The Acquisition of Discrete Segmental Categories: Data and Model

Emmanuel Dupoux
Ecole des Hautes Etudes en Sciences Sociales,
Ecole Normale Supérieure,
Centre National de la Recherche Scientifique, Paris
http://www.lscp.net

Abstract

The way in which we parse continuous speech into discrete phonemes is highly language-dependant. Here, we first report that this phenomenon not only depends on the inventory of phonetic distinctions in the language, but also on the inventory of syllabic types. This is illustrated by studies showing that Japanese listeners perceptually insert epenthetic vowels inside illegal consonant clusters in order to make them legal. We then argue that this raises a bootstrapping problem for language acquisition, as the learning of phonetic inventories and syllabic types depend on each other. We present an acquisition model based on the storing and analysis of phonetic syllabic templates. We argue that this model has the potential of solving the bootstrapping problem as well as a range of observation regarding perceptual categorization for speech sounds.

1. The categorization problem

One of the greatest challenge to the study of speech perception is to understand how the continuous signal is parsed into discrete units. The variation of air pressure that conveys speech is a continuous variable; yet, the perception that we have of it is categorical, that is, it rests on a finite number of discrete categories : words, syllables, segments. How is such a mapping performed is a central question in several areas of research ranging from speech perception, to language acquisition and automatic speech recognition.

The first difficulty in this mapping problem was singled out by researchers at Haskins labs and can be labeled the *acoustic/phonetic variability problem*. It is very difficult to find unambiguous cue or set of acoustic cues in the signal that would correspond to a given linguistic segment (see the book by Perkell and Klatt [1]). Lack of invariance is due to differences between talkers vocal tracts, variations in environmental characteristics (reverberation, noise, etc.), speed and style of speech, but also coarticulation effects, which dramatically change the acoustic realization of a given phonetic segment. The fact that acoustic/phonetic variability does not seem to cause problems to humans hearers (both adults and infants) still remains to be accounted for.

There is a second problem that was more recently highlighted, the *cross-linguistic variability problem*. The set of discrete categories that have to be recovered from the signal is not unique but varies dramatically across languages. Certain languages use only 3 vowels, while others use 20; certain languages use only 6 consonants, while other use almost a hundred. As a result, two distinct acoustic sounds will be heard as identical in one language, and as distinct in a different language. This makes the problem of finding out the acoustic correlates of discrete segments even more difficult. Yet, this problem is solved effortlessly by infants who manage to converge on the appropriate linguistic categories for their language within the first year of life [2,3], that is, before they have a very large comprehension lexicon. The mechanisms that underlie such a rapid acquisition still remain to be understood.

In the next section we review further data which show that speech perception not only depends on the inventory of speech sounds, but also on the syllabic types that are allowed in the language.

2. The effect of syllabic structure on speech perception

Speakers of different languages differ in their assessment of the segmental identity of different speech sounds. But they also disagree on **how many** segments there are in a given speech stream. Consider the word in English : “pepsi”. This word is typically considered to have two vowels and three consonants by English listeners. However, this word is reported as having **three** vowels by speakers of

Japanese ("pepusi"). In contrast, speakers of White Hmong will report only **two** consonants ("pesi"). The reason these changes are made to the English form is that neither of the other two languages allow for complex syllabic structures like the one found in /pep.si/. Such phenomena are very commonly observed in the field of loanword phonology. The generalization is that in borrowing words from other languages, speakers typically adapt the illegal forms so that the outcome becomes legal, by substituting, deleting or inserting entire segments[4].

We have recently gathered evidence that at least some of such adaptations are due to automatic processes arising in perception. Specifically, when native speakers of Japanese are presented with the nonword [ebzo], they report hearing the nonword [ebuzo] in the majority of the cases [5]. Furthermore, they have severe difficulties in distinguishing [ebzo] from [ebuzo] in an ABX discrimination [5] and in performing a lexical decision task with real words where the sound [u] has been deleted between consonants that make an illegal sequence (like in sokudo→sokdo) [6]. Furthermore, the Mismatch Negativity response elicited to the change between [ebzo] and [ebuzo] in French hearers is suppressed in Japanese hearers, showing that the two stimuli are already considered identical at very early stages during perception [7]. Finally, such language-dependant phenomena implicate areas in the Planum Temporale and in the Supramarginal Gyrus, areas of the brain involved in the early processing of speech sounds [8]. All these data indicate that there is a process *in perception* that transforms the acoustic signal for [ebzo] into an abstract representation of the form [ebuzo].

3. Bootstrapping phonetic categories

The above observations suggests that the processing of individual segments is dependant on higher order units, like syllables. This raises several puzzles. One relate to processing models, the other to theories of language acquisition. As regards processing models, most of them propose that segments or subsegmnetal units like features are the basic processing units[9,10]. In these models, the effects we report above are unaccounted for; they only way they could account for our result would be to supplement them with a syllabic level, which would interact very intimately with the segmental level. Indeed, in a language with a small inventory of syllabic types, even very tiny acoustic information will count as a vowel (if is occurs inside an illegal sequence of consonants), whereas the same information will simply be ignored in a language with more complex syllabic types. However, even patching the current processing models would leave the acquisition puzzle untouched.

Indeed, the complex interaction between acoustic cues for segments and syllabic context suggests that before learning the acoustic characteristics of segments, babies first need to learn the syllabic types of their language. Vice versa, it would seem that a prerequisite for learning the syllabic types of a language is to parse the signal into discrete consonants and vowels. This create a boostrapping problem of the chicken and egg type, for the acquisition of phonetic and syllabic categories.

There are several avenues to solve such a problem. For instance, Ramus has proposed that rythmic information can be reliably extracted from the acoustic signal, and that rhythm gives indication regarding the syllabic grammar[11]. One could then propose that guesses regarding the syllabic grammar are used to guide the establishment of segmental categories. Here, we propose a related but distinct idea that would propose that entire syllables are learned before segmental categories are established.

4. The SyllCat model.

The present proposal rests on the idea first expressed by Mehler [12] that syllables are natural perceptual units for infants. For instance, with his collaborators, Mehler shown that newborns can discriminate between strings of bisyllables and trisyllables across large variations in phonetic content, that is, they can count the number of syllables in a speech stream [13]. Such ability presumably rests on the acoustic correlates of vocalic nuclei which make them quite salient compared to consonants [14]. The SyllCat model presented in Figure 1 exploits the idea that segmentation into syllabic chunks is a primitive process and acoustically much simpler than segmentation into phonetic segments. The model is composed of two levels: a syllabic bank and a segmental bank, and is designed to account for both learning in the infant and processing in the adult.

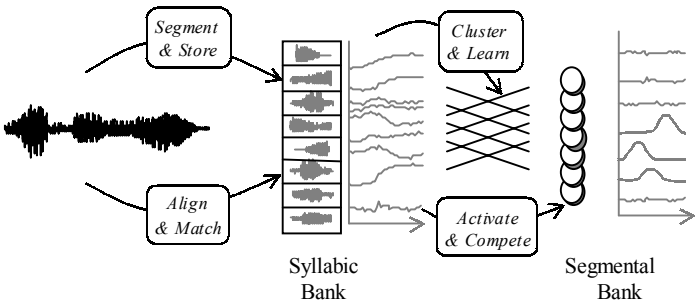

Figure 1. Outline of the SyllCat model composed of two banks of detectors, the instance-based acoustic syllable bank, and the abstract segmental bank. The graphs represent the time cours of activation of each detector.Elements with dotted lines correspond to the acquisition phase (mostly in infants), and with solid lines to the processing phase (infants and adults).

The infant starts out by segmenting out the acoustic counterpart of syllables from the signal and storing them as acoustic templates in an instance-based memory system. Such a segmentation process could rest on the extraction of energy or sonority contours as defined in [15]. Once a sufficient number of templates have been accumulated (or after a given time of exposure), infants start to analyse the structure of the similarity space between the various syllabic templates they have already stored. For instance, the templates for the syllables [ba], [be], [bi], [bo], [bu] have all similar spectral components in their initial parts but not in their final part. Vice versa for the syllables [pa], [ta], [ka], [ba], [da], [ga]. The analysis of this similarity structure could give rise, through a clustering algorithm, to the emergence of discrete segmental units (/b/, /a/, etc). Obviously the result of such clustering would then depend on the segmental inventory of the language. A language with a distinction between aspirated and unaspirated stops would give rize to two segmental detectors for each stop (/t/, and /t^h/). One can cast the result of such learning in a perceptron-type array of segment detectors, each connected to the array of syllabic detectors through weights that have been set according to the clustering algorithm.

In the adult, such a system would be stabilized, and would be little or not influenced by further language experience. A given speech input would then be matched with all the stored syllabic exemplars for each time slice, and the time series of degree of match vectors would then feed onto the segment detectors. After a stage of competitive activation, this would give rize to a unique segmental interpretation of the signal for each time slice. That is, the system would have achieved categorization into discrete and language-specific segmental categories. It has also the potential to account for compensation for within-syllable coarticulation effects, as each syllabic template contains detailed phonetic information regarding the entire transition between consonants and vowels in the articulatory gesture.

In such a model, the effects of syllabic structure on the perception of segments can be straightforwardly explained: if a sound in the environment contains a sequence of segment which is illegal in the language, there will be no syllable in the syllable bank that corresponds to it. As a result, the signal will weakly match a number of syllabic templates that approximates the input and which will jointly activate their segmental elements. The perceptual outcome will be the result of the educated guesses of all the syllabic detectors. In japanese for instance, the presentation of [ebzo] will strongly activate the syllabic detectors for [e] and then weakly [ba,be,bi,bo,bu], and then [zo]. One can see that the output of the syllabic bank militates in favor of the existence of a vowel between [b] and [z]. Which particular vowel will win out in the competition depends on many factors, the most important of which being the spectral proximity of each of the syllabic detectors with the particular phonetic cues present at the transition between [b] and [z]. As it turns out, [u] is the most centralized vowel in Japanese, and is also typically the shortest vowel. This is why [u] wins out in most of the case (although we have recently shown that presence of coarticulation can bias the perception towards other vowels). Of course, if the language contained the syllable [eb], one can see that no vowel would ever be inserted.

Further research is needed, in particular in the computational implementation of such a model to derive quantitative predictions and match them with the observed experimental results.

5. Acknowledgments

This work was supported in part by a CNRS grant Aide à Projet Nouveau entitled "Les surdités phonologiques: Etudes interlangues", and an interdisciplinary CNRS grant “Acquisition phonologique précoce : algorithmes et simulations”. I would like to thank Sharon Peperkamp and Janet Pierrehumbert for useful comments.

6. References

[1] Perkell, J.S. and Klatt, D. (eds.) (1986). *Invariance and Variability of Speech Processes*, Lawrence Erlbaum Assoc., Hillsdale, NJ

[2] Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, **7**, 49–63.

[3] Kuhl, P., Williams, K., Lacerda, F., Stevens, K. & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science,* **55**, 606-608.

[4] Peperkamp, S. & E. Dupoux (2003) Reinterpreting loanword adaptations: The role of perception. In: Proceedings of the 15th International Congress of Phonetic Sciences.

[5] Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, **25(6)**, 1568-1578.

[6] Dupoux, E., Pallier, C., Kakehi, K., & Mehler, J. (2001). New evidence for prelexical phonological processing in word recognition. *Language and Cognitive Processes*, **16**, 491-505,

[7] Dehaene-Lambertz, G., Dupoux, E., & Gout, A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of Cognitive Neuroscience*, **12**, 635-647.

[8] Jacquemot C., Pallier C., Lebihan D., Dehaene S. & Dupoux E. (2003). Phonological grammar shapes the auditory cortex: a functional Magnetic Resonance Imaging study. *Journal of Neuroscience*, **23(29)**, 9541-9546.

[9] McClelland JL, Elman JL (1986). The TRACE model of speech perception, *Cognitive Psychology* **18(1)**, 1-86.

[10] Norris, D. (1994). Shortlist : A connectionnist model of continuous speech recognition. *Cognition*, **52**, 189-234.

[11] Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, **73(3)**, 265-292.

[12] Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining Models of Lexical Access: The onset of word recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 236-262). Cambridge Mass: MIT Press.

[13] Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant perception. *Infant Behavior and Development*, **4**, 247-260.

[14] Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J. L. Morgan & K. Demuth (Eds.), Signal to Syntax: Bootstrapping from speech to grammar in early acquisition (pp. 101-116). Mahwah, NJ: Lawrence Erlbaum Associates.

[15] Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In Speech Prosody 2002, Aix-en-Provence.