In press: Proceedings of the 33rd Annual Cognitive Science Society, 2011

Templatic features for modeling phoneme acquisition

Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
Ecole des Hautes Etudes en Sciences Sociales, Ecole Normale Supérieure
29 rue d'Ulm, 75005 Paris, France
Guillaume Beraud-Sudreau (guillaumeberaud@gmail.com)
Laboratoire de Sciences Cognitives et Psycholinguistique, Centre National de la Recherche Scientifique
29 rue d'Ulm, 75005 Paris, France
Shigeki Sagayama (sagayama@hil.t.u-tokyo.ac.jp)
Departement of Information Physics and Computing
University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656 Japan

Abstract

We describe a model for the coding of speech sounds into a high dimensional space. This code is obtained by computing the similarity between speech sounds and stored syllable-sized templates. We show that this code yields a better linear separation of phonemes than the standard MFCC code. Additional experiments show that the code is tuned to a particular language, and is able to use temporal cues for the purpose of phoneme recognition. Optimal templates seem to correspond to chunks of speech of around 120ms containing transitions between phonemes or syllables.

Keywords: Early language acquisition, modeling, phonemes

Introduction

Infants spontaneously learn their ambient language at an amazing speed. During their first year of life, they construct abstract perceptual categories corresponding to the phonemes of their language. They lose the ability to distinguish fine phonetic variants that belong to the same phoneme category, and enhance their ability to distinguish between category contrasts (see a review in Kuhl, 2000). This is done without any supervision from the parents, before a substantial recognition lexicon has been built (12-month-olds are believed to recognize about 100 words), and before they can articulate correctly the phoneme categories they recognize. How do infants achieve this? One possibility is that they perform some kind of unsupervised statistical clustering of the ambient speech signals. Maye, Werker and Gerken (2002) showed that 6-month-old infants perform such computations, using artificial languages with either a monomodal statistical distribution or a bimodal distribution of phonetic cues.

Only a limited number of studies have addressed the computational mechanisms that could underlie such acquisitions. Guenther & Gjaja (1996) showed that Self-Organizing Maps have the potential to construct phoneme categories in an unsupervised fashion (see also Valhabba et al., 2007; Gauthier, Shi & Xu, 2007). Valhabba et al (2007) implemented an incremental version of Expectation Maximization on a Gaussian Mixture model, and showed that both the number of vowels and their statistical distributions can be inferred from the signal in an unsupervised fashion.

These studies, however, did not use raw speech signals, but rather a small number of parameters extracted by hand: e.g., the frequency of the first and second formants, vowel duration, etc. (Valhabba et al 2007). This presupposes that infants are equipped with fairly speech-specific perceptual abilities and, crucially, that they know how to segment the continuous stream into discrete segments like consonants, or vowels. This latter assumption is problematic given that such segmentation is not universal, but depends on the phonology of the language (Dupoux et al, 1999).

Varadarajan et al. (2008) is one of the few published paper that attempted to learn phonemes from raw speech. Using an optimized version of Successive State Splitting (SSS, Takami & Sagayama, 1992), they grew in an unsupervised fashion a large network of Hidden Markov Model (HMM) states. These states were shown to encode speech sounds with no loss of information compared to supervized HMMs, but there were two problems. First, the states of the HMM network did not correspond to phonemes, but rather to subphonemic units of the size of acoustic events (e.g. burst, closure, transition, etc). This is the *oversegmentation* problem. Second, even combining states into sequences did not yield phonemes, but rather, context dependant variants (contextual allophones). This is the *contextual variability* problem. Here, we address the first problem, the second problem being address in other work (Peperkamp et al, 2006; Martin et al. submitted; Boruta et al. 2011).

The oversegmentation problem of SSS, although disappointing, is not entirely surprising. State-of-the art supervised HMMs have the same problem: segments are typically modeled using three states, not a single state. The reason is that HMMs represent speech as local spectral feature vectors (e.g. Mel-Frequency Cepstral Coeficients – MFCC, computed over a 15-20ms window), whereas phonemes are realized as a complex articulatory trajectory spanning between 50 and 150ms, sometimes involving a sequence of events (constriction, release, changes in the source, etc.). Since HMMs are modeling speech sounds through Gaussians distributions (which are local), the only way to model phonemes accurately is to segment them into subparts. This problem is not limited to MFCC features, but would also apply to any local feature, like for instance

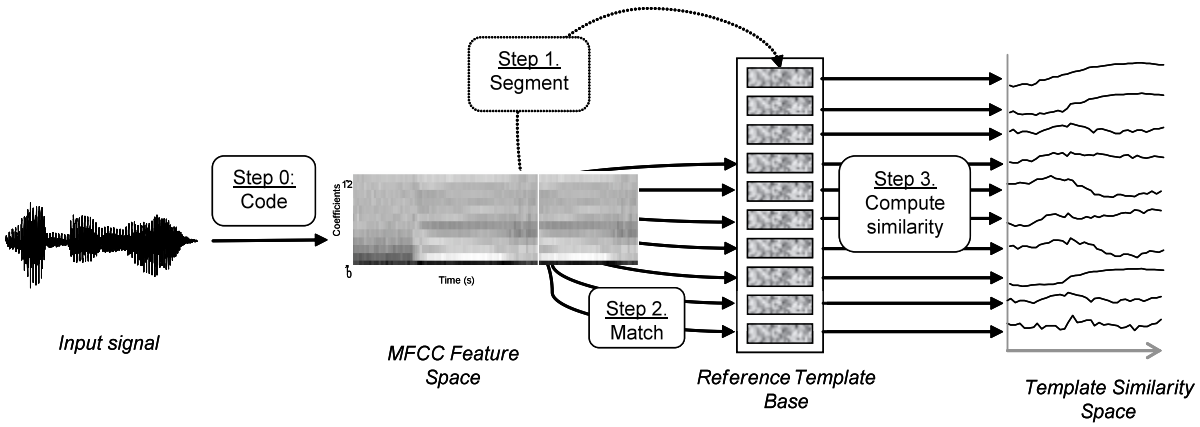

Figure 1. Outline of the High Dimensional Template Matching model. It is composed of an instance-based bank of reference templates, and has two processing modes: during early language experience (dotted lines), the templates are segmented out of the speech stream, during subsequent development and in adults, the signal is matched to the bank of templates (solid lines).

wavelet type functions (Smith & Lewicki, 2006), or features derived from auditory models (Chi et al, 2005).

To solve the oversegmentation problem, we propose to explore the feasibility of replacing low dimensional, low-level features with high dimensional, holistic or coarse grained features. We review some existing proposals.

Holistic/templatic features

Research into the human visual system has revealed that the brain analyzes shapes and objects in a series of hierarchical stages in which stimulus features of increasing complexity and size are extracted. Ullman et al. (2002) argued that the maximally informative features for the purpose of object classification are not local features, but rather features of intermediate complexity, that correspond to *fragments* of images or objects. The brain would store such fragments which form a high dimensional code adapted to a particular domain of object perception. Along a similar line, Edelman (1996) proposed that the brain represents a shape through its similarity to a number of reference shapes, that are stored as patterns of elementary features. Familiar and novel objects are then represented as points in a shape space computed from similarities to a set of reference objects.

Such proposals are only starting to be applied to speech. Liquid State Machines (Maas et al, 2002), or Echo State Systems (Jaeger, 2002) use recurrent networks or dynamic systems to recode a time-varying low dimensional signal into a high dimensional one which incorporates information spanning the recent past of the system. Such codes are more robust to noise than low-level featural approaches (Skowronski & Harris, 2007), however it is unclear how to optimize such representations. Coath & Denham (2005), proposed a model storing templates consisting of 100ms-200ms speech sounds, which are used as convolution filters. They argue that the high dimensional code obtained is more robust to variation due to time compression and speaker variation than classical features. Dupoux (2004) has proposed a similar approach based on the psycholinguistics of human infants, whereby processing is based on the segmenting and storing of syllable-sized templates, which are the basis for discovering the smaller and more abstract phonemes, which can in turn be used to recover the even more abstract linguistic features. In this paper, we explore a quantitative assessment of this last approach.

The algorithm

The idea of using examples from the problem set as the basis for representing further examples is at the core of Support Vector Machine models (Cortes & Vapnik, 1995). The present proposal is inspired by the idea that large units like syllables are natural perceptual units for infants and adults. For instance, Bertoncini & Mehler (1981) showed that neonates can count the number of syllables in a speech stream, before they have learned the phonemes of their language. The proposal is that, during their first year of life, infants build a large base of syllable-like templates, and at a later stage, compute the similarity between the incoming signal and the stored templates. The High Dimensional Template Matching model (HD-TMatch) presented in Figure 1 assumes that all sounds (templates and signal) are first coded in terms of low level features (Step 0). During the early acquisition phase, (Step 1), the model segments out chunk of speech of a given size and stores them as templates in an instance-based memory system. After the templates become fixed, speech sounds are matched to the templates (Step 2), and a similarity between each template and the signal is computed (Step 3). This model translates a time varying trajectory in acoustic space into a point in similarity space. As such, it has the potential to solve part of the oversegmentation problem since it matches whole trajectories instead of just a slice of time. It also has the potential to address convergence towards the native sounds since the stored templates belong to the native language. Note that the model is not committed to templates being exactly aligned to linguistically defined syllables; they could as well correspond to diphones, triphones, or acoustic chunks of syllable size.

- Step 0: Coding. The input was coded in terms of a frame every 5 ms consisting of 13 MFCC coefficients (Mermelstein, 1976) computed on overlapping 15 ms windows.
- Step 1. Templates Segmentation. The template base can vary according to three independent parameters: (a) *number of templates*. To be effective, templates have to be numerous enough to cover the range of possible sound combinations in the language. However, too many templates may hamper learning. (b) *template duration*: a template has to be long enough to contain significant dynamic properties, but not too long, otherwise the number of templates required for total language coverage explodes. (c) *template boundaries*: Templates can either be temporally aligned to structural properties of speech (syllable boundaries, peak of vowel nucleus, etc), or randomly segmented. Even though these 3 variables may interact, in the present study, we manipulate one while keeping the other two constant in artificial languages.
- Step 2. Template Matching. In the model, each template is matched to the signal in a parallel and independent way, as if each template were as an autonomous recognizer, looping through the signal in the attempt to recognizing itself. We use Dynamic Time Warping (DTW) (Myers and Rabiner, 1981) to find the optimal alignment of the template and the signal, hence obtaining a warping function for each template. Multiple passes through the templates are allowed if the signal is long enough.
- Step 3. Similarity. The warping function is used to extract two different types of signal: spectral similarity, and temporal distortion. Spectral similarity is based on the readout of the Euclidian distance between the MFCC coefficients of the signal and the aligned template for each frame (see Appendix). Temporal distortion is based on the local slope of the warping function: any deviation from a slope of 1 in the warping function is giving a cost in the temporal distortion between the template and signal. The output representation for a bank of N templates is hence a set of 2N time series, sampled every 5 ms.

Methodology

The aim of this paper is to compare the efficiency of templatic features compared to low-level ones for the purpose of phoneme classification. We assessed this using a linear separation test: a perceptron was trained on a set of labeled examples using the RPROP algorithm (Riedmiller, and Braun, 1993), and the performance of the classification was measured both on the training set and on a novel generalization set. Training and recognition of the phonetic categories was computed frame by frame using human labels, and the error rate was the percentage of misclassified frames over the training or generalization sets. This performance was compared to the results obtained with two baseline low-level featural codes. One is the raw MFCC (13 dimensions) used as input to the model. The second baseline is MFCC + Delta2 code, which corresponds to MFCC coefficients plus their first and second order derivatives (39 dimensions). This is a useful comparison since time derivatives of the MFCC coefficients are a standard way to improve on local featural codes to capture some of the dynamic properties of speech.

The algorithms were tested on two pseudo-languages, which we constructed with carefully balanced phoneme and syllable sets. Utterances of each pseudo-language were recorded by a male talker in a quiet and non reverberating environment, digitized, and converted into MFCC coefficients. All stimuli were hand labeled for the purpose of performing the linear separation test. The stimuli were then distributed into three sets. The first set was used to generate the templates, the second one for training the perceptron and a third one for generalization. Each test was performed 10 times, with a different random assignment of sets, in order to derive standard deviations for the error rates.

The *Monosyllabic language* contained 6 vowels /a e i u o y / and 6 consonants /R m s p t k/. These phonemes were combined to create 36 Consonant-Vowels (CV) syllables. The syllables were pronounced in isolation (as if they were monosyllabic words). Each syllable was recorded 54 times. The template set contained between 4 and 12 exemplars of each syllables, the training set contained 34 exemplars and the generalization set contained 8 exemplars.

The *Polysyllabic language* contained trisyllabic CVCV-CV words, composed of 8 phonemes /R ʃ d m e a i u/. These phonemes are arranged following the same CV structure than in the previous sets. The set was built in such a way that all the phonemes consonants or vowels were produced the same number of times, in every position. The template set contained 12 exemplars of each syllables (192 templates), the training set contained 34 exemplars and the generalization set contained 8 exemplars.

Results

Assessing the templatic code

We used the monosyllabic language for these experiments. The linear classification performance of MFCC and MFCC+Delta2 are used as baseline (Figure 2a). As seen in Figure 2b, template features using whole syllables as templates yields systematically better phoneme classification performance than the baselines. This shows that templatic features are both more informative than the MFCCs on which it is based, and outperform the MFCC time derivatives. Increasing the number of templates from 4 per syllable types to 12 per syllable types improves slightly the performance, but as the overall dimensionality grows from 144 to 432 dimensions, one can start to see evidence of over-fitting (i.e. a growing gap between training and generalization). Adding time distortion coding increases more the performance than adding more templates, suggesting that the temporal distortion adds a new and useful type of information. This is interesting, because temporal alignment parameters are typically thrown away in classical speech recognition systems (more on this below). In Figure 2c, we show that the gain in performance obtained by template coding is not due to high dimensionality alone.

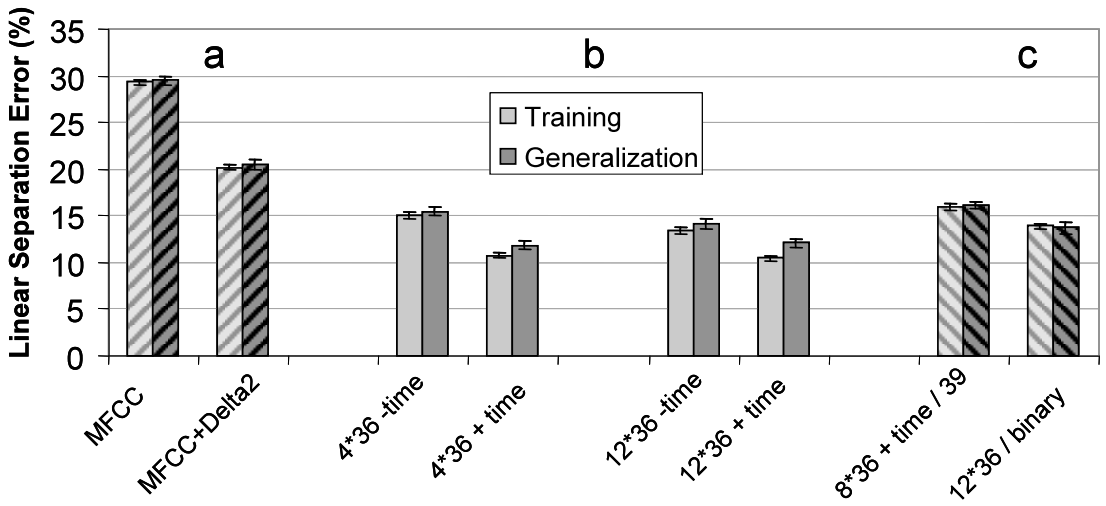

Figure 2. Percent error in a phoneme classification test using linear separation, for the training and the generalization sets, as a function of type of input code, using the Monosyllabic Language. Bars show one standard deviation over and above the mean. a. Baseline scores for the MFCC (12 dimensions) and MFCC+Delta2 (39 dimensions) codes. b. Scores for templatic codes. We used as templates 4 exemplars of each of the 36 syllable types (left) or 12 exemplars (right). The '-time' bars show the scores with spectral similarity only and the '+time' shows the score where time distortion has been added. c. Scores for compressed templatic codes. We projected a code using 8*36 templates (spectral+temporal) onto the first 39 principal components (left). We quantizated the spectral similarity of a code with 12*36 templates onto a binary code (0 or 1) (right).

Indeed, projecting the code obtained with 8 templates per syllable type (plus time distortion) onto the first 39 PCA dimensions still yields better performance than MFCC+Delta 2 despite the fact that the number of dimensions is the same. Finally, we quantized each dimension onto a binary code by using a threshold set at one standard deviation above the mean (the means and standard deviations are computed across the dimensions, for each time frame separately). This was done for a code using 12 templates per syllable, and the result was undistinguishable from that obtained using non quantized version, suggesting that the high dimension templatic code is intrinsically a (sparse) binary code.

Language specificity

How well does the template code capture language-specific properties? If this code was only increasing performance because of its high dimensionality, the particular set of templates used should be irrelevant. Here, we split the monosyllabic language into two disjoint sublanguages. The "easy" sublanguage used the maximally distinct consonants /R m s/ and vowels /a e i/. The "hard" sublanguage used the minimally distinct consonants /p t k/ and vowels /o u y/. Each sublanguage had only 9 syllable types. We used as a template set the syllables from one sublanguage, and tested either on new exemplars of the same language (appropriate templates) or exemplars from the other language (inappropriate templates). As shown in Table 1, using the inappropriate language for the template set yields a large drop in performance, and this both for the easy and hard sublanguage. An ANOVA ran across 10 simulations on the log probability of error for the generalization set showed a significant effect of sublanguage ($F(1,36)=537$, $p<.0001$), and appropriateness ($F(1,36)=410$, $p<.0001$), but no interaction between these two factors ($p>.05$). Appropriate templates were better than the MFCC+delta2 baseline ($F(1,36)=266$; $p<.001$), and inappropriate templates were worse score than baseline ($F(1,36)=41$, $p<.0001$). In brief, template features are optimally tuned to the language from which they are extracted; they are very good for the segments that belong to that language, and poor for 'foreign' segments. This mimicks the tuning process to native sounds which take place during early language acquisition (Kuhl, 2000).

Table 1. Percent error in phoneme classification (and standard error across simulations) in two sublanguages, easy and hard, as a function of the code used to represent the signals: the language-independent MFCC+Delta 2 code, and the templatic codes based on the appropriate or inappropriate sublanguage.

Code	Easy Language		Hard Language	
Baseline	Training	General.	Training	General.
MFCC + Delta 2	8.1% (0.3)	8.9% (0.7)	11.8% (0.3)	12.9% (0.7)
Appropriate Templates	4.8% (0.2)	5.2% (0.5)	8.2% (0.5)	10.2% (0.7)
Inappropriate Templates	8.0% (0.2)	9.4% (1.0)	14.0% (0.7)	16.6% (1.0)

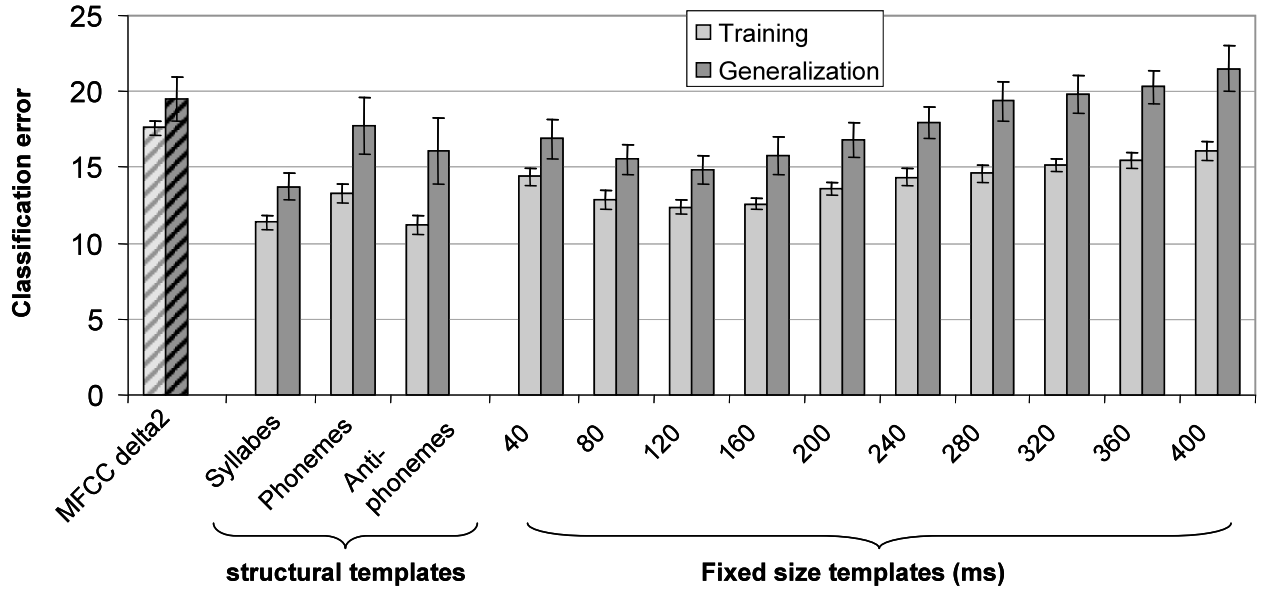

Figure 4. Effect of the size of the template on phoneme classification using a linear separation test. The language used was the Polysyllabic Language. The number of templates is fixed in all simulations.

Temporal cues and vowel duration

We found above that temporal distortion was useful even in an artificial language in which duration cues a priori carries little linguistic information. The usefulness of temporal distortion should be even more apparent in a language where such cues are used, like in Japanese, where vowel length is contrastive. We introduced a contrast in vowel duration in the "easy" sublanguage. It had 6 vowels /a e i a: e: i:/, the latter three being obtained by doubling the duration of the vowels in the original recording using pitch synchronous resynthesis. The results are shown in Table 2; evidently, template codes, especially with time distortion fare better than the MFCC controls.

Table 2. Percent error in phoneme classification (and standard error across simulations) in a language using contrastive vowel duration. The score is given for the training and generalization sets, and for a specific short versus long vowel contrast (generalization only).

Code	Training	Gener.
MFCC	32.3% (0.3)	32.5% (0.9)
MFCC + Delta 2	26.0% (0.4)	26.1% (0.8)
8*36 – time	20.3% (0.3)	23.9% (0.8)
8*36 + time	11.9% (0.2)	15.1% (0.7)

Size and nature of the templates

What is the optimal size of the templates? We used the *polysyllabic* language and tested structurally defined templates, syllables, phonemes and antiphonemes, segmented using human labels. Antiphonemes were defined as the final 50% part of one phoneme followed by the initial 50% part of the next. As shown in figure 4, syllabic templates yielded the best performance, but somewhat counter-intuitively, antiphonemes were better than phonemes. Second, we tested randomly segmented templates of a fixed duration. We found that randomly segmented templates can do almost as well as syllables, as long as they have a duration of around 120ms. This duration corresponds to a unit whose size is intermediate between syllables and phonemes. These two findings are compatible with the hypothesis that templates are optimal when they capture the *transition* parts between phonemes. The 120ms is also compatible with the optimal unit found by Coath and Denham (2005).

Conclusion

We have found that coding the speech signal in a high dimensional space of template similarity yields a significant improvement over standard MFCC features, even when temporal derivatives are used. In addition, time distortion derived from the DWT alignment process adds useful information over and above spectral similarity. This is especially true when the language makes use of contrastive durational cues. We found that the improvement of the templatic code is limited to the particular language used to make up the template sets. Templates of one language are ill suited to classify phonemes belonging to a different language. Finally, optimal templates seem to correspond to units around 100-200ms, containing at least the transitions between two adjacent phonemes.

Of course, all of these conclusions are limited by the experimental approach we used, which is to test our system on miniature languages, with restricted phoneme and syllable inventories. It remains to be shown whether such coding and conclusions scale up to real-sized languages, more coarticulated inputs such as spontaneous speech, and psychologically realistic learning procedures, such as incremental unsupervised clustering. Another point worth mentioning is that, because of the multiple DTWs, the complexity of the algorithm is in $o(n.l^2)$, where n is the

number of templates, and l is the utterance length. More work remains to be done to optimize this algorithm. Moreover, the usability of the code is limited by tractability issues regarding clustering algorithms in high dimensions. Potentially useful is the fact that templatic features can be reduced to binary vectors at little or no cost.

Overall, this supports the interest of coarse graine features for modeling speech perception (Coath & Denham, 2005; Skowronski & Harris, 2007), but more research is needed to add biological constraints to such models and derive new predictions for early language acquisition.

Acknowledgments

We thank Sharon Peperkamp, Peter Dayan and Paul Smolensky for very useful discussion and comments.

Appendix

A given stimulus S is aligned to a template T using DTW, and the two time axes are related through the warping function warp(t). We can read out DS,T(t), the Euclidian distance between the signal and warped template:

$$D_{S,T}(t)=\sqrt{\sum_i (s_i(t)-t_i(warp_{S,T}(t)))^2}$$

where $s_i(t)$and $t_i(t)$ are the MFCC coefficients of S and T at time t, respectively. We then define Sim(t), a time-dependant measure of template similarity:

$$\mathrm{Sim}_{S,T}(t)=\frac{1}{\alpha+D_{S,T}(t)}$$

where α is a constant used to avoid infinite values for a distance of zero ($\alpha = 10^{-3}$.). Finally, we define a time-dependant measure of temporal distortion:

$$\text{Dist-temp}_{S,T}(t)=\left|\log\left(warp'_{S,T}(t)\right)\right|$$

where warp'(t) is the smoothed slope of the warping function at time t, computed with a regression on 5 adjacent frames, and truncated to fit the interval $[10^{-3}, 10^{+3}]$.

References

Boruta, L., Peperkamp, S., Crabbé, B., & Dupoux, E. (2011). Testing the robustness of online word segmentation: effects of linguistic diversity and phonetic variation. Proceedings of CMCL, ACL, Portland, Oregon.

Chi, T., Ru, P., & Shamma, S.A. (2005). Multiresolution spectrotemporal analysis of complex sounds, *JASA*, 118, 887-906.

Coath M., & Denham, S.L., (2005). Robust sounds classifications through the representations of similarity using response fields derived from stimuli during early experience, *Biological Cybernetics*, 93(1), 22-30.

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks, *Machine Learning*, 20.

de Boer, B. & Kuhl, P.K. (2003) Investigating the role of infant-directed speech with a computer model. *ARLO* 4, 129–134.

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.

Dupoux, E., (2004). The Acquisition of Discrete Segmental Categories: Data and Model In *Proceedings of the 18th International Congress of Acoustics*, Kyoto, April 4-9.

Edelman, S. (1998). Representation is Representation of Similarities, *Behavioral and Brain Sciences* 21,449-498.

Gauthier, B., Shi, R., & Xu, Y., (2007). Learning phonetic categories by tracking movements, *Cognition,*103, 80-106

Guenther, F.H., & Gjaja, M.N. (1996). The perceptual magnet effect as an emergent property of neural map formation, *JASA*, 100(2), 1111–1121.

Jaeger, H. (2002). Adaptive nonlinear system identification with echo state networks. In S.T.S. Becker, & K. Obermayer (Eds.), *Advances in neural information processing systems* (pp. 593–600). Cambridge: MIT Press.

Kuhl, P.K. (2000). A new view of language acquisition. *PNAS*, 97, 11 850–11 857.

Maass, W., Natschlager, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14, 2531–2560.

Maye, J., Werker, J.F., & Gerken L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), B101-B111.

Mehler, J. & Bertoncini, J. (1981). Syllables as Units in Infant Perception, *Infant Behav. and Devel.*, 4, 271-284.

Myers, C.S., & Rabiner, L. R. (1981). A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 607, 1389-1409.

Peperkamp, S., Le Calvez, R., Nadal, J.P. & Dupoux, E. (2006). The acquisition of allophonic rules. *Cognition*, 101, B31-B41.

Riedmiller, M. & Braun, H. (1993), A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm, Proceedings of the IEEE International Conference on Neural Networks, San Francisco: IEEE.

Skowronski, M.D. & Harris, J.G. (2007). Automatic speech recognition using a predictive echo state network classifier. *Neural Networks*, 20, 414–423.

Smith E.C., Lewicki, M.S., (2006). Efficient auditory coding, *Nature,* 439, 978–982.

Takami, J. & Sagayama, S. (1992). A successive state splitting algorithm. In *ICASSP*, 66.6, 573–576.

Ullman, S., Vidal-Naquet, M. & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5(7), 682-687.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. & Amano, S. (2007) Unsupervised learning of vowel categories from infant-directed speech, *PNAS*, 104:33, 13273–13278.

Varadarajan, B., Khudanpur, S., & Dupoux, E. (2008). Unsupervised Learning of Acoustic Subword Units. In *Proceedings of ACL-08: HLT*, 165-168.

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