Fully Unsupervised Word Learning from Continuous Speech Using Transitional Probabilities of Atomic Acoustic Events

Okko Johannes Räsänen[1]

[1] Dept. Signal Processing and Acoustics, Aalto University School of Science and Technology, Finland
okko.rasanen@tkk.fi

Abstract

This work presents a learning algorithm based on transitional probabilities of atomic acoustic events (vector quantized spectral features). The algorithm learns models for word-like units in speech without any supervision, and without a priori knowledge of phonemic or linguistic units. The learned models can be used to segment novel utterances into word-like units, supporting the theory that transitional probabilities of acoustic events could work as a bootstrapping mechanism of language learning. The performance of the algorithm is evaluated using a corpus of Finnish infant-directed speech.

Index Terms: unsupervised learning, distributional learning, language acquisition, word segmentation

1. Introduction

Human children face the complexity of spoken language when they are starting to learn their native language during the first year of their life. Since spoken language consists mainly of continuous acoustic signals without pauses between words, segmentation and thereby learning of words is a difficult task without pre-existing knowledge of the language. Although some cues to word boundaries exist, they are mainly language specific. Therefore the word segmentation capabilities of human infants cannot be explained solely by innate word segmentation mechanisms, but a learning aspect has to be included. One of the most widely studied cues for word segmentation are the transitional probabilities (TPs) of subsequent speech sounds and the closely related phonotactic rules (e.g., [1-3]).

The idea behind the transitional probability analysis is that the probability of transition from one acoustic unit to another is higher inside a linguistically relevant pattern such as word than in the transitions across two patterns. Studies have shown that infants as young as 8 months of age can learn TPs of syllables in an artificial language and use these statistical dependencies to segment continuous speech stream into word like units [1]. Further experiments support the idea that the learned word-like-unit structures act as lexical candidates if they are presented in a proper linguistic context [2]. Lately, Pelucchi et al. [3] have shown that infants are able to use transitional probabilities also in real speech spoken in a foreign language, and that they also take into account backward probabilities of speech sounds. This suggests that knowledge of the phonemic or syllabic system of a language is not a necessity for distributional learning.

Based on these findings, it can be hypothesized that infants might bootstrap their word segmentation process by analyzing regularly recurring stretches of acoustic signals (that can be modeled with TPs between atomic acoustic units) without pre-existing phonemic knowledge (phones, syllables). These recurring segments act as preliminary lexical items that can be associated to multimodal/motor representations (functional aspect) and analyzed in further detail to facilitate speech perception (developmental aspect; see, e.g., PRIMIR-theory of language acquisition by Werker & Curtin [4]).

If, however, the TP framework were to be considered as a feasible method for the bootstrapping of linguistic learning, the parallel existence of a computational mechanism that is able to demonstrate such processing would be convenient. As for computational models of word segmentation, in [5] and [6] it was shown that word segmentation is possible in a weakly supervised learning framework where the learning agent receives multimodal support from a visual scene. By associating recurring segments of speech to objects in the visual scene through cross-situational learning, the agent learns to parse keywords from the incoming utterances. However, this learning paradigm did not lead to learning of words that were not systematically related to objects in the environment. Instead, only the keywords that were present *in both* audio and as visual categories were learned and segmented properly.

In the current work, we have modified the previously used Concept Matrix (CM) algorithm to perform truly unsupervised acquisition of word models without multimodal support. The new algorithm will be referred to as *self-learning concept matrices* (SLCM). We show that the algorithm is capable of acquiring spectrotemporal representations of recurring word-like units from speech without any *a priori* knowledge of speech sounds or words, and that these representations inherently segment novel utterances into word-like units.

2. Methods

The SLCM algorithm originates from the CM algorithm [6] that is based on the analysis of transitional probabilities between elements $a_j \in [1, 2, \ldots, N_A]$ in a discrete sequence $X = [a_1, a_2, \ldots, a_n]$. Each model c (or *concept* in a multimodal case) consists of a set of matrices that model transition probabilities at different temporal distances, or lags, $\boldsymbol{k} = \{k_1, k_2, .., k_K\}$.

In the previous work [5,6], the set of models $\boldsymbol{c} \in C$ that were updated during perception of an utterance was defined by a bag of tags that represented contextual information. For example, an utterance "*What a nice green dog and ball*!" would be paired with visual tags [ball] and [dog]. This type of weak supervision automatically grounds the auditory words with contextual knowledge, but as a shortcoming, the contextual information determines directly how many internal models are needed and forces all content in the audio signal to be updated into these models. Therefore, e.g., silence and function words did not acquire their own models unless they were somehow presented in the contextual information source.

This problem can be overcome by disabling the use of external tags for $\boldsymbol{c}$, and by letting the algorithm itself to decide

which model or models should be updated. By using the already learned models to recognize new input in a limited time window, it is possible to decide whether the input is novel or familiar. If the activation value of an existing model exceeds a pre-defined threshold, this model becomes updated by the contents in the analysis window. If no sufficiently high activation is present, a new model will be created for the signal in the analysis window. Initially, the system is created without any models and the first analysis window becomes the first model.

2.1. Novelty detection and learning

When a sequence $X = \{a_1, a_2, \ldots, a_n\}$ is used as input, the subsequence Ω of the first L elements $\Omega(1) = \{a_1, a_2, \ldots, a_L\}$ of the sequence is chosen and the transition frequencies between element pairs $f[a_i,a_j]$, $a \in [1, 2, \ldots, N_A]$, at lags $k_d \in \boldsymbol{k}$ in $\Omega(1)$ are stored into transition frequency matrices $f_c(a_i \mid a_j , k_d)$, where $c = 1$ for the first model, i.e., a separate matrix is created for each lag. Then the frequency matrices are normalized into transition probability matrices P^s by having:

$$P_c^{\mathrm{S}}(a_j \mid a_i,k_d) = \frac{f_c(a_j \mid a_i,k_d)}{\sum_{j=1}^{N_A} f_c(a_j \mid a_i,k_d)} \quad (1)$$

Then the analysis window is shifted S elements forward to position $\Omega(\mathrm{T} = 2) = \{a_{1+T*S}, a_{2+T*S}, \ldots, a_{L+T*S}\}$ and the previously learned models c are used to compute the transition probabilities of the new sequence by using the learned models:

$$A(c,n) = \frac{1}{K}\sum_{d=1}^{K} P_c^{\mathrm{S}}(\Omega[n] \mid \Omega[n-k_d],k_d) \quad (2)$$

i.e., the mean of TPs is computed across all lags $\boldsymbol{k}$. Then the mean probability of each model in $\Omega(T)$ is computed:

$$\hat{A}(c,T) = \frac{1}{L}\sum_{n=1}^{L} A(c,n) \quad (3)$$

Now, if the activation $\hat{A}(c,T)$ of any single model exceeds a pre-defined threshold δ, the TPs of the most activated model are updated according to (1) using the transitions in the sequence $\Omega(T)$. If no sufficiently high activation is achieved, a new model c_m is created using the transitions in $\Omega(T)$. The window is then again shifted L elements and the new subsequence $\Omega(T+1)$ is recognized using the learned models. This windowing process is repeated for the duration of entire training signal, leading to learning of a non-predefined number of models for patterns in the input sequence.

2.2. Enhanced segmentation and classification

In order to enhance contrast between of the learned TP models, the probability that a specific transition from a_i to a_j occurs in the case of model c and lag k, instead of any other models, is incorporated into the *activation matrix P* by having:

$$P_c(a_j \mid a_i,k_d) = \frac{P_c^{\mathrm{S}}(a_j \mid a_i,k_d)}{\sum_{g=1}^{N_C} P_g^{\mathrm{S}}(a_j \mid a_i,k_d)} - \frac{1}{N_C} \quad (4)$$

where N_C is the total number of models. The subtracted term $1/N_C$ ensures that non-informative transitions, i.e., transitions that are equally probable across all C, have a value of zero. The reason why (4) is not applied to novelty detection during learning is that it enforces a forced choice between the existing models. This leads to poor novelty detection performance since the probability mass of each transition across all models is always zero (note that activation values can be negative due to the subtraction of the constant). However, the normalization (4) has a significant impact on segmentation performance.

Now, when a novel utterance is represented, the activation of each model c at each moment of time t is computed

$$A(c,t) = \frac{1}{K}\sum_{d=1}^{K} P_c(X[t] \mid X[t-k_d],k_d) \quad (5)$$

This provides a temporally local activation estimate for each model. The activations are smoothed temporally using a simple moving average filtering in a 480 ms window. Only the most activated model for each moment of time is retained, leading to segmentation of the input into activation stretches of competing models, and segment boundaries are indicated by points in which the winning model changes from one to another (fig. 1).

3. Experiments

3.1. Material and data preprocessing

The speech material consisted of 2000 Finnish child-directed utterances from a female speaker, taken from the ACORNS corpus [7]. The corpus was designed to represent speech input to an infant under the age of one year, and contains simple sentence structures like "*Missä puhelin on nyt?*" ("*Where is the telephone now*?"). Together with all nouns, verbs, adjectives and pronouns, the size of the vocabulary is 38 words plus silence. Half of the utterances were recorded as infant directed speech (IDS) and the other half as adult directed speech (ADS). However, no distinction between these two modes was made in these experiments. 1700 randomly chosen signals were used for training and the remaining 300 were used for testing.

The audio signals were windowed with a Hamming window of length 25 ms and a window step size of 10 ms. Standard MFCC features (11 coefficients + energy) were extracted from each frame, and the relative weights of the energy and first cepstral coefficient (spectral tilt) were reduced by multiplying the coefficients by a factor of 0.25. A randomly chosen subset of MFCC vectors was used to create a vector quantization (VQ) codebook of size $N_A = 150$ using the k-means algorithm. All utterances were then converted to sequences of VQ-indices using the codebook and Euclidean distance metric, yielding one VQ label a_i for each 10 ms frame. The training set resulted in one long sequence (376628 frames) that encompassed all training data and was presented to the SLCM in a single pass. Each test utterance was represented by a separate VQ sequence to allow matching to manual reference on utterance-by-utterance basis.

3.2. Evaluation

The temporal accuracy of the word segmentation was evaluated by comparing the standard deviation σ_s of the distance from detected word boundaries to boundaries produced by automatic HMM-based forced-alignment segmentation. The mean number of insertions per annotated word was also computed to ensure that the apparent increases in the segmentation accuracy were not achieved by introduction of additional segment boundaries.

The contents and quality of the learned models were analyzed by computing the entropy of the distribution of word classes represented by each model. First, the temporal segments of speech where each model was most active were detected. Only segments exceeding 150 ms in length were included in further analysis. These segments were compared to the underlying word-level annotation in order to obtain a distribution

of underlying words that explained what words were actually spoken when the given model was active. Purity of a model's distribution was then measured by the Shannon entropy

$$H(c) = -\sum_{r=1}^{R} P(\alpha_r)\log_R P(\alpha_r) \quad (6)$$

that yields zero for a fully selective model and one for a totally unselective model. Here, R denotes the total number of words and α denotes the word in the reference annotation. Entropy was computed separately for all of the models $\boldsymbol{c}$, and the overall mean entropy H_C was computed by weighting the model entropies with the frequencies f_c that denote how many times the model c was activated in the test set:

$$H_c = \sum_{c=1}^{N_c} f_c H(c) / \sum_{c=1}^{N_c} f_c \quad (7)$$

In addition, word class entropy H_a was measured. H_a indicates how many models exist for a given annotated word class α, and it can be computed similarly to H_c by simply swapping c with α, and replacing R with N_C in (6), and then by computing the (unweighted) mean of these entropies across all words. For a low overall H_a, only a small number of models exist for each annotated word, whereas H_a that approaches one indicates that all models are equally representing all of the words.

3.3. Word segmentation results

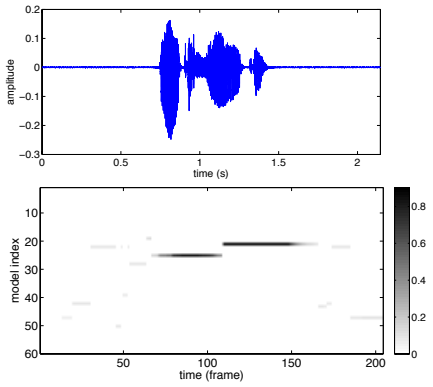

Figure 1: *Signal waveform for utterance "Näytä äiti" (top) and the corresponding model activations at different rows (bottom). Words "näytä" and "äiti" are represented by strong activations of two of the models.*

Lags $\boldsymbol{k} = \{1, 2, \ldots, 8\}$ were used in the experiments. The training produced a total of 60 models using parameter values of δ = 0.039, L = 600 ms, and S = 200 ms. Once the training data was processed as described in section 2.1, the learned models were normalized for recognition using (4) and the test set of 300 novel utterances was used as input to the recognition process (5). This yielded 1293 word segments exceeding 150 ms length (4.31 per utterance).

Figure 2 illustrates the overall entropies H_c and H_a of the process as a function of training time while figure 3 shows a surface-plot of the same entropies for all models separately. As can be seen, the entropies drop very rapidly in the beginning as new models are being created. Most of the models are already in place after 10 minutes of speech and only a small number of new models are formed later. As the amount of training time increases, selectivity of the existing models increases as well. After training over the entire training set, the overall selectivity achieves H_c = 0.22 and the annotation entropy H_a = 0.25.

Figure 4 shows the number of insertions per annotated word (left) and the mean segment boundary deviation from the reference (right) as a function of training time. Behavior of the curves is not strictly monotonically decreasing, especially near the beginning. However, the overall descending trend in both curves indicates that the segmentation becomes more accurate as more training data is introduced, and that this increase in accuracy is not obtained by introducing superfluous boundaries.

Table 1 shows the underlying word distributions for the 17 most selective models after removing models reacting mainly to silence (#h). There are nine word models that are reacting only to the corresponding word over 70 % of the time. Models for words *"katso"* (look), *"hassu"* (funny), *"pullo"* (bottle), and *"kirja"* (book) are especially selective, since they only react to silence in addition to the their own word.

The reason why silence is included in many of the models is the fact that many of the words are systematically positioned at the beginning or ending of an utterance, and the windowing mechanism in the learning process captures some of the silence preceding or following the words. In addition, the forced-alignment annotation of the corpus systematically deviates from SLCM at utterance endings due to the slowly fading "breathy" spectrum characteristic of words in sentence final position.

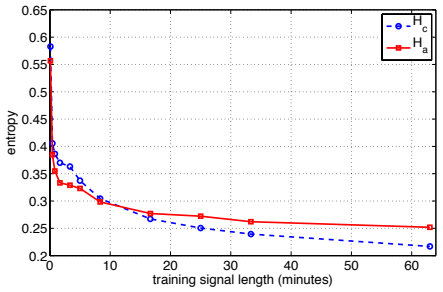

Figure 2: *Model entropy H_c and word class entropy H_a as a function of time trained (in minutes).*

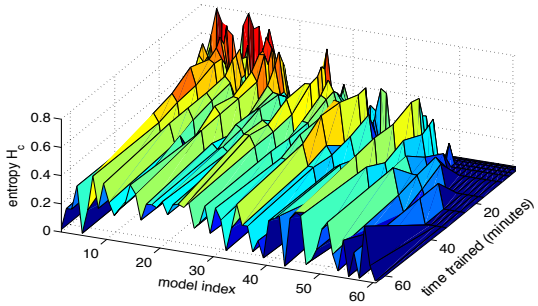

Figure 3: *Surface-plot of model entropies H_c as a function of training time. The zero-entropy area in the right-back corner is due to models that were formed only later on during the training.*

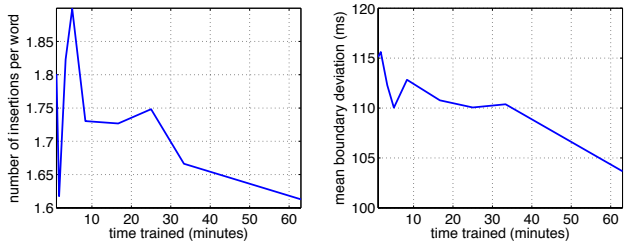

Figure 4: *Number of segment boundary insertions per annotated word (left) and mean boundary deviation from reference (right) as a function of training time.*

Table 1. *Correspondence between the learned models and the reference annotation. Each row represents the contents of a model i. N denotes the number of times a given model was activated in the test set. Only words with p>0.05 are shown.*

i	*N*	word 1	p_1	word 2	p_2
1	26	*katso*	0.9		
2	8	*pullo*	0.84	*on*	0.07
3	7	*vaippa*	0.83	*nyt*	0.13
4	16	*hassu*	0.82	*#h*	0.14
5	18	*pullo*	0.82	*#h*	0.17
6	26	*kirja*	0.8	*#h*	0.18
7	6	*nyt*	0.8	*on*	0.13
8	35	*on*	0.78	*puhelimen*	0.16
9	8	*nyt*	0.74	*isi*	0.13
10	27	*auto*	0.67	*#h*	0.27
11	20	*äiti*	0.66	*kylvyn*	0.08
12	12	*onpas*	0.65	*on*	0.15
13	33	*Johanna*	0.63	*äidin*	0.31
14	24	*hassu*	0.61	*#h*	0.2
15	6	*isi*	0.61	*nyt*	0.33
16	24	*ota*	0.59	*#h*	0.27
17	12	*kiva*	0.57	*isi*	0.22

When the detected speech segments are listened to, many of the models exhibit fairly accurate word segmentation in terms of subjective perceptual judgment. As can be expected based on table 1, some of the models are very pure and only rarely contain extraneous signal contents in addition to one specific word, whereas some of the models represent two different words. These multi-word models are often due to the windowing that spans partially across two short words when the model is first created. This causes the model to react to the two words either in isolation or combination, and subsequently the model will be updated to incorporate more and more detailed representations of the both words. This type of problem is especially pronounced for often co-occurring short words such as "*se on*" (*it is*), "*nyt on*" (*now is*), or "*nyt isi*" (*now daddy*).

4. Discussion and conclusions

The present work demonstrates that automatic word segmentation and learning of primitive ungrounded lexical items from real speech is possible without pre-existing linguistic or phonemic knowledge or contextual support by simply analyzing transitional probabilities between atomic acoustic events. This provides support to the distributional learning hypothesis (e.g., [1-2]) and PRIMIR theory of language acquisition [4].

The proposed algorithm is computationally straightforward and it is likely that, with further development, the performance of the algorithm can be enhanced. For example, by utilizing a varying length windowing synchronized to the temporal envelope of the speech could facilitate learning and increase model selectivity. Despite current shortcomings, the algorithm clearly demonstrates a capability for the incremental learning of internal representations from speech without supervision.

Previous approaches to unsupervised word learning have been reported by Park & Glass [8] and Aimetti [9]. Park & Glass [8] used dynamic time-warping (DTW) to find recurring stretches of speech signals and then linked these acoustically similar segments through graph clustering. A cognitively inspired system by Aimetti [9] also performs unsupervised acquisition of word models by using DTW-based detection of recurring units between acoustic episodes. The difference with DTW-based approaches and the SLCM is that a DTW-based system looks for repetitions on an utterance by utterance basis, requiring storage of feature representations of all utterances in memory, whereas the SLCM does not store episodic representations in full detail, but only stores statistical dependencies ("TPs") between atomic acoustic units in the case of each model and uses the obtained statistical models to recognize new inputs. This makes SLCM computationally very attractive, since the computational complexity does not increase with the input length, but only linearly as a function of number of learned models. However, the DTW approaches are also compatible with the idea of tracking transitional probabilities in speech, i.e., they succeed in the task if such statistical structure exists, even though the algorithms do not explicitly count and store probabilities of subsequent acoustic events.

Finally, despite the possibility for totally unsupervised learning of lexical candidates, it should not be forgotten that real linguistic development takes place in a much richer world where the learner is embedded in a tight interaction with its caregivers and the surrounding environment [10]. When compared to the unimodal learning situation as was used in this work, the interaction with the complex real world and other social agents actually imposes additional constraints and provides feedback that can aid in linguistic development (see, e.g., [11]). Also, the only way to acquire meaning for the auditory word forms is to ground them in combination with other perceptual systems and actions of the agent, something that was not studied in this work. It is also noteworthy that a real infant is exposed to a much larger amount of speech during infancy than what was used in this study, or any other known studies attempting to perform computational modeling of language acquisition.

5. Acknowledgements

This research was funded by the Nokia Research Center (NRC) Tampere Finland and the Finnish Graduate School in Language Studies (Langnet) funded by Ministry of Education of Finland. The author would like to thank Kris Demuynck for providing the HMM forced-alignment annotation and Toomas Altosaar for useful comments on the manuscript.

6. References

[1] Saffran, J.R., Aslin, R.N. and Newport, E.L., "Statistical Learning by 8-Month-Old Infants", Science, 274:1926-1928, 1996.

[2] Saffran J.R., "Words in the sea of sounds: the output of infant statistical learning", Cognition, 81:149-169, 2001.

[3] Pelucchi, B., Hay, J.F. and Saffran, J.R., "Eight-month-old infants track backward transitional probabilities", Cognition, 113(2):244-247, 2009.

[4] Werker, J.F. and Curtin, S., "PRIMIR: A Developmental Framework of Infant Speech Processing", Language Learning and Development, 1:197-234, 2005.

[5] Räsänen, O., Laine, U.K. and Altosaar, T., "Computational language acquisition by statistical bottom-up processing", Proc. Interspeech '08, 1980-1983, 2008.

[6] Räsänen, O., Laine, U.K. and Altosaar, T., "A noise robust method for pattern discovery in quantized time series: the concept matrix approach", Proc. Interspeech '09, 3035-3038, 2009.

[7] http://www.acorns-project.org.

[8] Park, A. and Glass, J.R., "Unsupervised word acquisition from speech using pattern discovery" Proc. ICASSP'06, 409-412, 2006

[9] Aimetti, G., "Modelling Early Language Acquisition Skills: Towards a General Statistical Learning Mechanism", Proc. EACL-2009, 2009.

[10] Meltzoff, A.N., Kuhl, P.K., Movellan, J. and Sejnowski, T. J., "Foundations for a New Science of Learning", Science, 325:284-288, 2009.

[11] Oudeyer, P.-Y. and Kaplan, F., "Discovering Communication", Connection Science, 18(2):189-206, 2006.