makes a signal sound like speech. It is thus basic to speech perception, defined as lexical access. Implications of this argument are that the perceptual system maps an informationally-rich signal directly onto lexical forms that are structurally rich, and that phonemes are not essential for lexical access. Some properties of such a view of speech perception are discussed.

1 INTRODUCTION

In this paper, I argue that speech perception takes place by reference to a mainly nonsegmental phonetic structure. I discuss first some obvious shortcomings of the standard view of phonetic structure, in which prosody and a linear sequence of phonemes form two largely separate strands. Next, I argue that models of phonetic structure and of perception should include detailed information about the dynamics of vocal tract behaviour, since these details contribute coherence and systematic information to the signal. Finally, I outline the main properties I think a nonsegmental phonetic model of speech perception should have.

2 THE STANDARD VIEW OF PHONETIC STRUCTURE

The standard view of phonetic structure is of a linear sequence of so-called segments superimposed on a rather independent prosodic strand. Most people acknowledge that this view is a vast oversimplification, but nevertheless it underlies almost all the most influential phonetic models of speech production and perception. Explanations of the relationship between this abstract picture and reality are vague. Relationships between segments and prosody are poorly understood and not well studied: the two tend to be analysed separately, even though we know they are really not separable. Timing, for example, contributes crucially to both segmental identity and prosody. And formal relationships between these phonetic and phonological constructs and the other constructs of linguistics, such as grammar, are almost nonexistent.

Segments, for most people, seem to be closely tied to phonemes, even though, as I understand it, the term segment is typically used precisely to avoid the term phone or phoneme. At its least theoretical, a segment is an ‘acoustic segment’, i.e. that part of the acoustic signal that corresponds most closely to the ‘main properties’ for a particular phoneme.

Segments that are easiest to identify have abrupt acoustic boundaries. Many correspond to changes in excitation source, and/or to spectral steady states. They are usually clearly visible in spectrograms: turbulence noise of fricatives, silence associated with oral stops (often with a noise transient), periodicity of sonorants, the steady states (and sometimes transitions) of vowels, nasals, prevocalic /l/, and so on.

Everyone acknowledges that these acoustic segments are not phonemes, or even phones. But there seems to be a willingness to let the relationship between the two remain murky, partly perhaps because the linear model is so neat, and, in these clearcut cases, there is a strong connection between acoustic segments and phonemic identity. However, the term ‘segment’ is extended to other sounds as well: /w/, /j/, various types of /r/, and postvocalic /l/ are also segments, and we say that they are hard to segment ‘because their boundaries are only arbitrarily distinguishable from those of neighbouring segments’.

Descriptions of coarticulation are also often vague about how it arises, although coarticulation is integral to recent models e.g. Articulatory Phonology, and [1].

Intonation has tended to be seen as having the opposite problem. The challenge has been not so much to find the acoustic correlates of a predefined set of discrete units in the more continuous, measurable F0 contour, as to establish what the discrete units should be.

In reality, acoustic correlates of linguistic units are typically complex, spread over relatively long sections of the signal, simultaneously contribute to more than one linguistic unit, and do not cluster into discrete bundles.

3 COHERENCE & SYSTEMATIC VARIATION IN SPEECH

Models of human speech perception have typically incorporated linguistic-phonetic constructs fairly uncritically. Thus they assume that the main challenge is to map the acoustic signal onto the discrete sequence of segments that correspond to phonemes. Intermediate stages such as distinctive features may or may not be included, and prosody has usually been neglected. The nondiscrete nature of the signal has been ignored or seen as a problem of noise (but cf. [2]). If we were to take the opposite approach, and use what we have learned about speech and speech perception to help define the properties that a model of phonetic structure should have, we might come up with something rather different.

3.1 Natural speech

When humans speak, there is a tight relationship between the behaviour of the vocal tract and the acoustic properties of the emitted sounds. Thus natural speech is acoustically coherent: it contains all sorts of acoustic-phonetic fine detail that reflects vocal tract behaviour. This fine detail, and the consequent acoustic coherence, is found in all aspects of speech. For example, it is found in correlations between the mode of glottal excitation and the behaviour of the upper articulators, especially at abrupt segment boundaries; in the amplitude envelope governing, for instance, perceptions of rhythm and of ‘integration’ between stop bursts and following vowels, and in coarticulatory effects on formant frequencies. (Other modalities, such as vision, can also contribute coherence, as I discuss briefly below. For simplicity, I restrict the present discussion to the acoustic signal.) All these types of effect contribute to acoustic variability. But the variation they contribute is systematic, or lawful variation, and adds information rather than noise to the signal.

We could say that systematic variation will only be called variation if we are bound to a view of speech as a linear sequence of phonemes that have a canonical, or pure, form, with clearcut temporal boundaries, and in which phonemes, excitation source, and prosodic variables are thought of as independent. These conceptually distinct strands are not separable in reality, and although there are reasons within linguistic theory to analyse them separately, maintaining a rigid separation may unnecessarily distort our thinking about speech perception and synthesis.

There is evidence from at least three fields of enquiry that coherence (or naturalness) of the speech signal is crucial to speech perception: from auditory psychophysics of the way the auditory system organises sounds into patterns, from speech synthesis by rule, and from speech perception itself.

3.2 Auditory psychophysics

Experiments show that when sounds have certain temporal and spectral relationships to one another, humans group them so that they form coherent patterns, such as alternating single notes and chords, or particular rhythms of a single tone. This phenomenon is called auditory streaming [4]. An auditory stream is perceived as coming from a single source. To use an older term from psycholinguistic studies, the sounds form a Gestalt. To cohere as an auditory stream, the sounds must be somewhat similar in frequency and timbre: a rhythmically-structured series of tones that differ greatly in pitch, say, is more likely to be heard as two independent streams. Changes in frequency or temporal relationships can drastically change the percept. Depending on the change made, the sounds may be heard as another pattern in the same stream, or they may break into a different number of streams, each with its own pattern, or as a chaos of unrelated events. In short, whether or not a time-varying signal (like speech) is heard as coming from a unitary source depends on tight spectral and temporal relationships between its various events. An example in speech is that we use the continuity of F0 to distinguish simultaneous vowels from each other [5].

3.3 Speech synthesis by rule

Those who work with synthetic speech have experienced the sense of incoherence due to inappropriate changes
likely to be heard as speech. Sounds that cannot come from a vocal tract but can nevertheless be interpreted, like sine wave speech [9], seem to be understood because the acoustic fundamental properties of speech, and hence of vocal tract movement: achieving the right timing, frequency and amplitude relationships is crucial. No one claims that a sine wave sounds natural, nor that it is easy to understand: these requirements demand that fine acoustic detail is added. And this detail follows the systematic variation caused by the way the vocal tract works.

Theories based on the acoustic signal incorporate vocal tract dynamics at least implicitly to the extent that they refer to time-varying properties. The theory of acoustic invariance, for example, stresses effects of the changing shape of the vocal tract that are reflected in constancy of relationships in frequency or amplitude across acoustic boundaries [10]. Postulated invariant properties transcend systematic variation in the signal, yet include it because the variation is part of each measure of invariance. The variation can be responded to as information rather than noise if we assume that the perceptual process continuously assigns probabilities rather than binary values to features, as I suggest below.

Experiments from other theoretical approaches also support the importance of vocal tract dynamics. The large literature on the importance of consonant transitions to phoneme identification is a prime example, while others show that the more subtle systematic variation also contributes to perceptual decisions. Some of these are mentioned below.

4.3 What constitutes perceptual information?

I make three assumptions about what constitutes perceptual information: that all speech-relevant information is potentially salient; that sensory input is interpreted in relational terms, and that the signal varies in the amount of information carried per unit time. The assumption that all speech-relevant information is potentially salient to the perceptual mechanism does not entail the claim that it is always all used. Whether it is used, and the extent to which it is used, depends on its quality and on what other information is available. Evidence supporting this view comes in many forms, including acoustic cue trading, the many demonstrations of the influence on sound or word identification of higher-order linguistic factors such as vocabulary size, predictability from context, and lexicality, and cross-modal influences on speech perception e.g. [13]. Of these, the last is perhaps most worth discussing.

In [13], /baba/ and /gaga/ were cross-matched such that the listeners heard /baba/ synchronised with a video of a mouth saying /gaga/, or vice versa. Responses were asymmetrical: the visual stimulus has a profound influence when the heard stops are bilabial, but when they are velar, the visual influence (of /baba/) is smaller and less consistent. The explanation rests in what the listener knows about the relative quality of each sensory channel. Acoustically, velar stops are fairly distinctive, whereas bilabials are not [14,15] and can easily be misheard. Clear sight of a closure being made inside the mouth can apparently cause the weak and hence potentially unreliable acoustic properties of a bilabial stop to be disregarded in favour of the more reliable visual information: when /baba/ is heard but /gaga/ is seen, the visual input dominates. When information from both channels is clear and hence reliable, the perceptual system gives weight to both, and produces combination g-b responses.

Evidence for the second assumption, that acoustic and visual information is interpreted in relational terms, is also widespread. Auditory streaming attests to its importance. Timing, by its nature, involves relational properties, as do aspects of perceived phone identity such as stop voicing and schwa identity. That relational properties are fundamental suggests that normally, sounds or features can only be interpreted in context. While that is a relatively new idea in acoustic studies of speech perception, it is not new in linguistic theory: relational properties underpin the entire phonetic and phonological structure. When the salient sensory cues are also expressed in relative terms, we have a consistent contrastive structure from sensory input up to the lexicon.

The third assumption, that the signal varies in the amount of information conveyed per unit time, requires no justification, but it does have important consequences for our thinking about perception. Let us first consider regions of the signal that are rich in information. These are sometimes called islands of reliability. They feature (with different names) in a number of theoretical approaches, including invariance theory [10], Temporal theory [16] and robust features [17]. While work on acoustic invariance has tended to emphasize dynamic properties, robust features are typically characterised in terms of properties that are constant for the duration of at least the major portion of a phone-sized acoustic segment. It is not clear that we need to choose between these approaches. While acoustic invariance seeks short-term properties that are minimally sufficient to provide evidence for a particular feature, each
4.4 Is the phoneme necessary for such weak acoustic cues to identify speech?

Phonemes in natural and synthetic speech adjacent vowels and even to nonadjacent consonants can spread throughout conditions, natural variation in formant low information can contribute valuable more inference about the message than others e.g. some phones are inherently partly to fashions of inquiry. Some regions of the signal that are not rich in articulation [14,18], and when long is contribution to the percept of place of stop function noise, which, when short, can contribute to the original interpretation (see [23]).

Intermediate between the acoustic signal and the lexicon. An intermediate classification seems only worthwhile if it reduces processing load. It must be reasonably error-free, and allow information to be thrown away. But acoustic information seems to be held until quite late in the identification sequence. For example, listeners can back-track to reinterpret acoustic information quite a long time after a misperception, reconstructing an entire phrase and seeming to 'hear' that the reconstruction is more satisfactory than the original interpretation (see [23]).

4.5 Outline of a nonsegmental model of speech perception

A model that assumes feature information such as sight of the speaker's face, if available. I have also argued that all input provides potentially valuable information, that its quality is evaluated during the process of making perceptual decisions, and that relational (context-dependent) properties are fundamental. These arguments lead me to question whether a phonemic stage is necessary to lexical access. Why not map more detailed properties of the signal directly onto words? This proposal is not original (cf [21,22]), but a reason for making it are worth examining, in addition to those made by e.g. [18] that acoustic cues are not always straightforwardly combined into phonetic features, nor features into phonemes.

An obligatory phonemic stage must be intermediate between the acoustic signal and the lexicon. An intermediate classification seems only worthwhile if it reduces processing load. It must be reasonably error-free, and allow information to be thrown away. But acoustic information seems to be held until quite late in the identification sequence. For example, listeners can back-track to reinterpret acoustic information quite a long time after a misperception, reconstructing an entire phrase and seeming to 'hear' that the reconstruction is more satisfactory than the original interpretation (see [23]).

Another argument is that some phonemic sequences map uniquely onto words only after the acoustic offset of some candidates e.g. 'plum' vs 'plumber' in I saw the plum on the tree [24]. The listener seems able to keep both lexical options available [25], but it seems risky to keep only phoneme strings without detailed sensory information, and contrary to evidence of late integration of different sources of information e.g. [26]. A less commonly made argument comes from language acquisition. Children seem to learn to talk by imitating the sound pattern of what they hear, without a complete phonological (or syntactic) analysis [27,28]. If that is the case, then presumably they operate without a fully systematic phonemic inventory, and if children start by doing that, it is difficult to see that they should be obliged to change as they get older. I suggest that it is possible but not obligatory to interpret the signal in terms of phonemes before lexical access. Normally, the phonemic interpretation will come after lexical access, perhaps to the extent that the person is literate.

Modelling phonemes as only an optional route to solve the conundrum in which allophonic information is crucial to feature identity and word segmentation but must be ignored in order to assign phoneme status. In a model in which phonemes are not central, we preserve the perceptual cueing value of variation due to phonetic context and connect speech processes by relating the input directly to phonological structure. Thus we preserve the information about syllable-dependent variation in the spectral and temporal properties of phones that is crucial to lexical access. This can have interesting phonological implications. Take the patterns of clear vs dark vs vocalized /l/ found in several varieties of English. Thus /l/ is [l] in standard Southern British English, but [l] is rapidly gaining ground as a stylistic option for some speakers, and is the only option for others. In standard phonological theory, all these accents are said to have a phoneme /l/ which can fall either before or after a syllabic nucleus. But are these 'l's the same for speakers who have only the vocalized version syllable-? For such speakers, the vocalized version is subject to linking phenomena which cannot occur in the dark /l/ version: consider legal fees, [lɪɡ̃fɪz], but legal aid, [lɪɡəd]. This suggests to me that, in these contexts, vocalized and word-initial /l/ are phonologically distinct. An important consequence of distinguishing syllable position of phones or features is that syllabic constituency is not only signalled, but preserved throughout the interpretation. Correct assignment of syllabic constituency seems basic to correct word segmentation, but this information is lost in a phonemic string unless phonotactic constraints are violated.

4.6 Outline of a nonsegmental model of speech perception

The model I suggest follows that proposed by [23]. Here, I develop some nonsegmental aspects of the model, for a bottom-up channel from the sensory signal to words. The role of higher-order knowledge and the model's interactive aspects are neglected here due to space constraints; they are discussed in [23].

One consequence of seeking a model that is closely tied to vocal tract dynamics is that it will use a rich acoustic structure with many redundancies. Another is that time must be explicitly represented, with both rapid events and slower time-changing information contributing to perceptual decisions. The model assumes that all speech-relevant information is used, weighted according to its apparent value. The sensory input is interpreted in terms of linguistically-relevant units, and lexical items are represented as complex structures involving those same units.

These lexical structures comprise syllables and their constituents, together with information that maps onto higher-order structures of prosodic and grammatical trees. Thus intonation and rhythm guide decisions and focus attention onto stressed syllables [29]. Lexical structures include some set of features as terminal elements. These features are unconventional; they take probability rather than binary values, and are distributed across time rather than bundled into units with discrete boundaries. Probabilities attached to features can change within as well as between syllabic constituents. Thus weak cues from small coarticulatory effects are represented. For example, the probability of a feature [high] is significantly greater than zero in the nucleus of the syllable before a high vowel, but it is higher still (normally 1) in the nucleus of the syllable containing that high vowel.

A model that assumes feature probabilities but neglects the weak cueing function of coarticulatory effects can assume that the lexical representation is in terms of resting levels, thresholds, and supra-threshold activation; any input value greater than the threshold activates the feature. But to accommodate coarticulatory cues, it seems necessary to limit the range of expected probabilities for each feature. When the effect of interest is weak (e.g. slight vowel raising due to coarticulation with a high vowel in the next syllable), the lower and upper limits of the range will be less than 1 for the relevant feature, here [high]. When the effect of interest is the primary
property of that part of the signal (e.g. a high vowel), then the upper limit will be 1, and the lower limit will be determined by contextual influences from other parts of the utterance. There must also be knowledge of relative probability of features across acoustic segments [23].

The input signal is represented as a set of prelexical features (or possibly loose clusters of features) whose values are also represented as probabilities. The signal is continuously monitored for information on each unit, giving rise to continuous modulation of probability levels of pre-lexical features, which in turn affects activation level of lexical items. Thus the model tracks time functions and hence vocal-tract dynamics (and their acoustic consequences), rather than only event sequences.

Lexical access involves taking the best match between input and stored probabilities for features. Unambiguous stimulus input is given great weight, and the system is based on choosing the most central properties of the system, rather than only event sequences. Thus the model tracks time functions and hence vocal-tract dynamics (and their acoustic consequences), rather than only event sequences.

Lexical access involves taking the best match between input and stored probabilities for features. Unambiguous stimulus input is given great weight, and the system is based on choosing the most central properties of the system, rather than only event sequences. Thus the model tracks time functions and hence vocal-tract dynamics (and their acoustic consequences), rather than only event sequences.

Lexical access involves taking the best match between input and stored probabilities for features. Unambiguous stimulus input is given great weight, and the system is based on choosing the most central properties of the system, rather than only event sequences. Thus the model tracks time functions and hence vocal-tract dynamics (and their acoustic consequences), rather than only event sequences.

Lexical access involves taking the best match between input and stored probabilities for features. Unambiguous stimulus input is given great weight, and the system is based on choosing the most central properties of the system, rather than only event sequences. Thus the model tracks time functions and hence vocal-tract dynamics (and their acoustic consequences), rather than only event sequences.