# Exploring the Relative Role of Bottom-up and Top-down Information in Phoneme Learning

**Abdellah Fourtassi[1] , Thomas Schatz[1,2] , Balakrishnan Varadarajan[3] , Emmanuel Dupoux[1]**

{abdellah.fourtasi; emmanuel.dupoux}@gmail, thomas.schatz@laposte.net, bvarada2@jhu.edu

[1] Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS, Paris, France
[2] SIERRA Project-Team, INRIA/ENS/CNRS, Paris, France
[3] Center for Language and Speech Processing, JHU, Baltimore, USA

## Abstract

We test both bottom-up and top-down approaches in learning the phonemic status of the sounds of English and Japanese. We used large corpora of spontaneous speech to provide the learner with an input that models both the linguistic properties and statistical regularities of each language. We found both approaches to help discriminate between allophonic and phonemic contrasts with a high degree of accuracy, although top-down cues proved to be effective only on an interesting subset of the data.

## 1 Introduction

Developmental studies have shown that, during their first year, infants tune in on the phonemic categories (consonants and vowels) of their language, i.e., they lose the ability to distinguish some within-category contrasts (Werker and Tees, 1984) and enhance their ability to distinguish between-category contrasts (Kuhl et al., 2006). Current work in early language acquisition has proposed two competing hypotheses that purport to account for the acquisition of phonemes. The *bottom-up hypothesis* holds that infants converge on the linguistic units of their language through a similarity-based distributional analysis of their input (Maye et al., 2002; Vallabha et al., 2007). In contrast, the *top-down hypothesis* emphasizes the role of higher level linguistic structures in order to learn the lower level units (Feldman et al., 2013; Martin et al., 2013). The aim of the present work is to explore how much information can ideally be derived from both hypotheses.

The paper is organized as follows. First we describe how we modeled phonetic variation from audio recordings, second we introduce a bottom-up cue based on acoustic similarity and top-down cues based of the properties of the lexicon. We test their performance in a task that consists in discriminating within-category contrasts from between-category contrasts. Finally we discuss the role and scope of each cue for the acquisition of phonemes.

## 2 Modeling phonetic variation

In this section, we describe how we modeled the representation of speech sounds putatively processed by infants, before they learn the relevant phonemic categories of their language. Following Peperkamp et al. (2006), we make the assumption that this input is quantized into context-dependent phone-sized unit we call *allophones*. Consider the example of the allophonic rule that applies to the French /r/:

$$/\text{r}/ \rightarrow \begin{cases} [\chi]\ / & \text{before a voiceless obstruent} \\ [\text{ʁ}] & \text{elsewhere} \end{cases}$$

Figure 1: Allophonic variation of French /r/

The phoneme /r/ surfaces as voiced ([ʁ]) before a voiced obstruent like in [kanaʁ ʒon] ("canard jaune", yellow duck) and as voiceless ([χ]) before a voiceless obstruent as in [kanaχ puχpʁ] ("canard pourpre", purple duck). Assuming speech sounds are coded as allophones, the challenge facing the learner is to distinguish the allophonic variation ([ʁ], [χ]) from the phonemic variation (related to a difference in the meaning) like the contrast ([ʁ],[l]).

Previous work has generated allophonic variation using random contexts (Martin et al., 2013). This procedure does not take into account the fact that contexts typically belong to natural classes. In addition, it does not enable to compute an acoustic distance. Here, we generate linguistically and acoustically controlled allophones using Hidden Markov Models (HMMs) trained on real audio recordings.

### 2.1 Corpora

We use two speech corpora: the Buckeye Speech corpus (Pitt et al., 2007), which consists of 40 hours of spontaneous conversations with 40 speakers of American English, and the core of the Corpus of Spontaneous Japanese (Maekawa et al., 2000) which also consists of about 40 hours of recorded spontaneous conversations and public speeches in different fields. Both corpora are time-aligned with phonetic labels. Following Boruta (2012), we relabeled the japanese corpus using 25 phonemes. For English, we used the phonemic version which consists of 45 phonemes.

### 2.2 Input generation

#### 2.2.1 HMM-based allophones

In order to generate linguistically and acoustically plausible allophones, we apply a standard Hidden Markov Model (HMM) phoneme recognizer with a three-state per phone architecture to the signal, as follows.

First, we convert the raw speech waveform of the corpora into successive vectors of Mel Frequency Cepstrum Coefficients (MFCC), computed over 25 ms windows, using a period of 10 ms (the windows overlap). We use 12 MFCC coefficients, plus the energy, plus the first and second order derivatives, yielding 39 dimensions per frame. Second, we start HMM training using one three-state model per phoneme. Third, each phoneme model is cloned into context-dependent triphone models, for each context in which the phoneme actually occurs (for example, the phoneme /ɑ/ occurs in the context [d–ɑ–g] as in the word /dɑg/ (“dog”). The triphone models are then retrained on only the relevant subset of the data, corresponding to the given triphone context. These detailed models are clustered back into inventories of various sizes (from 2 to 20 times the size of the phonemic inventory) using a linguistic feature-based decision tree, and the HMM states of linguistically similar triphones are tied together so as to maximize the likelihood of the data. Finally, the triphone models are trained again while the initial gaussian emission models are replaced by mixture of gaussians with a progressively increasing number of components, until each HMM state is modeled by a mixture of 17 diagonal-covariance gaussians. The HMM were built using the HMM Toolkit (HTK: Young et al., 2006).

#### 2.2.2 Random allophones

As a control, we also reproduce the random allophones of Martin et al. (2013), in which allophonic contexts are determined randomly: for a given phoneme $/p/$, the set of all possible contexts is randomly partitioned into a fixed number $n$ of subsets. In the transcription, the phoneme $/p/$ is converted into one of its allophones ($p_1$,$p_2$,..,$p_n$) depending on the subset to which the current context belongs.

## 3 Bottom-up and top-down hypotheses

### 3.1 Acoustic cue

The bottom-up cue is based on the hypothesis that instances of the same phoneme are likely to be acoustically more similar than instances of two different phonemes (see Cristia and Seidl, in press) for a similar proposition). In order to provide a proxy for the perceptual distance between allophones, we measure the information theoretic distance between the acoustic HMMs of these allophones. The 3-state HMMs of the two allophones were aligned with Dynamic Time Warping (DTW), using as a distance between pairs of emitting states, a symmetrized version of the Kullback-Leibler (KL) divergence measure (each state was approximated by a single non-diagonal Gaussian):

$$A(x,y) = \sum_{(i,j)\in DTW(x,y)} KL(N_{x_i}||N_{y_j}) + KL(N_{y_j}||N_{x_i})$$

Where $\{(i,j) \in DTW(x,y)\}$ is the set of index pairs over the HMM states that correspond to the optimal DTW path in the comparison between phone model $x$ and $y$, and $N_{x_i}$ the full covariance Gaussian distribution for state $i$ of phone $x$. For obvious reasons, the acoustic distance cue cannot be computed for Random allophones.

### 3.2 Lexical cues

The top-down information we use in this study, is based on the insight of Martin et al. (2013). It rests on the idea that true lexical minimal pairs are not very frequent in human languages, as compared to minimal pairs due to mere phonological processes. In fact, the latter creates variants (alternants) of the same lexical item since adjacent sounds condition the realization of the first and final phoneme. For example, as shown in figure 1, the phoneme /r/ surfaces as [χ] or [ʁ] depending on whether or not the next sound is a voiceless obstruent. Therefore, the lexical item /kanar/ surfaces as [kanaχ] or [kanaʁ]. The lexical cue assumes that a pair of words differing in the first or last segment (like [kanaχ] and [kanaʁ]) is more likely to be the result of a phonological process triggered by adjacent sounds, than a true semantic minimal pair.

However, this strategy clearly gives rise to false alarms in the (albeit relatively rare) case of true minimal pairs like [kanaχ] ("duck") and [kanal] ("canal"), where ([χ], [l]) will be mistakenly labeled as allophonic.

In order to mitigate the problem of false alarms, we also use Boruta (2011)'s continuous version, where each pair of phones is characterized by the number of lexical minimal pairs it forms.

$$B(x, y) = |(Ax, Ay) \in L^2| + |(xA, yA) \in L^2|$$

where $\{Ax \in L\}$ is the set of words in the lexicon $L$ that end in the phone $x$, and $\{(Ax, Ay) \in L^2\}$ is the set of phonological minimal pairs in $L \times L$ that vary on the final segment.

In addition, we introduce another cue that could be seen as a normalization of Boruta's cue:

$$N(x, y) = \frac{|(Ax,Ay)\in L^2|+|(xA,yA)\in L^2|}{|\{Ax\in L\}|+|\{Ay\in L\}|+|\{xA\in L\}|+|\{yA\in L\}|}$$

# 4 Experiment

## 4.1 Task

For each corpus we list all the possible pairs of attested allophones. Some of these pairs are allophones of the same phoneme (allophonic pair) and others are allophones of different phonemes (non-allophonic pairs). The task is a same-different classification, whereby each of these pairs is given a score from the cue that is being tested. A good cue gives higher scores to allophonic pairs.

## 4.2 Evaluation

We use the same evaluation procedure as in Martin et al. (2013). It is carried out by computing the Receiver Operating Characteristic (ROC) curve (varying the z-score threshold and computing the resulting proportions of misses and false alarms). We then derive the Area Under the Curve (AUC), which also corresponds to the probability that given two pairs of phones, one allophonic, one not, they are correctly classified. A value of 0.5 represents chance and a value of 1 represents perfect performance.

In order to lessen the potential influence of the structure of the corpus (mainly the order of the utterances) on the results, we use a statistical resampling scheme. The corpus is divided into small blocks (of 20 utterances each). In each run, we draw randomly with replacement from this set of blocks a sample of the same size as the original corpus. This sample is then used to retrain the acoustic models and generate a phonetic inventory that we use to re-transcribe the corpus and re-compute the cues. We report scores averaged over 5 such runs.

## 4.3 Results

Table 1 shows the classification scores for the lexical cues when we vary the inventory size from 2 allophones per phoneme in average, to 20 allophones per phoneme, using the Random allophones. The top-down scores are very high, replicating Martin et al.'s results, and even improving the performance using Boruta's cue and our new Normalized cue.

| | English | | | Japanese | | |
|---|---|---|---|---|---|---|
| Allo./phon. | M | B | N | M | B | N |
| 2 | 0.784 | 0.935 | **0.951** | 0.580 | 0.989 | **1.00** |
| 5 | 0.845 | 0.974 | **0.982** | 0.653 | 0.978 | **0.991** |
| 10 | 0.886 | 0.974 | **0.981** | 0.733 | 0.944 | **0.971** |
| 20 | 0.918 | 0.961 | **0.966** | 0.785 | 0.869 | **0.886** |

Table 1 : Same-different scores for top-down cues on Random allophones, as a function of the average number of allophones per phoneme. M=Martin et al., B=Boruta, N= Normalized

Table 2 shows the results for HMM-based allophones. The acoustic score is very accurate for both languages and is quite robust to variation. Top-down cues, on the other hand, perform, surprisingly, almost at chance level in distinguishing between allophonic and non-allophonic pairs. A similar discrepancy for the case of Japanese was actually noted, but not explained, in Boruta (2012).

| | English | | | | Japanese | | | |
|---|---|---|---|---|---|---|---|---|
| Allo./phon. | A | M | B | N | A | M | B | N |
| 2 | **0.916** | 0.592 | 0.632 | 0.643 | **0.885** | 0.422 | 0.524 | 0.537 |
| 5 | **0.918** | 0.592 | 0.607 | 0.611 | **0.908** | 0.507 | 0.542 | 0.551 |
| 10 | **0.893** | 0.569 | 0.571 | 0.571 | **0.827** | 0.533 | 0.546 | 0.548 |
| 20 | **0.879** | 0.560 | 0.560 | 0.559 | **0.876** | 0.541 | 0.543 | 0.543 |

Table 2 : Same-different scores for bottom-up and top-down cues on HMM-based allophones, as a function of the average number of allophones per phoneme. A=Acoustic, M=Martin et al., B=Boruta, N= Normalized

In the following section, we will discuss the counter-intuitive result obtained above, and explore why top-down scores degrade when realistic allophones are used.

## 5 Analysis

### 5.1 Why does the performance drop for realistic allophones?

When we list all possible pairs of allophones in the inventory, some of them correspond to lexical alternants ([χ], [ʁ]) → ([kanaχ] and [kanaʁ]), others to true minimal pairs ([ʁ], [l]) → ([kanaʁ] and [kanal]), and yet others will simply not generate lexical variation at all, we will call those: *invisible* pairs. For instance, in English, /h/ and /ŋ/ occur in different syllable positions and thus cannot appear in any minimal pair. As defined above, top-down cues are set to 0 in such pairs (which means that they are systematically classified as non-allophonic). This is a correct decision for /h/ vs. /ŋ/, but not for invisible pairs that also happen to be allophonic, resulting in false negatives. In tables 3, we show that, indeed, invisible pairs is a major issue, and could explain to a large extent the pattern of results found above. In fact, the proportion of visible allophonic pairs ("allo" column) is way lower for HMM-based allophones. This means that the majority of allophonic pairs in the HMM case are invisible, and therefore, will be mistakenly classified as non-allophonic.

| | **Random** | | | | **HMM** | | | |
|---|---|---|---|---|---|---|---|---|
| | English | | Japanese | | English | | Japanese | |
| Allo./phon. | allo | ¬ allo | allo | ¬ allo | allo | ¬ allo | allo | ¬ allo |
| 2 | 92.9 | 36.3 | 100 | 83.9 | 48.9 | 25.3 | 37.1 | 53.2 |
| 5 | 97.2 | 28.4 | 99.6 | 69.0 | 31.1 | 14.3 | 25.0 | 25.9 |
| 10 | 96.8 | 19.9 | 96.7 | 50.1 | 19.8 | 4.23 | 21.0 | 14.4 |
| 20 | 94.3 | 10.8 | 83.4 | 26.4 | 14.0 | 1.89 | 12.4 | 4.04 |

Table 3 : Proportion (in %) of allophonic pairs (allo), and non-allophonic pairs (¬ allo) associated with at least one lexical minimal pair, in Random and HMM allophones.

There are basically two reasons why an allophonic pair would be invisible ( will not generate lexical alternants). The first one is the absence of evidence, e.g., if the edges of the word with the underlying phoneme do not appear in enough contexts to generate the corresponding variants. This can happen when the corpus is so small that no word ending with, say, /r/ appears in both voiced and voiceless contexts. The second, is when the allophones are triggered on maximally different contexts (on the right and on the left), as illustrated below:

$$/\mathrm{p}/ \rightarrow \begin{cases} [\mathrm{p}_1]\ /\ \mathrm{A}\_\_\mathrm{B} \\ [\mathrm{p}_2]\ /\ \mathrm{C}\_\_\mathrm{D} \end{cases}$$

When A doesn't overlap with C and B does not overlap with D, it becomes impossible for the pair ([p$_1$], [p$_2$]) to generate a lexical minimal pair. This is simply because a pair of allophones needs to share at least one context to be able to form variants of a word (the second or penultimate segment of this word).

When asked to split the set of contexts in two distinct categories that trigger [p$_1$] and [p$_2$] (i.e., A__B and C__D), the random procedure will often make A overlap with B and C overlap with D because it is completely oblivious to any acoustic or linguistic similarity, thus making it always possible for the pair of allophones to generate lexical alternants. A more realistic categorization (like the HMM-based one), will naturally tend to minimize within-category distance, and maximize between-category distance. Therefore, we will have less overlap, making the chances of the pair to generate a lexical pair smaller. The more allophones we have, the bigger is the chance to end up with non-overlapping categories (invisible allophonic pairs), and the more mistakes will be made, as shown in Table 3.

### 5.2 Restricting the role of top-down cues

The analysis above shows that top-down cues cannot be used to classify all contrasts. The approximation that consists in considering all pairs that do not generate lexical pairs as non-allophonic, does not scale up to realistic input. A more intuitive, but less ambitious, assumption is to restrict the scope of top-down cues to contrasts that do generate lexical variation (lexical alternants or true minimal pairs). Thus, they remain completely agnostic to the status of invisible pairs. This restriction makes sense since top-down information boils down to knowing whether two word forms belong to the same lexical category (reducing variation to allophony), or to two different categories (variation is then considered non-allophonic). Phonetic variation that does not cause lexical variation is, in this particular sense, orthogonal to our knowledge about the lexicon.

We test this hypothesis by applying the cues only to the subset of pairs that are associated with

| | English | | | | | | | Japanese | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Individual cues | | | | Combination | | | Individual cues | | | | Combination | |
| Allo./phon. | * (%) | A | A* | B* | N* | A*+B* | A*+N* | * (%) | A | A* | B* | N* | A*+B* | A*+N* |
| 2 | 26.6 | 0.916 | 0.965 | 0.840 | 0.950 | 0.971 | **0.994** | 60.92 | 0.885 | 0.909 | 0.859 | 0.906 | 0.918 | **0.946** |
| 4 | 14.3 | 0.918 | 0.964 | 0.858 | 0.951 | 0.975 | **0.991** | 30.88 | 0.908 | 0.917 | 0.850 | 0.936 | 0.934 | **0.976** |
| 10 | 4.24 | 0.893 | 0.937 | 0.813 | 0.939 | 0.960 | **0.968** | 16.06 | 0.827 | 0.839 | 0.899 | **0.957** | 0.904 | 0.936 |
| 20 | 1.67 | 0.879 | 0.907 | 0.802 | 0.907 | **0.942** | 0.940 | 5.02 | 0.876 | 0.856 | 0.882 | **0.959** | 0.913 | 0.950 |

Table 4 : Same-different scores for different cues and their combinations with HMM-allophones, as a function of average number of allophones per phonemes.

| | English | | | | | | | Japanese | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Individual cues | | | | Combination | | | Individual cues | | | | Combination | |
| Size (hours) | * (%) | A | A* | B* | N* | A*+B* | A*+N* | * (%) | A | A* | B* | N* | A*+B* | A*+N* |
| 1 | 9.87 | 0.885 | 0.907 | 0.741 | 0.915 | 0.927 | **0.969** | 34.78 | 0.890 | 0.883 | 0.835 | 0.915 | 0.889 | **0.934** |
| 4 | 18.3 | 0.918 | 0.958 | 0.798 | 0.917 | 0.967 | **0.989** | 48.00 | 0.917 | 0.939 | 0.860 | 0.937 | 0.938 | **0.973** |
| 8 | 21.3 | 0.916 | 0.964 | 0.837 | 0.942 | 0.971 | **0.992** | 51.71 | 0.915 | 0.940 | 0.889 | 0.937 | 0.954 | **0.977** |
| 20 | 24.4 | 0.911 | 0.960 | 0.827 | 0.936 | 0.969 | **0.994** | 58.12 | 0.921 | 0.954 | 0.865 | 0.912 | 0.945 | **0.971** |
| 40 | 26.6 | 0.916 | 0.965 | 0.840 | 0.950 | 0.971 | **0.994** | 60.92 | 0.885 | 0.909 | 0.859 | 0.906 | 0.918 | **0.946** |
| $\infty$ | 34.82 | | | | | | | 72.16 | | | | | | |

Table 5 : Same-different scores for different cues and their combinations with HMM-allophones, as a function of corpus size. * (%) refers to the proportion of the subset of contrasts associated with at least one minimal pair. The cues applied to this subset are marked with an asterisk (*)

at least one lexical minimal pair. We vary the number of allophones per phoneme on the one hand (Table 4) and the size of the input on the other hand (Table 5). We refer to this subset by an asterisk (*), by which we also mark the cues that apply to it. Notice that, in this new framing, the M cue is completely uninformative since it assigns the same value to all pairs.

As predicted, the cues perform very well on this subset, especially the N cue. The combination of top-down and bottom-up cues shows that the former is always useful, and that these two sources of information are not completely redundant. However, the scope of top-down cues (the proportion of the subset * ) shrinks as we increase the number of allophones. Table 5 shows that this problem can, in principle, be mitigated by increasing the amount of data available to the learner. As we were limited to only 40 hours of speech, we generated an artificial corpus that uses the same lexicon but with all possible word orders so as to maximize the number of contexts in which words appear. This artificial corpus increases the proportion of the subset, but we are still not at 100 % coverage, which according the analysis above, is due (at least in part) to the irreducible set of non-overlapping pairs.

## 6 Conclusion

In this study we explored the role of both bottom-up and top-down hypotheses in learning the phonemic status of the sounds of two typologically different languages. We introduced a bottom-up cue based on acoustic similarity, and we used already existing top-down cues to which we provided a new extension. We tested these hypotheses on English and Japanese, providing the learner with an input that mirrors closely the linguistic and acoustic properties of each language. We showed, on the one hand, that the bottom-up cue is a very reliable source of information, across different levels of variation and even with small amount of data. Top-down cues, on the other hand, were found to be effective only on a subset of the data, which corresponds to the interesting contrasts that cause lexical variation. Their role becomes more relevant as the learner gets more linguistic experience, and their combination with bottom-up cues shows that they can provide non-redundant information. Note, finally, that even if this work is based on a more realistic input compared to previous studies, it still uses simplifying assumptions, like ideal word segmentation, and no low-level acoustic variability. Those assumptions are, however, useful in quantifying the information that can ideally be extracted from the input, which is a necessary preliminary step before modeling *how* this input is used in a cognitively plausible way. Interested readers may refer to (Fourtassi and Dupoux, 2014; Fourtassi et al., 2014) for a more learning-oriented approach, where some of the assumptions made here about high level representations are relaxed.

## Acknowledgments

This project is funded in part by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the Ecole de Neurosciences de Paris, and the Région Ile de France (DIM cerveau et pensée). We thank Luc Boruta, Sanjeev Khudanpur, Isabelle Dautriche, Sharon Peperkamp and Benoit Crabbé for highly useful discussions and contributions.

## References

Luc Boruta. 2011. Combining Indicators of Allophony. In *Proceedings ACL-SRW*, pages 88–93.

Luc Boruta. 2012. *Indicateurs d'allophonie et de phonémicité*. Doctoral dissertation, Université Paris-Diderot - Paris VII.

A. Cristia and A. Seidl. In press. The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*.

Naomi H. Feldman, Thomas L. Griffiths, Sharon Goldwater, and James L. Morgan. 2013. A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4):751–778.

Abdellah Fourtassi and Emmanuel Dupoux. 2014. A rudimentary lexicon and semantics help bootstrap phoneme acquisition. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)*.

Abdellah Fourtassi, Ewan Dunbar, and Emmanuel Dupoux. 2014. Self-consistency as an inductive bias in early language acquisition. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Patricia K. Kuhl, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9(2):F13–F21.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of japanese. In *LREC*, pages 947–952, Athens, Greece.

Andrew Martin, Sharon Peperkamp, and Emmanuel Dupoux. 2013. Learning phonemes with a proto-lexicon. *Cognitive Science*, 37(1):103–124.

J. Maye, J. F. Werker, and L. Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82:B101–B111.

Sharon Peperkamp, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41.

M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and Fosler-Lussier. 2007. Buckeye corpus of conversational speech.

G.K. Vallabha, J.L. McClelland, F. Pons, J.F. Werker, and S. Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273.

Janet F. Werker and Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49 – 63.

Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. 2006. *The HTK Book Version 3.4*. Cambridge University Press.