PHONEME-LIKE UNITS AND SPEECH PERCEPTION

Terrance M. Nearey
Department of Linguistics, University of Alberta, Edmonton, AB, Canada

ABSTRACT
In English, the perception of syllables and words can be largely predicted from the perception of ‘smaller’ phoneme-like units. Experiments reviewed by Allen [1] show that the correct identification of nonsense CVC syllables in noise can be extremely well predicted from the marginal correct identification rates of their constituent phonemes. Simulations by Nearey [2] suggest that this result can be readily achieved when syllable patterns can be ‘factored’ into phoneme parts, while representations with even moderately idiosyncratic relationships between stimuli and syllables cannot reproduce such results. Parametric experiments with synthetic speech, wherein listeners hear syllables that span two or more categories of two or more segments (e.g., bad, bed, bat, bet) also provide evidence for phoneme-factorability [3, 4, 5]. This paper summarizes how models with phoneme-like units as the core phonetic-transducers supplemented with stimulus-independent bias terms that can provide a good account of key aspects of listeners’ classification of words and nonsense syllables alike. However, substantial work remains to be done to bridge the gap between static pattern recognition models used in phonetics and current dynamic lexical access approaches, such as MERGE [6]. Some of the issues involved are outlined.

1. INTRODUCTION
For the following discussion, the term ‘basic phonological unit of speech perception’ will be taken to mean the mechanism that performs all or at least most of the work of initial transduction of stimulus properties (or their psychophysical reflexes) to discrete symbolic elements. The ‘size’ and generality of such basic units are still unsettled issues in speech perception. Classical distinctive feature theory [9] is highly compositional, positing small reusable elements that map onto very general stimulus properties. At the other end of the spectrum, Goldinger’s [10] episodic lexicon opts for large, idiosyncratic units; namely, collections of traces of exemplars of specific words of individual speakers.

Two streams of evidence reviewed below support a specific kind of compositionality I will refer to as the factorability of larger symbolic elements, such as words or syllables into phoneme-like units. The first stream involves speech reception, the intelligibility of natural speech in noise. Simulation of such experiments by Nearey [2] confirm that this pattern works over nonsense) can be achieved when stimulus properties associated with syllables are strictly factorable into phoneme parts. However, slightly more idiosyncratic representations of syllables lead to very different results.

The second stream of research stems from experiments with synthetic speech that study the simultaneous perception of consonants and vowels. Results here also demonstrate that listeners’ classification of syllables and words can also be largely reduced to the perception of their phoneme parts through the use of simple pattern recognition models [4,5].

After reviewing this evidence, much of which is elaborated in [2,3], the relationship of these results to lexical access is discussed. It is clear that the spirit of the hypothesis of phoneme factorability is compatible with a bottom-up framework like MERGE [6]. However, much work must be done before a concrete link between the two streams of research can be achieved [7].

2. SPEECH IN NOISE
2.1. Speech Reception Experiments
Allen [1] provides a concise account of the early work on speech intelligibility led by Fletcher at Bell Labs and of subsequent developments. Under varying noise conditions, intelligibility of spoken words and texts can be well predicted from listeners’ identification of phonetically balanced nonsense CVCs. Correct identification of such syllables can be accurately predicted from constituent phoneme scores as follows:

\[ P_s = P_{C_1} P_{C_2} \]

where \( P_s \) is the probability of correct identification of an entire CVC syllable, and \( P_{C_1} \), \( P_{C_2} \) are the marginal probabilities of correct identification of the constituent initial, medial and final phonemes. In many such experiments the probabilities for the three phonemes slots happen to be approximately equal (\( P_{C_1} = P_{C_2} = P_v \)), so that (1) is approximately \( P_s = (P_p)^3 \), where \( P_p \) is the average phoneme identification rate, \( (P_{C_1} + P_{C_2} + P_v) / 3 \). Boothroyd and Nittroeur [8] introduced a useful generalization of Fletcher’s formula:

\[ P_s = (P_p)^j \]

where the value \( j \) is called the \( j \)-factor. Consistent with Fletcher’s work [8] also find that for nonsense words, \( j \) (estimated simply as the mean of \( \log P_s / \log P_p \)) typically very close to 3.0. However, for real words, \( j \) is only about 2.5. (Data in [8] show \( M = 3.05, SD = .14 \) for nonsense and \( M = 2.43, SD = .07 \) for words across noise conditions. In the speech reception literature [1, 8], \( j \) is interpreted as an estimate of the number of informationally independent units that compose a syllable. If \( j \) is smaller than 3 for CVCs, then the three units are not fully independent, and syllables
are being recognized better than predicted by the recognition rates of their constituent phonemes.

Allen [1] argues that the value \( j = 3.0 \) in nonsense CVCs suggests that phones are perceived independently. The modest deflation of the \( j \)-factor for real words results from lexical redundancy or 'context entropy', reflecting statistical properties of symbol distributions in the lexicon, independent of stimulus properties. Allen claims these results (and others involving independence of information in different frequency bands) argue against the use of larger basic acoustic units, such as word templates in automatic speech recognition. Nearey [2] conducted a series of simulations with syllable patterns with more idiosyncratic stimulus properties (that cannot be reduced to phonemic parts), to see if the number of independent informational units as measured by the \( j \)-factor might be substantially smaller than 3.0 for all CVCs.

2.2. Simulations

2.2.1. Phoneme Factorable Syllables

In [2], 'pseudosynthetic' cue patterns were constructed for each of 1000 CVC syllables, with 10 possible segments in each phoneme slot (the same numbers as in [8]). Eight cues were manipulated. Four (F1 and F2 vowel targets; initial and final consonant voicing duration) were absolute invariants, with fixed values for each phoneme. The rest (initial and final 'dominant frequency' [spectral center of gravity of burst for stops, F2 steady-state for sonorant consonants], and F2 transition onset for initials and F2 offset for finals) exhibited 'embryonic coarticulation' through relational invariants, that with locus-equation style constraints between vocalic and consonantal cues. More explicitly, although each syllable was represented as a point in the 8-dimensional cue space, those points were constrained to lie in an orderly pattern consistent with simple constraints expressed at the phoneme level.

A single Bayesian classifier was used in all simulations. It was provided with sufficient statistics of Gaussian distributions for the cue patterns used to generate the stimuli (i.e., distinct mean vectors for each of the 1000 syllables and the variances of a common diagonal covariance matrix due to the added noise, discussed below). It is noteworthy that overall correct syllable identification rate (\( P_n \)) of this classifier is unchanged when individual consonant and vowel labels are randomly scrambled among the syllables. However, under such random perturbations, average phoneme identification rates (\( P_n \)) are reduced to 1/10, and calculated \( j \)-factors can be shown to approach 1.0 [2].

Varying signal-to-noise ratios were simulated by adding random normal deviates (sampled afresh on each virtual trial and with higher variences in noisier conditions) to the mean (prototype) cue pattern for each syllable on each virtual trial. A random 13% of the CVCs were designated as real words (roughly the ratio of words to nonsense in [8]). Simulations showed that it was easy to obtain \( j \) near 3.0 for nonsense syllables if and only if the means of the syllabic templates were constrained to be 'phoneme-factorable' by conforming exactly to the (absolute and relational) invariants of the type described above. It was also easy to simulate the slightly reduced \( j \) factor of 2.5 for real words by means of a single bias factor (after Broadbent [13]) that inflated the estimated likelihood for each syllable that represented a simulated real word by a ratio of 4.5:1 compared to simulated non-words.

2.2.2. Non-factorable syllables

\( J \)-factors similar to those of the real speech in [8] could be obtained when the underlying syllable-level patterns comply with certain phoneme-factorable patterns. But could other, relatively non-factorable, patterns produce similar results? If syllabic patterns are randomly related to their phoneme spellings, they will exhibit \( j \)-factors near 1.0. But what happens if there is a less radical dissociation? Suppose syllabic patterns are roughly, but not perfectly, compositional: a conventional phonemic description captures a coarse family resemblance among syllables, but there is a residue of unique, arbitrary stimulus properties tied to each syllable. Additional simulations in [2] show that even a very small residue of this kind results in \( j \)-factors much smaller than 3.0. Syllable cue-patterns, which were largely, but not entirely functions of their phoneme parts were simulated by starting with the pure phoneme-factorable patterns of section 2.2.1. These were then modified by adding a fixed randomly chosen value to the value of each cue for each syllable template. Given these fixed (but no longer factorable) templates, channel noise was again simulated by adding Gaussian deviates to each template mean on each virtual trial.

These simulations showed that even a 'soupçon of syllabic je ne sais quoi' is sufficient to cause a large deflation of the \( j \)-factor for both simulated words and non-words. If factorable syllabic templates are perturbed by Gaussian distributions with a standard deviation of about 5% of the original ranges of the cues, the means for such mildly 'syllable contaminated' templates are still very strongly correlated \( (r > .97) \) with those of the corresponding seed templates of the factorable model. In such a case, simulated nonsense syllables show \( j \)-factors on average less than the 2.5 value for real words found in [8] while those for real words are reduced \( j \)-factor of approximately 2.0.

Additional simulations were run with a larger fraction of irreducible syllabic effects. The means of the "20% syllable contaminated" models were still fairly strongly correlated with those of their factorable seed templates \( (r > .78) \). In this case the \( j \)-factors are severely deflated, averaging roughly 1.5 for both words and non-words.

3. FACTORABILITY IN SYNTHETIC SPEECH

Mermelstein [11] conducted an experiment with synthetic /BVC/ spanning the English words bad, bed, bat, and bet by varying F1 frequency and the duration of
the vocoid. Mermelstein’s analysis indicated that both cues affected perception of both the vowel and the final consonant. However, he also concluded that vowel and consonant choices were made independently of each other. Formally, \( P_{VC} = P_v P_C \).

In an analysis of a similar but more focused experiment, Whalen [12] claims that Mermelstein’s hypothesis cannot fully cover the newer results. Nearey [5] confirms this but his reanalysis of Whalen’s data using polytomous logistic regression, shows that a slight modification to Mermelstein’s model accommodates Whalen’s new data. Nearey’s revised model starts with bottom-up evaluation of stimulus properties by phoneme-sized elements, but then adds (stimulus-independent) biases favoring the specific VC combinations at and ed. These rhyme biases serve a role identical to that of frequency biases in Broadbent [13] or of Massaro’s contextual features [14].

Logistic regression is a powerful tool in assessing factoriality in parametric synthesis experiments and it has been successfully applied to a number of other cases, discussed in detail [3,4,5]. Nearey [3] has also shown how simpler graphical and numerical methods of j-factor analysis from [8] can be applied in such cases. For example, Whalen’s bad-bet experiment can be approached through equation (1) as follows. Because of the nature of the stimuli, the first consonant is always the same, so \( P_{C1} = 1.0 \) for \( C_1 = /bl/ \). In speech reception experiments (with natural speech), probabilities are usually calculated only with respect to the ‘correct’ syllable and are aggregated over all words (and again over all non-words) in a single noise condition. For parametric experiments, one can calculate the relevant quantities for each syllabic choice for each individual stimulus. In this case, response probabilities for \( C_1, V \) and \( C_2 \) are often very unequal, so it is appropriate to redefine such an ‘average’ phoneme probability, \( P_p \), as the geometric (rather than arithmetic) mean of the probabilities of response to the constituent phoneme. This ensures that when (1) is true, \( P_r = (P_{p'})^3 \) where \( P_{p'} = \left( P_{C1} / P_{C1} \right)^{\frac{1}{3}} \), for arbitrary phoneme probabilities. (For numerical reasons [8], only syllables with response probabilities between .05 and .95 are used below).

Applying these methods Whalen’s data yields estimates of \( j \) of 3.19 for bad, 2.83 for bat, 2.86 for bed, and 3.16 for bet (SDs ranged from 0.13 to 0.20). A j-factor of less than 3.0 indicates that responses to that word are greater than predicted by phoneme independence. Graphic analysis of j-factor results (using methods adapted from [8], see [3]) shows patterns quite similar to that observed for real versus nonsense syllables of section 2.2.1 above.

Other experiments also show j-factor values close to the theoretical independence value. Nearey’s [4] d/bVC/ data show j-factors ranging from 2.78 to 3.05 over 10 response syllables (SDs range from 0.24 to 0.45). Massaro and Cohen [14] ran an experiment where F2 and F3 patterns spanned /bla/, /dra/, /bla/, *dla/. Although the analysis presented by the authors suggests strongly the need for a substantial bias term (via their contextual feature mechanism) against the illegal */dla/. Nearey [3] shows j-factor plots of this data showing values that cluster tightly about the pure independence j = 3 curve. Numerical calculation yields j-factors of 2.94 to 3.05 over the four syllables (SDs ranging from 0.14 to 0.15).

Whalen [12] presents an experiment manipulating a fricative noise and F2 of a steady state vocoid, spanning the words sue, see, shoe, and see. He demonstrates that these data also fail to meet Mermelstein’s strict independence condition. Nearey [5] shows that the data yield to the same kind of diphone biased logistic as in the bad-bet case, but the biases in question are larger. This is confirmed by a simple j-factor analysis. Since the words in question here are two segments long, the target j-factor for independence is 2.0. The syllables are nearly a full point larger for the syllables shoe (j = 2.50) and see (2.55) than for sue (1.55) and she (1.57; SDs range from .14 to .31 over the four syllables).

Finally, data from the experiment of Repp and colleagues [15] shows j-factor patterns that are wildly different from any discussed so far. This experiment involves a choice set over the phrases gray ship, gray chip, great ship and great chip by manipulating silence following the first vowel and duration of the fricative noise of the sibilant. If a ‘zero’ coded for the final pseudo consonant of gray, the data can be scored as if there were two phonemes in two positions (\{t\} versus zero in the first position and alveopalatal affricate versus alveopalatal fricative in the second). Here, ignoring the fixed segments gray _ip, and treating 2.0 as the target value for independence, we find j-factors ranging from 1.4 to 2.8 a spread of 1.4 ‘effective information units’, with standard deviations ranging from .28 to .63. Graphic presentation of the data shows the situation to be chaotic and very unlike any other data set examined.

4. FACTORABILITY AND LEXICAL ACCESS

Nearey [3,4,5] has shown that all the examples considered above, except that from [15], are demonstrably compatible with categorization models where stimulus properties interact only with phoneme-sized elements, and where all additional fine-tuning is provided by stimulus-independent biases. Nearey’s models are closely related to many of Massaro’s FLMP. Massaro and Cohen [1995] have also shown that some lexical effects can be accounted for by similar bias mechanisms (see also [13]). These findings appear to be very compatible with bottom-up lexical access models like MERGE [6], where phonetic transduction takes place at an early stage of processing. However, models like Nearey’s or Massaro’s are far from a form that could actually be attached as a front end to MERGE-like models.

As noted in [4], the models I have applied to synthetic speech experiments are really simple static pattern classifiers. Neatly packaged cues are provided as fixed-dimensional input vectors. Time, if it is used at all, is pre-measured and delivered as just another scalar quantity in that input. This kind information is readily

\[ \text{SWAP} \]
available for synthetic stimuli, since it is part of the recipe for their construction. Obtaining input measurements from speech waveforms (even synthetic ones) is by no means trivial. Nonetheless, there are techniques, such as dynamic programming, that should allow an initial approximation to such an approach. In the near future, we hope to implement such a procedure, starting from techniques described in [17]. It will be instructive (and probably sobering) to try to build a model that inputs waveforms and produces predictions of listener’s behavior using existing modeling frameworks.

Even if such a system meets with modest success, there is still a long way to go. Everything I have described so far applies only to asymptotic categorization behavior; it says nothing about reaction times or any other dynamic properties of listeners’ reaction to stimuli. It is well known that many context effects, including lexical status, tend to increase at longer reaction times. Massaro and Cohen [16] have demonstrated that aspects of such temporally varying effects can be produced by simulating the growth over time of the strength of contextual properties. Although this is suggestive (as are one or two other examples from the literature), it is still a long way from showing how such effects can be implemented in a system that starts with a waveform and extracts all information (including that serves as ‘context’) in plausible, temporally evolving ways. This problem must be addressed if phonetic and lexical access models are ever to come together. The only working models for lexical access are those created by speech technologists [7]. We probably need to spend a little more time looking over their shoulders.

5. ACKNOWLEDGEMENTS

Thanks to R. Smits, J. McQueen and A. Cutler, P. Ladefoged and J. Allen for discussion and encouragement. Work supported by SSHRC. Author’s address: Terrance M. Nearey, University of Alberta, Edmonton AB, Canada T6G 2E7. Email: t.nearey@ualberta.ca.

6. REFERENCES