Computational language acquisition by statistical bottom-up processing

Okko Räsänen[1], Unto K. Laine[1], Toomas Altosaar[1]

[1] Department of Signal Processing and Acoustics, Helsinki University of Technology,
Speech Technology Team

`orasanen@cc.hut.fi, Unto.Laine@tkk.fi, Toomas.Altosaar@tkk.fi`

Abstract

Statistical learning of patterns from perceptual input is an increasingly central topic in cognitive processing including human language acquisition. We present an unsupervised computational method for statistical word learning by analysis of transitional probabilities of subsequent phone pairs. Results indicate that word differentiation is possible with this type of approach and are in line with previous behavioral findings.

Index Terms: computational language acquisition, speech segmentation, speech clustering, statistical learning

1. Introduction

Child language acquisition is a central topic in several fields of scientific research. It can be investigated from several perspectives, e.g., by organizing follow-up studies with regular linguistic skills assessments, using modern brain imaging techniques for observation of neural activation in linguistic tasks, or by conducting psychological experiments that assess behavioral responses of the test subjects to specific stimuli. The common goal of these different approaches is to create a uniform and coherent understanding of human language learning and language related cognitive processing

In the past few decades development in technology has enabled the use of computational models for testing and development of more accurate hypotheses regarding human cognition and behavior. Several models for speech perception and working memory have been successful in imitating several aspects of human cognitive functions, although they also all have their limitations. While a good test of the plausibility of any theoretical model is to test it by implementing it as a functional algorithm, experimental models can also be used for hypothesis development that can be further verified with behavioral studies. Modern brain imaging techniques also provide important information regarding the processing that takes place in our brains. However, it is not a straightforward task to determine the direct mappings between neural processes and their behavioral and computational counterparts. In the end, if computational models are to be considered descriptive, they should support behavioral data and should not contradict that what is known about neurological processes existing in the primate brain.

In the Acquisition of Communication and Recognition Skills (ACORNS) project the central aim is to develop computational mechanisms for creating an artificial agent that is capable of acquiring human verbal communication behavior. The area of focus is on dynamic learning by interaction with the environment, or more precisely, multi-modal learning where acoustic input is coupled with conceptual information about the learner's surroundings. Unlike many traditional ASR systems, the learning agent (LA) is expected to imitate child language acquisition and possibly supply new insights to the wide theoretical framework of human language processing.

In this paper we present one possible approach for creating inner representations from the acoustic speech input waveform of that has many analogies to cognitive aspects of language learning. Bottom-up speech segmentation yielding phone-sized units and segmental data clustering enable the statistical analysis of phones. We will show that the use of only transitional probabilities of subsequent phone pairs enables differentiation of words from the speech stream.

2. Theoretical background

It is well-established that infants are capable of distinguishing phonetic contrasts of basically any language during their first months of their lives [1]. However, as the child matures this ability endures only in the languages that the infant is continuously being exposed to [2,3]. Interestingly enough, the contrast resolution capability is recovered very quickly for non-used languages if the infant interacts for even short periods of time face-to-face in a language immersion situation [4]. If infants are exposed to only the sound stream without face-to-face interaction, no changes in resolution capability occur. This is closely related to a mutual necessity of context and speech that is considered as a prerequisite for word learning in humans [5], and may also be referred to as multimodal learning since our knowledge about our surroundings is obtained only through sensory input.

How does the infant acquire the necessary skills for perception of speech? Saffran and her colleagues have shown that infants exploit the statistical properties of speech signals for learning. In one of their studies they showed that infants use transition probabilities of neighboring speech sounds for segmentation of words from fluent speech [6]. Later, they determined that this type of statistical learning is not only confined to linguistic material in infants and adults, but also that acoustic tone stream segmentation can be performed by using similar statistical cues [7]. Infants also learn word-like units from artificially generated non-sense word sequences [8] and even statistical dependencies of non-adjacent units [9]. This type of statistical learning can take place without conscious awareness, which does not exlude the idea of exploiting phone- or syllable-level units in speech processing even when, e.g., illiterate adults may not be able to distinguish syllables from speech [10]. This seems to suggest that there is a general learning process that extracts information from the recurrent patterns found in the sensory input in order to construct structural descriptions of the surrounding world.

During the first months of their lives, infants pay more attention to the short-term spectral and temporal properties of

auditory (speech) signals. It has been suggested that as their inner language related representation becomes enriched, infants' tendency to process larger temporal units starts to dominate [11,12] and the speech recognition process shifts from the statistical analysis of phones and syllables towards words. If statistical analyses are able to capture the essential content of small speech units (e.g., phones), and if it is possible to compress this information into new representations, then these representations can potentially be used as inputs for higher-level statistical analyses that enable the discovery of larger scaled temporal dependencies, e.g., word-like units or grammatical regularities. Experimental evidence for this type of higher level statistical processing exists. For example, Saffran [13] has suggested that humans may use statistical information for acquisition of abstract components such as grammatical rules found in natural languages.

From the neurological perspective, the statistical processing of incoming data is closely related to the plasticity of the brain (see, e.g., [14]) that is known to be most effective during early childhood and which degrades as a function of age. Neural plasticity and statistical learning offer a possible explanation for a so-called critical period of language acquisition. As the person becomes older, large-scale reorganization of the neural networks becomes more difficult as the neural connections have become strongly consolidated for specific statistical regularities (frequency effect). Statistical adaptation already takes place at the feature extraction level in the sensory input channel and can be exemplified, e.g., by the formation of a solely vertically sensitive visual cortex V1 found in cats that have grown-up in an environment containing only vertical contrasts [15]. These findings support the notion that front-end processing in perception adapts to the properties of sensory input in order to provide for an efficient mapping from environmental stimuli towards inner representative units that can be used in statistical analysis.

One should keep in mind that many other mammalian species are able to process speech sufficiently well in order to discover the relations between human spoken commands and their ensuing actions (classical conditioning with verbal stimulus) even when their sound perception mechanisms are very differently specialized from human hearing. It may even be possible to hypothesize that this type of stimulus segmentation (extraction of coherent units), adaptation, and associative statistical learning is provided by universal mechanisms implemented in the cortical and/or sub-cortical structures and is shared by many if not all mammalian species.

3. Computational approach to learning

To model the language acquisition process from the statistical learning perspective, the system must be able to convert acoustic waveform signals into rich descriptive units that can be used for statistical analysis. The goal is to experiment with statistical pattern discovery for word recognition. Experimental evidence from [6] suggests that this can be accomplished using small functional units of speech: phones, or in this case, phone-like units that are obtained from blind segmentation.

3.1 Processing stages in machines and humans

In this computational approach, the first stages of sound perception are approximated by a so-called blind bottom-up process where no external knowledge or previous stimuli intervene with the process: the system is "born" without any linguistic knowledge (the only *a priori,* innate aspect being the structure allowing the learning). Each incoming utterance is first segmented into phone- (or possibly diphone-) sized units by a blind speech segmentation algorithm (see [16]). Segmentation output contains approximately 75 % of the boundaries within 20 ms of the manual reference boundaries without notable over-segmentation.

The segments are described by their spectral properties as single units and clustered with an incremental clustering algorithm that uses the spectral cross-correlation of segments as a distance measure in order to create a coarse classification resulting in phone-like categories. This creates a symbolic sequence representation for each incoming utterance that can be used in statistical analysis and word discovery.

This algorithmic phone segmentation process can be considered as a processing step that infants are innately capable of since it has been shown that they can discriminate phone contrasts already at birth [1] as well as extract statistical information from relations of neighboring sound units [6,7]. Mechanisms for spectral descriptions of sound segments are also innate, as frequency properties of sounds are extracted in separate bands in the cochlea ending up as tonotopic maps in the auditory cortex [17].

To find an analogy between the clustering and cognitive processes, it may be useful to consider such incremental clustering as a sort of neural adaptation to properties of sensory input: processing of incoming stimuli is shaped by all previous inputs that continuously update the phonetic categories to which the incoming phones are mapped to. Categories are only created and sustained for those spectral segments that the system is being exposed to and unused clusters are pruned away if they are rarely activated. In other words, incremental clustering is an unsupervised mechanism for the categorical perception of sounds that builds discrete categories dynamically on the basis of acoustic input (N.B. nothing so far has been limited to speech alone, as the segmentation can be applied in principle to any type of sound material). This is especially interesting since categorical perception is a very central phenomenon in all aspects of human perception [18].

Statistical learning can be built upon categorical activations. Cluster based categories form a symbolic alphabet if a unique label is chosen for each category (e.g., 1, 2, 3… *N*). Each new acoustic utterance produces a sequence of symbols that contains a discrete description of signal content and simultaneously causes slight adaptations to take place in the mapping process from waveform to categories. Different pattern discovery methods can now be applied for detecting structures in the symbol sequences. These structures can be reused for statistical analysis to obtain higher-level descriptions in a hierarchical manner. To move towards this goal, this paper presents a very simple finite state-machine that is capable of differentiating words by exploiting the transitional probabilities of subsequent cluster pairs (cf., [6]).

From the cognitive point of view this pattern discovery process may represent subsequent neural processing steps that are involved in categorical perception, or, it may be the interaction between activation in categorically related substrates and a general mechanism responsible for building statistical models. However, learning and language capability in general are very complex problems in terms of cognitive processing, and it is not very well understood what mechanisms truly underlie implicit statistical learning.

3.2 The bottom-up algorithm

The blind segmentation algorithm that is currently under development uses a linear frequency (FFT) representation of a signal for the detection of segment boundaries. The signal is first windowed into short 6 ms frames and the cross-correlation of frames is utilized to enable detecting sudden changes in the signal. Segment boundaries are hypothesized to exist at locations where the change in the spectral properties exceeds a manually defined threshold level. For feature extraction, spectral tilt and mean energy are removed from the FFT frames and the vectors normalized to zero mean unit vectors. Each segment is divided into onset and final sections (initial 40 % and final 60 % of the duration) and the algorithm picks the five most contrastive spectral vectors for both sections and averages them into two spectral representations for each segment.

These segmental descriptions are then clustered using a simple incremental algorithm that computes the cross-correlation of the incoming segment's spectral vector to all existing clusters and merges it to the best match if the merging threshold d_{min} is exceeded. If no suitably close matching cluster is found, a new cluster is created. Clusters retain information only about their cluster centroid that is the mean spectrum of all merged segments. The initial and final sections of the segments are clustered into separate spaces and lower thresholds are used for the final sections to avoid sparseness, as phone endings are more dependent on context and contain more variation. A numerical label is assigned for each cluster to provide for a symbolic sequential description of the input. A more comprehensive description of the segmentation and clustering algorithm and its performance can be found in [16].

4. Experiments

4.1 Training procedure

A computational experiment to examine word learning via cluster transition probability analysis was conducted. Speech material used in the experiments was recorded as part of the ACORNS project and consisted of two thousand Finnish utterances spoken by two different female speakers. Each utterance consisted of a carrier sentence containing an embedded keyword. In total there were nine identical Finnish keywords (e.g., *vaippa* (eng. *diaper*), *kylpy* (*bath*), *isi* (daddy), *äiti* (*mother*),* etc.) for both speakers and one unique keyword (the infant's name) for each speaker, keywords recurring 200 times for each speaker. Half of the utterances were recorded as infant directed speech (IDS) and half of them as adult directed speech (ADS) but no distinction between these two modes was made in this experiment.

Multimodality that enables associative learning was simulated through contextual tagging: in the training phase, in addition to the spoken utterance, a symbolic tag corresponding to a keyword was simultaneously presented to the system. This simulates a situation where the caregiver is talking to the infant and directs the child's attention to a salient object, e.g., by saying "*Look, a bottle*" and showing a bottle simultaneously.

Speech is first automatically segmented into phone-like units and features for each segment are extracted. A cluster space is created for the onsets and the final sections of segments by presenting all the utterances from the test material to the system. This simulates a situation where the infant is exposed to large amounts of native speech that enables adaptation to properties of the language specific sounds before learning

actual words begins. Small clusters, defined as clusters containing less than two segments, are removed. Clusters that have moved closer to each other than a specified merging threshold are merged. This approach resulted in approximately 100 clusters, each typically containing hundreds to several thousands of segments. Finally, the speech material was clustered into this pre-formed cluster-space again, leading to a symbolic sequence being generated for each utterance.

To calculate and store the phone transition probabilities, a simple finite state machine was implemented. The algorithm takes one symbol sequence $\mathbf{U}$ of an utterance at a time and reads the corresponding keyword tag from a metafile. If the keyword has not been seen before, the algorithm creates a new "concept matrix" $\mathbf{C}$ of size M x M where M is the size of the cluster alphabet. This matrix is a (stochastic) transition matrix where transitions between each two subsequent symbols in the corresponding sequence are added. All transitions occurring in the training data are added to the matrices in a similar manner and the transition probabilities from phones are normalized to sum up to one. Every time a keyword tag appears in the data, the transition probabilities for the corresponding matrix are updated. This process leads to N_k concept matrices where N_k is the number of keywords in the training material.

4.2 Testing procedure

A sequence $\mathbf{U}$ corresponding to the utterance being tested is windowed using a sliding window, providing a sub-sequence $\mathbf{S_{i,l}}$ for each step. Taking into account all transitions from a phone to the next one in $\mathbf{S_{i,l}}$ and summing up the corresponding values from concept matrices $\mathbf{C_k}$ leads to a probability for each concept k for each sub-sequence $\mathbf{S_{i,l}}$. If the probability for a transition in a concept matrix is zero, a small penalty to the probability sum is introduced. The process is repeated for every possible sub-sequence window location in $\mathbf{U}$ and for several different sub-sequence lengths $l = l_0,...,L$ of $\mathbf{S_{i,l}}$ where l_0 and L are chosen manually. The computation is made in parallel for both the onset and final sections of the segments and the transition probabilities of these two are combined. The concept corresponding to the largest probability is chosen as the word hypothesis. If the word hypothesis corresponds to the tag existing in the annotation, the word is considered as correctly recognized. The recognition accuracy was estimated using a N-fold approach where utterances were randomly divided into training and evaluation material and the results were averaged over several trials.

4.3 Findings

The best results were obtained with a relatively low merging threshold (in terms of phone differentiation; $d_{min} = 0.4$ for onset and $d_{min} = 0.25$ for endings). Higher thresholds led towards overly sparse transition matrices where different realizations of the same word started to follow entirely different paths. Lower thresholds, on the other hand, started to affect the required ability to differentiate between different phones.

The first experiment was conducted using 2000 utterances from a single speaker. The algorithm learns the transitional probability structure very quickly, leading to significantly above chance accuracy even when only a few utterances per each keyword have been seen (fig. 1). Recognition accuracy achieved a 74.5 % correctly recognized keyword level when 1800 (90 %) of the utterances were used for training and the remaining 10 % were used for evaluation. The increase in recognition accuracy as a function of trained utterances did not seem to saturate even at
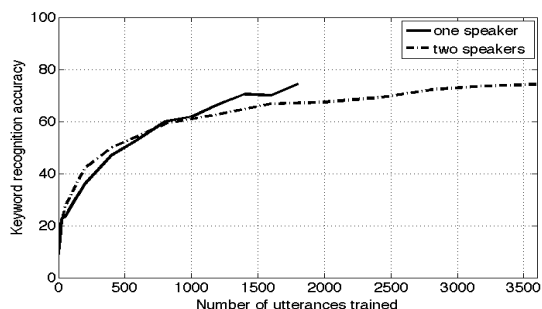
Figure 1: *Recognition accuracy.*

this point, but the amount of training material available for a single speaker dictates the upper limit for this type of experiment, an issue that may be addressed by future development of the speech corpora.

When material from two speakers are combined (a total of 4000 utterances) the clustering process, using the same thresholds as with one speaker, creates a slightly larger symbolic alphabet due to the increased variety of speech existing in the material. This also increases the sparseness of the transition matrices. When utterances from different speakers are mixed in a random order (fig 1., dashed line) the algorithm again achieves a very similar recognition accuracy (74.3 %) where 90 % of the utterances were used for training. Also, the recognition accuracy is only ≈10 % worse when a similar number of utterances are used for training as in the single speaker case.

5. Conclusions

It was shown that blind speech segmentation producing phone-like segments, segment feature extraction, followed by segment clustering, leads to a relatively simple description of the underlying speech signal that can be used for statistical analysis. By taking into account only the transitional probabilities of subsequent phone pairs, and ignoring large-scale temporal order, and, e.g., prosodical aspects of speech, it is still possible to obtain relatively high word recognition accuracies with a limited vocabulary. This effect is notably similar to behavioral results obtained in [8,9] regarding implicit statistical learning in which infants were able to segment words from fluent speech by using the statistical relationship between neighboring sounds after only two minutes of exposure to the artificial speech stream.

A similar approach for unsupervised word discovery was reported earlier by ten Bosch & Cranen [19]. The main difference between the algorithm presented here and their approach is that they used a DTW matching technique for word discovery, in which a new sequence was compared to previous sequences to find the most likely path. They also used k-means clustering for labeling, which leads to a more accurate classification of the prevailing data, but requires the presence of all classifiable data during the initial clustering process if the number of clusters is kept constant. Despite these differences, their results also strongly support the idea of bottom-up statistical learning without pre-defined linguistic constraints.

Finally, it should be noted that the computational system used in these experiments is still under development. Finding more sophisticated approaches for all levels of processing will be one of the main goals along with the development of an entirely self-directed learning mechanism. The potential of accessing higher level linguistic structures such as syntax- or even semantics by hierarchical unsupervised statistical learning will also be an interesting research subject in the future.

7. References

[1] Blumstein S., Mehler J., Bertoncini J. & Bilejljac-Babic R.: Discrimination in neo-nates of very short CVs. Journal of Acoustical Society of America, 82, 1987

[2] Trehub S.: The Discrimination of Foreign Speech Contrasts by Infants and Adults. Child Development, 47(2), 1976

[3] Werker J. & Tees R.: Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavioral Development., 7, 49-63, 1984

[4] Kuhl P., Tsao F-M. & Liu H-M.: Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. In Proc. National Academy of Sciences., 100(15), 9096-9101, 2003

[5] Benedict H.: Early lexical development: Comprehension and production. Journ. of Child Language, 6, 183-200, 1979

[6] Saffran J., Aslin R. & Newport E.: Statistical Learning by 8-Month-Old-Infants. Science, 274, 1926-1928, 1996

[7] Saffran J., Johnson E., Aslin R. & Newport E.L.: Statistical learning of tone sequences by human infants and adults. Cognition, 70, 27-52, 1999

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

[9] Newport E. & Aslin R.: Learning at a distance I. Statistical learning of non-adjacent dependencies. Cognitive Psychology, 48, 127-162, 2004

[10] Morais J., Cary L., Alegría J. & Bertelson P.: Does awareness of speech as a sequence of phones arise spontaneously? Cognition, 7, 323-331, 1979

[11] Stager C. & Werker J.: Infants listen for more phonetic detail in speech perception than in word-learning tasks. Nature, 388, 1997

[12] Swingley D.: Statistical clustering and the contents of the infant vocabulary. Cognitive Psychology, 50, 86-132, 2005

[13] Saffran J.: The Use of Predictive Dependencies in Language Learning. Journal of Memory and Language, 44, 493-515, 2001

[14] Buonomano, D. & Merzenich M.: Cortical plasticity: from synapses to maps. Annual Review of Neuroscience, 21, 149-186, 1998

[15] Blakemore C. & Cooper G.: Development of the brain depends on the visual environment. Nature, 228, 477-478, 1970

[16] Räsänen O: Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture. Master's thesis, http://lib.tkk.fi/Dipl/2007/urn010123.pdf, 2007

[17] Aitkin L.: The Auditory Cortex. Chapman and Hall, 1990

[18] Harnad S. (ed.): Categorical Perception: The Groundwork of Cognition. Cambridge University Press, 1987

[19] Ten Bosch L. & Cranen B.: A computational model for unsupervised word discovery. In Proc. Interspeech, 2007