Combining Indicators of Allophony

Luc Boruta
Univ. Paris Diderot, Sorbonne Paris Cité, ALPAGE, UMR-I 001 INRIA, F-75205, Paris, France
LSCP, Département d'Études Cognitives, École Normale Supérieure, F-75005, Paris, France
luc.boruta@inria.fr

Abstract

Allophonic rules are responsible for the great variety in phoneme realizations. Infants can not reliably infer abstract word representations without knowledge of their native allophonic grammar. We explore the hypothesis that some properties of infants' input, referred to as indicators, are correlated with allophony. First, we provide an extensive evaluation of individual indicators that rely on distributional or lexical information. Then, we present a first evaluation of the combination of indicators of different types, considering both logical and numerical combinations schemes. Though distributional and lexical indicators are not redundant, straightforward combinations do not outperform individual indicators.

1 Introduction

Though the phonemic inventory of a language is typically small, phonetic and phonological processes yield manifold variants[1] for each phoneme. Words too are affected by this variability, yielding different realizations for a given underlying form. Allophonic rules relate phonemes to their variants, expressing the contexts in which the latter occur. We are interested in describing procedures by which infants, learning their native allophonic grammar, could reduce the variation and recover words. Combining insights from both computational and behavioral studies, we endorse the hypothesis that infants are good distributional learners (Maye et al., 2002; Saffran et al., 1996) and that they may 'bootstrap' into language tracking statistical regularities in the signal.

We seek to identify which features of infants' input are most reliable for learning allophonic rules. A few indicators, based on distributional (Peperkamp et al., 2006) and lexical (Martin et al., submitted) information, have been described and validated *in silico*.[2] Yet, other aspects have barely been addressed, e.g. the question of whether or not these indicators capture different aspects of allophony and, if so, which combination scheme yields better results.

We present an extensive evaluation of individual indicators and, based on theoretical and empirical desiderata, we outline a more comprehensive framework to model the acquisition of allophonic rules.

2 Indicators of allophony

We build upon Peperkamp et al.'s framework: the task is to induce a two-class classifier deciding, for every possible pair of segments, whether or not they realize the same phoneme. Discrimination relies on indicators, i.e. linguistic properties which are correlated with allophony. As a model of language acquisition, this classifier is induced without supervision.

In line with previous studies, we assume that infants are able to segment the continuous stream of acoustic input into a sequence of discrete segments, and that they quantize each of these segments into one of a finite number of phonetic categories. Quantization is a necessary assumption for the framework to apply. However, the larger the set of phonetic categories, the closer we get to recent 'single-stage' approaches (e.g. work by Dillon et al., in preparation) where phonological categories are acquired directly from raw infant-directed speech.

[1] We use *allophony* as an umbrella term for the continuum ranging from typical allophones to mere coarticulatory variants.

[2] See also the work of Dautriche (2009) on acoustic indicators of allophony, albeit using adult-directed speech.

Proceedings of the ACL-HLT 2011 Student Session, pages 88–93,
Portland, OR, USA 19-24 June 2011. ©2011 Association for Computational Linguistics

2.1 Distributional indicators

Complementary distribution is a ubiquitous criterion for the discovery of phonemes. If two segments occur in mutually exclusive contexts, the two may be realizations of the same phoneme.

Bearing in mind that the signal may be noisy, Peperkamp et al. (2006) looked for segments in near-complementary distributions. Using the symmetrised Kullback–Leibler divergence (henceforth KL), they compared the probability distributions of how often the contexts of each segment occur. In a follow-up study, Le Calvez (2007) compared KL to other indicators, namely the Jensen–Shannon divergence (JS) and the Bhattacharyya coefficient (BC).[3]

2.2 Lexical indicators

Adjacent segments can condition the realization of a word's initial and final phonemes. If two words only differ by their initial or final segments, these segments may be realizations of the same phoneme. Instantiating the general concept of functional load (Hockett, 1955), lexical indicators gauge the degree of contrast in the lexicon between two segments.

Using the simplest expression of functional load, Martin et al. (submitted) defined a Boolean-valued indicator, FL, satisfied by a single pair of minimally different words. As a result, FL is sensitive to noise. We define a finer-grained variant, FL*, which tallies the number of such pairs. Moreover, as words get longer, it becomes increasingly unlikely that such word pairs occur by chance. Thus, for any such pair, FL* is incremented by the length of those words.

We also propose an information-theoretic lexical indicator, HFL, based on Hockett's definition of functional load. HFL accounts for the fraction of information content, represented by the language's word entropy, that is lost when the opposition between two segments is neutralized. The 'broken typewriter' function used for neutralization guarantees that values lie in $[0, 1]$ (Coolen et al., 2005).

3 Corpora and experimental setup

In the absence of phonetic transcriptions of infant-directed speech, and as the number of allophones infants must learn is unknown (if assessable at all), we use Boruta et al.'s (submitted) corpora. They created a range of possible inputs, applying artificial allophonic grammars[4] of different sizes (Boruta, 2011) to the now-standard CHILDES 'Brent/Ratner' corpus of English (Brent and Cartwright, 1996). We quantify the amount of variation in a corpus by its allophonic complexity, i.e. the ratio of the number of phones to the number of phonemes in the language.

Lexical indicators require an ancillary procedure yielding a lexicon. Martin et al. approximated a lexicon by a list of frequent n-grams. Here, the lexicon is induced from the output of an explicit word segmentation model, viz. Venkataraman's incremental (2001) model, using the unsegmented phonetic corpora as the input. Though, obviously, infants can not access it, we use the lexicon derived from the CHILDES orthographic transcripts for reference.

4 Indicators' discriminant power

As the aforementioned indicators have been evaluated using various languages, allophonic grammars and measures, we present a unified evaluation, conducted using Sing et al.'s (2005) ROCR package.

4.1 Evaluation

Non-Boolean indicators require a threshold at and above which pairs are classified as allophonic. We evaluate indicators across all possible discrimination thresholds, reporting the area under the ROC curve (henceforth AUC). Equivalent to Martin et al.'s ρ, values lie in $[0, 1]$ and are equal to the probability that a randomly drawn allophonic pair will score higher than a randomly drawn non-allophonic pair; .5 thus indicates random prediction.

Moreover, we evaluate indicators' misclassifications at the discrimination threshold maximizing Matthews' (1975) correlation coefficient: let α, β, γ and δ be, respectively, the number of false positives, false negatives, true positives and true negatives, MCC $= (\gamma\delta - \alpha\beta)/\sqrt{(\alpha+\gamma)(\beta+\gamma)(\alpha+\delta)(\beta+\delta)}$. Values of 1, 0 and -1 indicate perfect, random and inverse prediction, respectively. This coefficient is more appropriate than the accuracy or the F-measure

[3] As for the actual computations, we use the same definitions as Le Calvez (2007) except that, as BC increases when distributions overlap and $0 \leq$ BC ≤ 1, we actually use $1-$BC.

[4] Because all allophonic rules implemented in the corpora are of the type $p \rightarrow a \ / \ \_ \ c$, FL and FL* only look for words minimally differing by their last segments.

when, as here, the true classes have very different sizes.[5] Using this optimal, MCC-maximizing threshold, we report the maximal MCC and, as percentages, the accuracy (Acc), the false positive rate (FPR) and the false negative rate (FNR).

4.2 Results and discussion

Indicators' AUC corroborate previous results for distributional indicators: they perform almost identically and do not accommodate high allophonic complexities at which they perform below chance (Figure 1.a) because, as suggested by Martin et al., every segment has an extremely narrow distribution and complementary distribution is the rule rather than the exception. By contrast, all three lexical indicators are much more robust even if, as predicted, FL's coarseness impedes its discriminant power (Figure 1.b).[6] The reason why FL* outperforms HFL may be due to the very definition of HFL's broken typewriter function: as the segments, e.g. $\{x, y\}$, are collapsed into a single symbol, the indicator captures not only minimal alternations like $wx \sim wy$, but also word pairs such as $xy \sim yx$.

AUC curves suggest that, for each type, indicators converge at medium allophonic complexity. Thus, misclassification scores are reported in Table 1 only at low (2 allophones/phoneme) and medium (9) complexities. Previous observations are confirmed by MCC and accuracy values: though all indicators are positively correlated with the underlying allophonic relation, correlation is stronger for lexical indicators. Surprisingly, zero FPR values are observed for some lexical indicators, meaning that they make no false alarms and, as a consequence, that all errors are caused by missed allophonic pairs.

5 Indicators' redundancy

None of the indicators we benchmarked in the previous section makes a perfect discrimination between allophonic and non-allophonic pairs of segments.

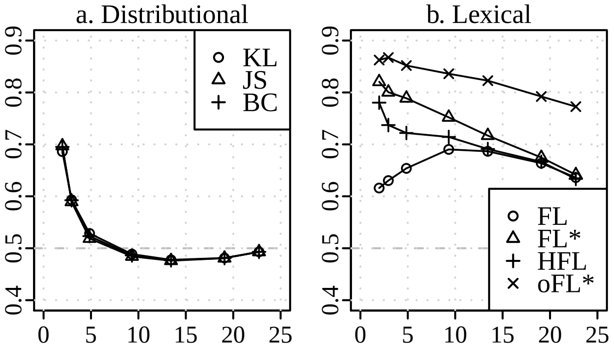

Figure 1: Indicators' AUC as a function of allophonic complexity. The dashed line indicates random prediction.

	2 allophones/phoneme				9 allophones/phoneme			
	MCC	Acc	FPR	FNR	MCC	Acc	FPR	FNR
KL	.095	*88.2*	11.3	58.5	.017	*90.7*	07.8	88.8
JS	.097	*86.4*	13.1	53.7	.014	*93.3*	05.1	93.0
BC	.097	*86.8*	12.8	54.4	.016	*89.9*	08.6	88.1
FL	.048	*37.3*	63.2	**13.6**	.116	*73.1*	26.8	**35.2**
FL*	**.564**	**99.3**	**00.0**	67.3	**.563**	**98.6**	**00.4**	53.0
HFL	.301	99.1	**00.0**	87.8	.125	*94.1*	04.5	78.7

Table 1: Indicators' performance at low and medium complexities, using the MCC-maximizing thresholds. Boldface indicates the best value. Italics indicate accuracies below that of a dummy indicator rejecting all pairs.

Yet, if some segment pairs are misclassified by one but not all (types of) indicators, a suitable combination should outperform individual indicators. In other words, combining indicators may yield better results only if, individually, indicators capture different subsets of the underlying allophonic relation.

5.1 Evaluation

To get a straightforward estimation of redundancy, we compute the Jaccard index between each indicator's set of misclassified pairs: let D and L be sets containing, respectively, a distributional and a lexical indicator's errors, $J(D, L) = |D \cap L| / |D \cup L|$. Values lie in $[0, 1]$ and the lower the index, the more promising the combination. To distinguish false positives from false negatives, we compute two Jaccard indices for each possible combination.

5.2 Results and discussion

Jaccard indices, reported in Table 2, emphasize the distinction between false positives and false negatives. False negatives have rather high indices: most

[5] If p phonemes have on average a allophones, out of the $pa(pa-1)/2$ possible pairs, only $pa(a-1)/2$ are allophonic, and a dummy indicator that rejects all pairs achieves a constant accuracy of $1-(a-1)/(pa-1)$, which is greater than 98% for any of our corpora. Besides, the computation of precision, recall and the F-measure do not take true negatives into account.

[6] These indicators perform similarly using the orthographic lexicon: we only report AUC for FL* (referred to as oFL*), as it gives the upper bound on lexical indicators' performance.

allophonic pairs that are not captured by distributional indicators are not captured either by lexical indicators, and *vice versa*. By contrast, there is little or no redundancy in false positives, even at medium allophonic complexity: though random pairs can be incorrectly classified as allophonic, the error is unlikely to recur across all types of indicators.

It is also worth noting that though JS performs slightly better than KL and BC, the exact nature of the distributional indicator seems to have little influence on the performance of the combination.

6 Combining indicators

As distributional and lexical indicators are not completely redundant, combining them is a natural extension. However, not all conceivable combination schemes are appropriate for our task. We present our choices in terms of Marr's (1982) levels of analysis.

At the computational level, a combination scheme can be either disjunctive or conjunctive, i.e. each indicator can be either sufficient or (only) necessary. Aforementioned indicators were designed as necessary but not sufficient correlates of phonemehood. For instance, while a phoneme's allophones have complementary distributions, not all segments that have complementary distributions are allophones of a single phoneme. Therefore, we favor a conjunctive scheme,[7] even if this conflicts with abovementioned results: most errors are due to missed allophonic pairs but a conjunctive scheme, where every indicator must be satisfied, is likely to increase misses.

At the algorithmic level, a combination scheme can be either logical or numerical. A logical scheme uses a logical connective to join indicators' Boolean decisions, typically by conjunction according to our previous decision. By contrast, a numerical scheme tries to approximate interactions between indicators' values, merging them using any monotone increasing function; discrimination then relies on a single threshold. In practical terms, we use multiplication as a numerical counterpart of conjunction.

6.1 Evaluation

Setting aside the following minor adjustments, we use the same protocol as for individual indicators.

		2 allo./phon.		9 allo./phon.	
		FP	FN	FP	FN
KL	FL	.096	**.071**	.113	.359
JS	FL	.113	.076	.071	**.355**
BC	FL	.110	.075	.118	.358
KL	FL*	**.000**	.595	**.008**	.520
JS	FL*	**.000**	.548	**.005**	.525
BC	FL*	**.000**	.556	**.007**	.517
KL	HFL	**.000**	.667	.087	.788
JS	HFL	**.000**	.612	.033	.781
BC	HFL	**.000**	.620	.089	.787

Table 2: Indicators' redundancy at low and medium allophonic complexities, estimated by the Jaccard indices between their false positives (FP) and false negatives (FN). Boldface indicates the best value.

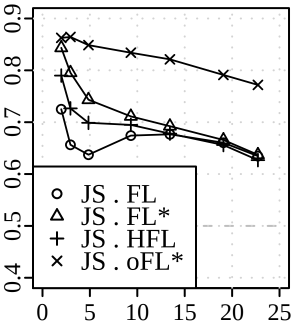

Figure 2: Indicators' AUC as a function of allophonic complexity, for the multiplicative combination scheme. The dashed line indicates random prediction.

Logical combinations require one discrimination threshold per combined indicator. As it facilitates comparison with previous results, we report performance at the thresholds maximizing the MCC of individual indicators (rather than at the thresholds maximizing the combined MCC) .

Numerical combinations are sensitive to differences in indicators' magnitudes. Equal contribution of all indicators may or may not be a desirable property, but in the absence of *a priori* knowledge of indicators' relative weights, each indicator's values were standardized so that they lie in $[0, 1]$, shifting the minimum to zero and rescaling by the range.

6.2 Results and discussion

It is worth noting that, while the performance of combined indicators is still good (Table 3), it is less satisfactory than that of the best individual indicators. Moreover, even if misclassification scores

[7] This generalizes Martin et al.'s attempt at combination: they used FL as a high-pass lexical filter prior to the use of KL.

		Logical combination: conjunction								Numerical combination: multiplication							
		2 allophones/phoneme				9 allophones/phoneme				2 allophones/phoneme				9 allophones/phoneme			
		MCC	Acc	FPR	FNR	MCC	Acc	FPR	FNR	MCC	Acc	FPR	FNR	MCC	Acc	FPR	FNR
KL	FL	.104	*92.9*	06.5	67.3	.037	*94.7*	03.6	91.3	.104	*92.9*	06.5	67.3	.116	*73.1*	26.7	**35.2**
JS	FL	.109	*91.7*	07.8	**62.6**	.032	*96.2*	02.1	94.6	.110	*91.5*	07.9	**61.9**	.116	*73.1*	26.7	**35.2**
BC	FL	.109	*91.9*	07.5	63.3	.038	*94.5*	03.9	**90.8**	.109	*92.8*	06.6	66.0	.116	*73.1*	26.7	**35.2**
KL	FL*	.457	**99.2**	**00.0**	78.9	.207	**98.2**	00.1	93.3	.526	**99.3**	**00.0**	71.4	.371	**98.4**	**00.1**	81.6
JS	FL*	**.465**	**99.2**	**00.0**	78.2	.153	**98.2**	**00.0**	95.7	**.548**	**99.3**	**00.0**	66.0	**.393**	**98.4**	00.2	78.3
BC	FL*	**.465**	**99.2**	**00.0**	78.2	**.211**	**98.2**	00.1	93.0	.535	**99.3**	**00.0**	68.7	.388	**98.4**	**00.1**	79.0
KL	HFL	.348	99.1	**00.0**	87.8	.078	*97.0*	01.3	93.5	.363	99.1	**00.0**	84.4	.117	*90.3*	08.4	73.7
JS	HFL	.348	99.1	**00.0**	87.8	.068	*97.9*	00.3	96.5	.359	99.1	00.1	83.7	.119	*90.4*	08.4	73.9
BC	HFL	.348	99.1	**00.0**	87.8	.077	*96.9*	01.4	93.2	.361	99.1	**00.0**	85.7	.119	*90.3*	08.4	73.5

Table 3: Performance of combined distributional and lexical indicators, at low and medium allophonic complexity. Boldface indicates the best value. Italics indicate accuracies below that of a dummy indicator rejecting all pairs.

show that conjoined and multiplied indicators perform similarly, disparities emerge at medium allophonic complexity: while multiplication yields better MCC and FNR, conjunction yields better accuracy and FPR. In that regard, observing FPR values of zero is quite satisfactory from the point of view of language acquisition, as processing two segments as realizations of a single phoneme (while they are not) may lead to the confusion of true minimal pairs of words. Indeed, at a higher level, learning allophonic rules allows the infant to reduce the size of its emerging lexicon, factoring out allophonic realizations for each underlying word form.

Furthermore, AUC curves for the multiplicative scheme (Figure 2),[8] most notably FL's, suggest that distributional indicators' contribution to the combinations appears to be rather negative, except at very low allophonic complexities. One explanation (yet to be tested experimentally) would be that they come into play later in the learning process, once part of allophony has been reduced using other indicators.

7 Conclusion

We presented an evaluation of distributional and lexical indicators of allophony. Although they all perform well at low allophonic complexities, misclassifications increase, more or less seriously, when the average number of allophones per phoneme increases. We also presented a first evaluation of the combination of indicators, and found no significant difference between the two combination schemes we defined. Unfortunately, none of the combinations we tested outperforms individual indicators.

For comparability with previous studies, we only considered combination schemes requiring no modification in the definition of the task; however, learning allophonic pairs becomes unnatural when phonemes can have more than two realizations. Embedding each indicator's segment-to-segment (dis)similarities in a multidimensional space, for example, would enable the use of clustering techniques where minimally distant points would be analyzed as allophones of a single phoneme.

Thus far, segments have been nothing but abstract symbols and, for example, the task at hand is as hard for [a] ∼ [ḁ] as it is for [ɥ] ∼ [k]. However, not only do allophones of a given phoneme tend to be acoustically similar, but acoustic differences may be more salient and/or available earlier to the infant than complementary distributions or minimally differing words. Therefore, the main extension towards a comprehensive model of the acquisition of allophonic rules would be to include acoustic indicators.

Acknowledgments

This work was supported by a graduate fellowship from the French Ministry of Research. We thank Benoît Crabbé, Emmanuel Dupoux and Sharon Peperkamp for helpful comments and discussion.

[8] We do not report a threshold-free evaluation for the logical scheme. As it requires the estimation of the volume under a surface, comparison between schemes becomes difficult. Moreover, as the exact definition of the distributional indicator does not affect the results, we only plot combinations with JS.

References

Luc Boruta, Sharon Peperkamp, Benoît Crabbé, and Emmanuel Dupoux. Submitted. Testing the robustness of online word segmentation: effects of linguistic diversity and phonetic variation.

Luc Boruta. 2011. A note on the generation of allophonic rules. Technical Report 0401, INRIA.

Michael R. Brent and Timothy A. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

Anthony C. C. Coolen, Reimer Kühn, and Peter Sollich. 2005. *Theory of Neural Information Processing Systems*. Oxford University Press.

Isabelle Dautriche. 2009. Modélisation des processus d'acquisition du langage par des méthodes statistiques. Master's thesis, INSA, Toulouse.

Brian Dillon, Ewan Dunbar, and William Idsardi. In preparation. A single stage approach to learning phonological categories: insights from inuktitut.

Charles Hockett. 1955. A manual of phonology. *International Journal of American Linguistics*, 21(4).

Rozenn Le Calvez. 2007. *Approche computationnelle de l'acquisition précoce des phonèmes*. Ph.D. thesis, UPMC, Paris.

David Marr. 1982. *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman.

Andrew T. Martin, Sharon Peperkamp, and Emmanuel Dupoux. Submitted. Learning phonemes with a pseudo-lexicon.

Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta, Protein Structure*, 405(2):442–451.

Jessica Maye, Janet F. Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):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.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941.

Anand Venkataraman. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):351–372.