**Words and Wordoids**

Janet Pierrehumbert

Northwestern University

Workshop on Early Phonological Acquisition

Carry-le-Rouet

October 8, 2001

1) **Phonetics is horrible**

English differs from other languages in how vowel length and fricative length trade off as cues for fricative voicing. Some English /s/ count as /z/ in Finnish. (Flege and Hillenbrand, 1986).

Glottalization as cued by amplitude and f0 functions as a consonant in English (Hillenbrand and Houde 1996, Pierrehumbert and Frisch 1996) but as a contrastive vowel feature in Coatzospan Mixtec (Gerfen and Baker, in press). A comparison of perceptual results shows that the phonetic ranges of these phonological categories overlap.

English and Thai listeners differ in perceptual compensation for coarticulatory nasalization. (Beddor and Krakow, 1999).

English and Shona listeners differ in perceptual compensation for V-V coarticulation (Beddor et al., in press).

2) There is no known case of analogous phonemes showing identical range of variation and perceptual boundaries in two languages.

Regularities in production are generally reflected in perception

---à The categorization of the phonetic space is learned.

3) **Prominence markers are also language particular**

In English, phrasal prominence lengthens the most stressed syllable of the head word of the phrase. In Finnish, the lengthening goes on the second mora of the head word. When the stressed syllable is light (monomoraic), the lengthening is manifested on the post-stress syllable (Lehtonen, 1970).

Urdu and Danish mark stressed syllables with an f0 minimum. The f0 goes up on the next (post-stress) syllable. (Hussain 1997, Gronnum 1992).

This pattern also exists as a pragmatically marked contour in English and German (Pierrehumbert and Steele 1990, Kohler 19XX.)

Downstepped f0 patterns exist in Japanese and English. It is the most common pattern in Japanese and British English. It is pragmatically marked in American English. (Pierrehumbert and Hirschberg, 1990). Japanese has SOV order (implying initial prominence, according to
Nespor et al. 1996). British and American English have SVO order (implying final prominence).

4, 5) F0 contours showing that stress in Urdu is marked with a L*. on the stressed syllable Reproduced from Hussain (1997).

6) F0 contours showing that stress in Danish is marked with a L*+H. Reproduced from Gronnum.

7) F0 contours showing the contrast between a downstepped and non-downstepped contour in Japanese. Reproduced from Pierrehumbert and Beckman (1988)

8, 9) F0 contours showing the contrast between a downstepped and non-downstepped contours in English. Reproduced from Liberman and Pierrehumbert (1984).

10) Phonetic learning involves learning probability distributions over the phonetic space.

--The phonetic space is a high dimensional cognitive map on which a metric of proximity/similarity is defined.

--Phonological elements are labels over this map.

--These labels include positional information, not just segmental identity. Within any given context, segmental distributions are well separated. But they overlap quite a bit across contexts.

Ex: The spectrum of schwa in "classify" is similar to /h/ in "hog". Glottalized /t/ in "kit" is similar to the attack on "Italy". The /z/ in "matches" is similar to the /s/ in "sit".

A concrete proposal: Exemplar theory. Exemplar theory provides a way to incrementally update probability distributions and to compute the net effect of these distributions on perceptual categorization. (Johnson 1997 etc.). Other models with multiple levels of representation and incremental updating can do the same job.

11) Figure reproduced from Johnson (1997) showing basic concepts of exemplar theory. The input is an auditory coding of the speech signal. A covering map provides an analog representation of the phonetic space. Category nodes are labels over this map.

12) Figure from Pierrehumbert (2001a) illustrating statistical classification of incoming stimuli, according to exemplar theory. Figure.

13) INTERPRETATION of CLASSIFICATION FORMULA

New token activates exemplars as a function of how close each is in the parameter space.

The strength of the activated exemplars cumulates in activating labels.

Labels compete through mutual inhibition.

14) EMPIRICAL SUPPORT FOR EXEMPLAR THEORY

I) Experimental findings that speech rate, talker identity, voice characteristics, intonation, and f0 are all retained in long-term memories of words. (Reviewed in Bradlow et al. 1999,)
II) Typological studies showing that languages differ in arbitrarily fine phonetic detail. (Reviewed in Pierrehumbert, Beckman, and Ladd 2001)

III) Prototype effects.

IV) Frequency effects, including: frequency bias in identification; speed-frequency relations; why frequency bias is stronger for acoustically poor stimuli.

15) Figure reproduced from Johnson, Fleming and Wright (1993) showing the vowel hyperspace effect.

16) **Basic Production Model (equations in Pierrehumbert, 2001a)**

Select a label. (More frequent labels selected more often).

Take a random sampling of the exemplar distribution for that label. (Sampling can be biased to model versus effects).

Identify the neighborhood of the selected exemplar.

The average properties of this neighborhood constitute the production goal.

The production goal is executed with noise (e.g. noise in the motor control system).

17-20) Figures reproduced from Pierrehumbert (2001a) showing category collapse as a result of persistent hypo-articulation of a marked category (a phonetically unstable category of low relative frequency). [Stage 1, Stage 2, Stage 3, Stage 4].

21) Figure reproduced from Pierrehumbert (in press) showing hypothetical example of five-way allophony with a robust statistical pattern. [Figure]

22) Figure reproduced from Pierrehumbert (in press) showing hypothetical example of five-way allophony in which overlap of categories leads to statistical collapse of one or more categories. [Figure]

23) Figure reproduced from Suzuki showing category sharpening through a perception-only neural mechanism.

24) And in fact:

Maye et al. (2000, in press) have shown that people exposed to a bimodal phonetic distribution in a hypothetical language impute a lexically contrastive function to the two modes.

Those exposed to a unimodal distribution with the same phonetic range do not impute a contrastive function.

Similar findings obtain for infants.

25) **Relation of phonetic knowledge to the lexicon.**

Common/previous understanding: Bottom up phonemic analysis of the speech stream. Phonemic
string(s) presented to the lexicon for possible matches via a competitive lexical network.

This is impossible because:

--Phonemic decisions reflect lexical information in a task-dependent fashion

26) Figure reproduced from McQueen et al. (2000) showing the structure of the MERGE model of speech perception and lexical access (for full discussion of the model, see Norris et al. 2000).

27) B) Acoustic embeddings do not always correspond to phonemic embeddings.

"two" is NOT found in "tune"
"rune" is NOT found in "poltroon"
"win" is NOT found in "wintery"
"best" IS found in "festoon"
"Peru" IS found in "macaroon"
"blight" IS found in "slight"
"mud" IS found in "plasma adjustment"

Detecting phonemes involves parsing the whole speech signal. Phonemes are not reliably detectable through bottom-up analysis of local invariances.


D) Evidence has accumulated for long-term encoding of word-specific phonetic detail (review in Pierrehumbert, in press)

--People automatically imitate speaker-particular details of words they have heard many times in a particular voice. (Goldinger 1996, 2000).


--Semantic gangs of words can be "left behind" in a historical phoneme shift, if they are used predominately in a certain social circle or type of situation (Yaeger-Dror and Kemp 1992, Yaeger-Dror 1996).

--Morphological paradigm effects include subphonemic detail (Steriade 2000).

29) Conclusions:

The representation which is created bottom-up in speech perception and passed to the lexicon for
possible matches is extremely detailed.

It could be a fine phonetic transcription; however, there is no real evidence that detail is lost; it could be an annotated spectrogram.

The great detail of this representation (paradoxically) increases the need for bottom-up processing, e.g. for a Fast Phonological Preprocessor ("FPP"; Pierrehumbert, 2001). It is necessary to establish possible alignment points between the speech stream and the lexicon, otherwise the search space includes competition between extremely many, extremely similar, minutely staggered, candidates.

Therefore, the primary function of the FPP is to project possible word boundaries.

30) Figure modifying MERGE architecture to show a possibly quantitative input encoding with bottom-up phonological labeling such as possible word boundaries and location of the head mora.

31) Walk through of spectrogram illustrating need for bottom up hypothesis of word boundaries in order to prevent multiple, slightly staggered evaluations of the same word candidate against the lexicon.

32) Allophones and prosodic junctures are probability distributions over speech (a continuous physical process).

Words are probability distributions over allophones and prosodic junctures, and possibly over phonetic parameters.

"Phonemes" are a meta-level, integrating lexical and allophonic information. The "phonemic level" presumably also includes other phonological categories such as syllables and feet.

NB: Morphemes and morphophonological alternations are also meta-lexical.

33) 9 month-olds are sensitive to phonotactics (defined on diphones) (Jusczyk et al. 1994, Mattys et al. 1999).

Problem

Do they know enough words to know what combinations occur word internally and which don't? Pierrehumbert (2001b) estimates 3200 words needed to learn major nasal-obstruent phonotactics.

Do they have phonemes in the sense of the MERGE model?

Proposal

Sequential statistics defined on allophones (e.g. labels over the phonetic space, not phonemes). Allophones are vague compared to adults'. Subsequent development shows both refinement of allophones and projection of phonemes.

Sequential statistics trained on speech not on the lexicon (e.g. on tokens not types).

34) A simple but nonhuman language
8 consonants = \{p, t, k, b, d, g, f, s\}
5 vowels = \{a, e, o, u, i\}

Word --> t V (C V)* s

Some words: tapes, tos, tubudus, tabutis

Continuous speech: tapestostubudustabutis

Observations: /st/ is the only cluster in the language.

It is the most frequent diphone.

It is a 100% reliable boundary cue.

This language is nonhuman because: Human languages have maximal contrast sets (maximal
perplexity) at word onsets. This language has minimal perplexity at word onsets.

Low probability diphones cause babies to impute boundaries. They do this before they have
a lexicon. When they get a lexicon, they can maintain and refine the phonotactically based
boundary detection.

35) PHONOLOGICAL CONSTRAINTS for which implicit stochastic knowledge has been
demonstrated:

Onsets. Rhymes. Syllable junctures. Stress templates. Vowel harmony. OCP (place/manner
Frisch et al. 2001).

These descriptors have no preferred unit and they cross-cut each other. Example:

Can constraints be any arbitrary fragments of phonological description for which lexical statistics
can be established?

36) Results on onsets, rhymes, and junctures confined to diphones.

Long-distance constraints either:target two phonemes separated by irrelevant intervening
material (e.g OCP constraints) or abstract across phonemes (e.g. stress rules, vowel harmony)

CLAIM (Pierrehumbert, 2001b): Phonological constraints are "coarse-grained" because overly
complex/detailed constraints are statistically unstable across variation in vocabulary.

Method: Monte Carlo simulation of vocabulary learning over Celex monomorphemes.
(Monomorphemes define the class of lexical items lacking internal word boundaries.)

Result: Relative goodness of nasal-obstruent clusters learnable from 1/3 of the total word list.
Frequency difference between /gri#/ and /kri#/ not reliably learnable.

37) DIPHONES versus TRIPHONES in general.

Wordlikeness judgments of nonsense words such as "spulsh" and "geltch" are affected by
diphone statistics and lexical neighbors. Triphone statistics have no demonstrable effect (Bailey and Hahn, 2001).

-à Look at statistical trainability of diphones and triphones generally.

38) Can you predict the existence of a diphone (as a well-formed word-internal combination) from the frequencies of the phonemes in it?

Results of a calculation over Celex monomorphemes.

("Exists" == at least one token. "Predicted to exist" == predicted to have a count of 1 or more.)

<table>
<thead>
<tr>
<th></th>
<th>Absent</th>
<th>Exist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted absent</td>
<td>48</td>
<td>7</td>
</tr>
<tr>
<td>Predicted to exist</td>
<td>582</td>
<td>732</td>
</tr>
</tbody>
</table>

Answer: No, very few of the absences are predicted.

Why: Sonority sequencing.

Could you learn these patterns from the available data?

Easily. 1369 possible bigrams. 51,257 tokens of diphones in the training set.

39) Can you predict the existence of a triphone (as a well-formed word-internal combination) from the frequencies of the diphones in it?

<table>
<thead>
<tr>
<th></th>
<th>Absent</th>
<th>Exist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted absent</td>
<td>42 776</td>
<td>892</td>
</tr>
<tr>
<td>Predicted to exist</td>
<td>2 463</td>
<td>4522</td>
</tr>
</tbody>
</table>

Answer: A high degree of success. Most absences are predicted.

Could you learn these patterns from the available data?

No. Compare 50 653 candidate triphone candidates to 39 977 triphone tokens in the training set.

What about the 2 463 and 892 diagonal entries? Mainly counts near one.

-à Can diphone statistics predict the frequency of triphones?

40) Figure from Pierrehumbert (forthcoming) showing median predicted triphone count against
median count observed, plus upper and lower quartiles of observed counts. Figure

41) Analogous figure for diphones showing that prediction of diphone frequencies from phoneme frequencies is much poorer. Figure

42) DIPHONES ARE WHERE THE ACTION IS

Acoustic landmark theory (K. Stevens, 1998)

Transitions from consonants (especially obstruents) are special.

Physics: Rapid spectral change due to changing shape of the vocal tract.

Pressure buildup during occlusion makes releases loud.

Psychoacoustics: Automatic gain control of the ear makes vowel onsets perceptually salient.

Articulatory phonology (Browman and Goldstein, 1986, 1992)

Overlapping of gestures for adjacent sounds leads to assimilations and neutralizations.

43) Conclusions:

Human language exhibits confluence across levels.

Confluence makes incremental learning possible.

Examples include:

Acoustic landmarks -- diphones -- lexicon size/ grammar complexity.

Surface statistics -- positional licensing of contrasts in words.

44) REFERENCES


In Cutler, A., McQueen, J. M. and Zondervan, R. Proceedings of SWAP (Spoken Word Access Processes), Nijmegen, Max Planck Institute for Psycholinguistics. 155-159.


Pierrehumbert,J. (2001b) Why phonology is so coarse grained. In J. McQueen and A. Cutler,
eds. SWAP special issue, Language and Cognitive Processes.


