A Computational Model of Word Learning from Multimodal Sensory Input

Deb Roy (dkroy@media.mit.edu)
Media Laboratory, Massachusetts Institute of Technology, 20 Ames Street,
Cambridge, MA 02139, USA

Abstract

How do infants segment continuous streams of speech to discover words of their language? Current theories emphasize the role of acoustic evidence in discovering word boundaries (Cutler 1991; Brent 1999; de Marcken 1996; Friederici & Wessels 1993; see also Bolinger & Gertsman 1957). To test an alternate hypothesis, we recorded natural infant-directed speech from caregivers engaged in play with their pre-linguistic infants centered around common objects. We also recorded the visual context in which the speech occurred by capturing images of these objects. We analyzed the data using two computational models, one of which processed only acoustic recordings, and a second model which integrated acoustic and visual input. The models were implemented using standard speech and vision processing techniques enabling the models to process sensory data. We show that using visual context in conjunction with spoken input dramatically improves learning when compared with using acoustic evidence alone. These results demonstrate the power of inter-modal learning and suggest that infants may use evidence from visual and other non-acoustic context to aid in speech segmentation and spoken word discovery.

Introduction

Around their first birthday, infants first begin to use word [1] which refer to salient aspects of their environment including objects, actions, and people. They learn these words by attending to the sights, sounds, and other sensations. The acquisition process is complex. Infants must successfully segment spoken input into units which correspond to the words of their language. They must also identify semantic categories which correspond to the meanings of these words. Remarkably, infants are capable of all these processes despite continuous variations of natural phenomena and the noisy input provided by their perceptual systems.

This paper presents a computational model of early word learning which addresses three interrelated problems: (1) Segmentation of fluent speech without a lexicon in order to discover spoken words, (2) Categorization of context corresponding the referents of words, and (3) Establishment of correspondence between spoken words and contextual

[1] The term “word” is used throughout this paper in accordance with Webster's Dictionary: “A speech sound or combination of sounds having meaning and used as a basic unit of language and human communication.”

categories. These three problems are treated as different facets of one underlying problem: to discover structure across spoken and visual inpu[2].

This model has been implemented using standard speech and vision processing techniques. It is able to learn from microphone and camera input (Roy 1999; Roy 2000). We used the model to evaluate the benefit of inter-modal structure for the problem of speech segmentation and word discovery. To gauge the relative usefulness of integrating visual context, we also implemented a uni-modal system which discovered words based on only acoustic analysis (i.e. without access to visual input). Our evaluations demonstrate that dramatic gains in performance are attained when inter-modal information is leveraged. These results suggest that infants would also benefit from attending to multimodal input during even the earliest phases of speech segmentation and spoken word discovery. This work differs from previous computational models of language learning (eg. Gorin 1995; Feldman et. al. 1996; Siskind 1996) in that both linguistic and contextual input are derived from physical sensors rather than relying on human generated symbolic abstractions.

CELL: A Model of Learning from Audio-Visual Input

We have developed a model of cross-channel early lexical learning (CELL), summarized in Figure 1. This model discovered words by searching for segments of speech which reliably predicted the presence of visually co-occurring shapes. Input consisted of spoken utterances paired with images of objects. This approximated the input that an infant might receive when listening to a caregiver while visually attending to objects in the environment.

A speech processor converted spoken utterances into sequences of phoneme probabilities. We built in the ability to categorize speech into phonemic categories since similar abilities have been found in pre-linguistic infants after exposure to their native language (Kuhl et al. 1992; Werker & Tees 1983). At a rate of 100Hz, this processor computed the probability that the past 20 milliseconds of speech belonged to each of 39 English phoneme categories or silence. The phoneme estimation was achieved by training an artificial recurrent neural network similar to (Robinson 1994). The network was trained with a database of phonetically transcribed speech recordings of adult native English speakers (Seneff & Zue 1996). Utterance boundaries were automatically located by detecting stretches of speech separated by silence.

A visual processor was developed to extract statistical representations of shapes from images of objects. The visual processor used `second order statistics' to represent object appearance, as suggested by theories of early visual processing (Julesz 1971). In a first step, edge pixels of the viewed object were located. For each pair of edge points, the normalized distance between points and the relative angle between the two edge points were computed. All distances and angles were accumulated in a two-dimensional histogram representation of the shape (the `second order statistics'). Three-dimensional shapes were represented with a collection of two-dimensional shape histograms, each derived from a particular view of the object. To gather visual data for evaluation experiments, a robotic device was constructed to gather images of objects automatically (Figure 2). The robot took images of stationary objects from various vantage points. Each

[2] In this paper we only discuss learning from audio-visual input. The underlying model is able to learn from any combination of input modes, i.e. the model is not dependent on speech or vision. See (Roy 1999) for more details.

object was represented by 15 shape histograms derived from images taken from 15 arbitrary poses of the robot. The chi-squared divergence statistic was used to compare shape histograms, a measure that has been shown to work well for object comparison (Schiele & Crowley 1996). Sets of images were compared by summing the chi-square divergences of the four best matches between individual histograms.

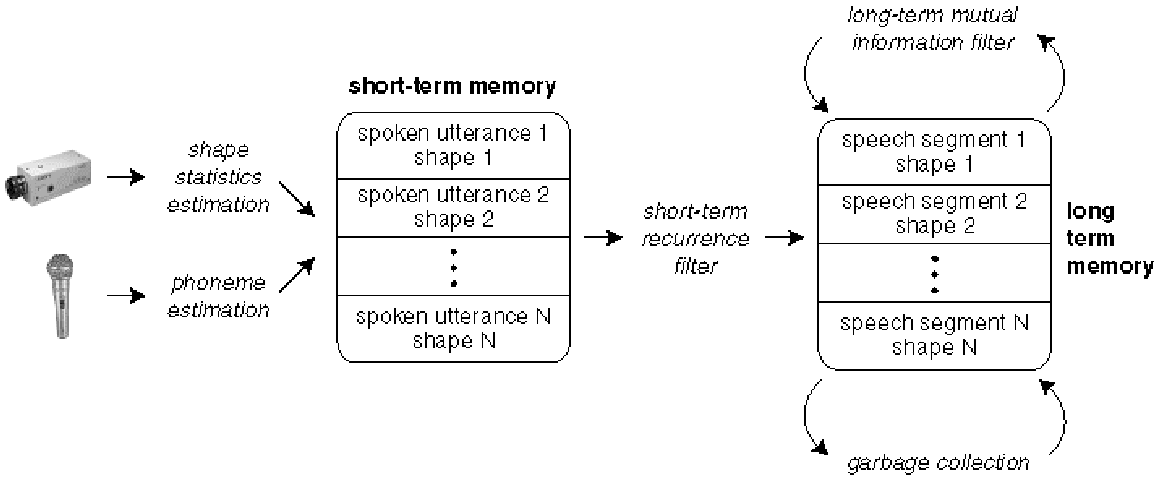
long-term mutual information filter
short-term memory
shape statistics estimation
phoneme estimation
spoken utterance 1 shape 1
spoken utterance 2 shape 2
spoken utterance N shape N
short-term recurrence filter
speech segment 1 shape 1
speech segment 2 shape 2
speech segment N shape N
long term memory
garbage collection

Figure 1: The CELL model. Camera images of objects are converted to statistical representations of shapes. Spoken utterances captured by a microphone are mapped onto sequences of phoneme probabilities. The short term memory (STM) buffers phonetic representations of recent spoken utterances paired with representations of co-occurring shapes. A short-term recurrence filter searches the STM for repeated sub-sequences of speech which occur in matching visual contexts. The resulting pairs of speech segments and shapes are placed in a long term memory (LTM). A filter based on mutual information searches the LTM for speech-shape pairs which usually occur together, and rarely occur apart within the LTM. These pairings are retained in the LTM, and rejected pairings are periodically discarded by a garbage collection process.

Phonemic representations of multi-word utterances and co-occurring visual representations were temporarily stored in a short term memory (STM). The STM had a capacity of five utterances, corresponding to approximately 20 words of infant-directed speech. As input was fed into the model, each new [utterance,shape] entry replaced the oldest entry in the STM. A short-term recurrence filter searched the contents of the STM for recurrent speech segments which occurred in matching visual contexts. The STM focused initial attention to input which occurred closely in time. By limiting analysis to a small window of input, computational resources for search and memory for unanalyzed sensory input are minimized as is required for cognitively plausibility.

To determine matches, an acoustic distance metric was developed (Roy 1999) to compare each pair of potential speech segments drawn from the utterances stored in STM. This metric estimated the likelihood that the segment pair in question were variations of similar underlying phoneme sequences and thus represented the same word. The chi-squared divergence metric described earlier was used to compare the visual components associated with each STM utterance. If both the acoustic and visual distance were small, the segment and shape were copied into the LTM. Each entry in the LTM represented a hypothesized prototype of a speech segment and its visual referent.

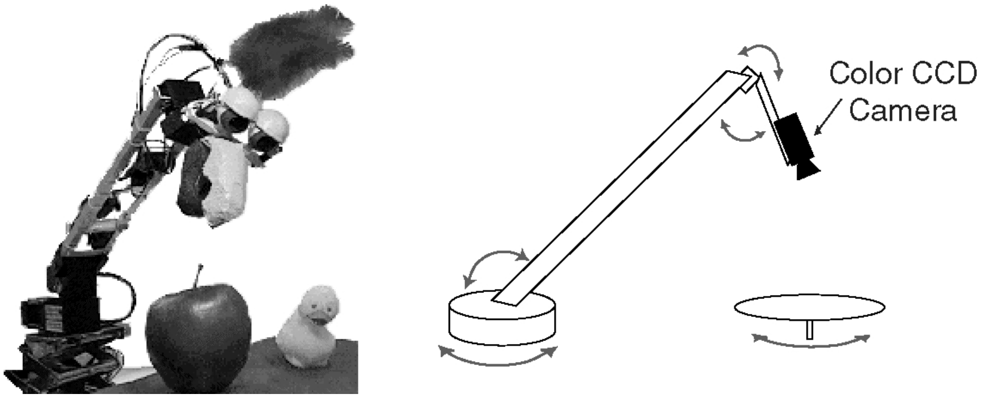

Figure 2: A robot was built to capture images of objects from multiple vantage points. The schematic on the right shows the five degrees of freedom of the imaging system including a turntable for rotating objects. As can be seen from the photograph on the left, the system was designed as a synthetic character to experiment with notions of embodied human-computer interfaces (see Roy, 1999; Roy et al. 1997).

Infant-directed speech usually refers to the infant's immediate context (Snow 1977). When speaking to an infant, caregivers rarely refer to objects or events which are in another location or which happened in the past. Guided by this fact, a *long-term mutual information filter* assessed the consistency with which speech-shape pairs co-occurred in the LTM. The mutual information (MI) between two random variables measures the amount of uncertainty removed regarding the value of one variable given the value of the other (Cover & Thomas 1991). Mutual information was used to measure the amount of uncertainty removed about the presence of a specific shape in the learner's visual context given the observation of a specific speech segment. Since MI is a symmetric measure, the converse was also true: it measured the uncertainty removed about the co-occurrence of a particular speech segment given a visual context. Speech-shape pairs with high MI were retained, and periodically a garbage collection process removed hypotheses from LTM which did not encode associations with high MI.

RECUR: A Model of Learning from Acoustic Input

For comparative purposes, we developed a second model, *RECUR*, which segmented speech using only acoustic information (Figure 3). The acoustic processing in RECUR was identical to that in CELL allowing us to compare them with the same evaluation data.

RECUR discovered words by searching for recurrent sequences of speech sounds. The underlying idea of the model, common in current theories of speech segmentation (Brent 1999; de Marcken 1996), is that the learner views language as constructed by an underlying process which concatenates words to generate utterances. By noticing sub-sequences of speech which often recur, the learner can detect common words and begin to segment fluent speech at word boundaries.

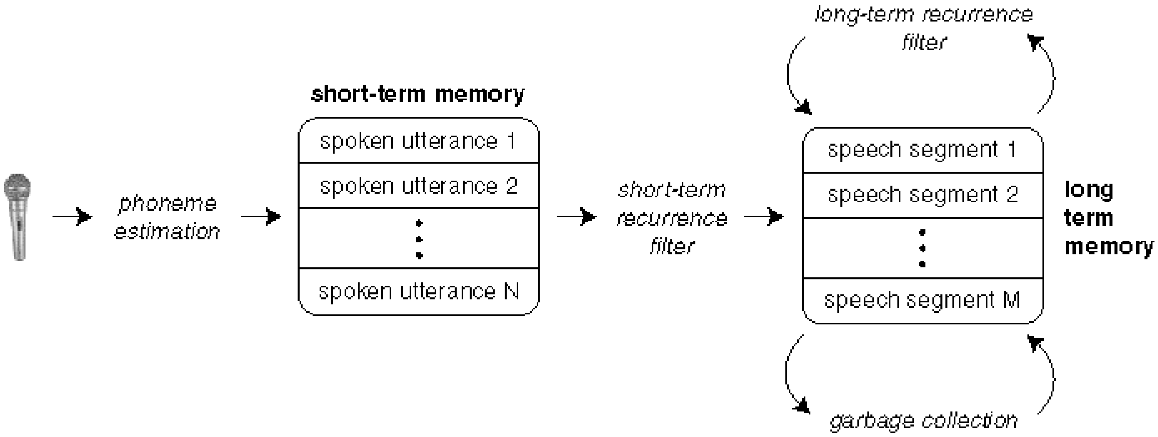

Figure 3: The RECUR model. Acoustic waveforms recorded by a microphone are converted to phoneme probabilities. Utterances are buffered by a short term memory (STM) and provide input to a recurrence filter which searches for repeated sequences of speech within the STM. The result is a set of speech segments which are stored in the long term memory (LTM). A second recurrence filter searches for entries in LTM which are repeated often across long spans of time. Such repetitions are evidence that the segment represents a word of the target language and is retained in LTM. A garbage collection process periodically removes segments from LTM which fail to pass through the long-term recurrence filter.

Infants are unlikely to search for all possible matches of speech segments across all spoken utterances which they have heard. Such recurrence analysis would require huge amounts of memory for verbatim speech, and would demand impractical computational resources. As suggested by theories of human memory (Miller 1956),our model eases the resource requirements by first searching for recurrent phonemic sequences in a short term window of input. The model performed an exhaustive search for repeated segments in the STM each time a new utterance was added. Recurrent speech sequences were extracted from the STM and copied into LTM. A second recurrence detector compared all LTM segments to one another using the same acoustic distance metric used on the STM. Segments in the LTM which were phonemically similar to many other speech segments in LTM were retained as reliable word candidates. Periodically, unlikely hypotheses which did not match other entries in the LTM were removed by a garbage collection process.

Evaluation

We evaluated the models by collecting speech and images similar to what an infant might observe during natural play with caregivers. Each model was then effectively “put in the infant’s place” to test whether it would learn words similar to what an infant might be expected to learn. A study involving six caregivers and their prelinguistic infants was conducted to gather a corpus of infant-directed speech. The participants were asked to engage in play centered around seven types of objects. The speech was then coupled with sets of images of these objects and used as input for the model.

All six participants were female and responded to a classified advertisement placed in a local newspaper. The infants (five males, one female) ranged in age from 8-11 months. Participants were asked to interact naturally with their infants while playing with a set of age-appropriate objects. We chose seven classes of objects commonly named in early infant speech (Huttenlocher & Smiley 1994): balls, toy dogs, shoes, keys, toy horses, toy cars, and toy trucks. A total of 42 objects, six objects from each class, were obtained.

Each caregiver participated in six sessions of play with their infants over a two day period. For each of the six sessions, participants were provided with a set of seven objects, one from each of the seven object classes. The order in which object sets were provided was randomized across participants. The objects were placed in a box marked ``in-box" at the start of each session. Participants were asked to take out one object at a time, play with it, and then return it to an ``out-box". They were *not* told to teach their infants words. Participants were free to choose the order in which objects were selected for play, and the duration of play with each object.

The speech recording of each session was automatically segmented into utterances. The robotic armature was used to gather a set of images of each object from various angles, approximating what the infant saw during play sessions. A set of 209 images were captured of each object from varying perspectives resulting in a database of 8,778 images. For each utterance, we randomly selected 15 images of the object which was in play when the utterance was recorded. These 15 images were paired with the utterance and presented as input to the models. The models were run separately on the speech recordings of each caregiver. For each caregiver, the model generated a set of output speech segments (RECUR), or speech-shape pairs (CELL). By testing both models with identical spoken input, we were able to determine the value, if any, in additionally providing visual context.

For each speech segment identified by the model, we evaluated two measures of accuracy. Measure 1 assessed segmentation accuracy: Did the segment start and end at English word boundaries? Measure 2 assessed word discovery: Did the speech segment correspond to a single English word? We considered words with attached articles and inflections as acceptable by Measure 2. We also allowed initial and final consonant errors for Measure 2, but not Measure 1.

Measure 1, segmentation accuracy, posed an extremely difficult challenge when dealing with acoustic data. For RECUR, 7 ± 5 % of segment boundaries corresponded to boundaries of English words. In contrast, 28 ± 6 % of segment boundaries extracted by CELL were chosen at actual words boundaries. For Measure 2, word discovery, almost three out of four speech segments (72 ± 8 %) acquired by CELL were single, complete English words. In contrast, performance for RECUR dropped to 31 ± 8 %.

The output of CELL was measured for semantic accuracy (Measure 3): How often did an output speech segment pass Measure 2 and also get paired with a semantically appropriate visual prototype? Since RECUR did not process visual data, this measure could not be meaningfully applied to its output. CELL achieved 57 ± 10 % on this measure. This result shows that the visual semantics derived from context was connected to appropriate words in a significant number of cases (random guessing of the meaning of a speech segment would yield a maximum of 14%). Recall from earlier that we had set out to address three problems of early word learning: word discovery, contextual categorization, and establishing word-context correspondences. CELL achieves each in a unified framework. Speech segments corresponding to prototypes of spoken words are extracted from continuous speech. Visual prototypes corresponding to words are identified and associated, in many cases, with appropriate spoken words.

Conclusion

CELL was able to learn spoken words and their visual groundings from multimodal sensory data. Comparisons with the acoustic-only RECUR model demonstrate the benefit of incorporating cross-modal information into the word discovery process. The inter-modal structure lead to a 2.3-fold increase in word discovery accuracy compared with analyzing structure within the acoustic channel alone. For speech segmentation, the improvement was even larger, four-fold. These result have implications for understanding language acquisition in infants. Rather than segment speech as a preparatory step towards acquiring sound-to-meaning mappings, a more efficient strategy could be to combine the segmentation process with the mapping process at the earliest stages of language learning. The additional structure from the contextual channels may accelerate the overall process of early lexical acquisition.

We often think of learning as consisting of discrete stages. In the case of learning early words, two natural alternatives come to mind. On one hand, perhaps infants learn early concepts and then look for spoken labels to fit these concepts. On the other hand, they might first learn salient speech sequences and then look for their referents. Our model and experiments verify that a more closely knit process in which the two "stages" in fact occur together is advantageous for the learner. By taking this approach, the learner is able to leverage information captured in the structure between streams of input.

Acknowledgements

This work has benefited greatly from discussions with Alex Pentland, Steven Pinker, Allen Gorin, Rupal Patel, and Bernt Schiele.

References

Bolinger, D.L. and Gertsman, L.J. Disjuncture as a cue to constraints, *Word*, 13, 246-255 (1957).

Brent, M.R. An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery, *Machine Learning*, 1999.

Cover, T.M. and Thomas, J.A. *Elements of Information Theory*, Wiley-Interscience, New York (1991).

Cutler, A. Segmentation problems, rhythmic solutions. In L. Gleitman and B. Landau, editors, The Acquisition of the Lexicon, chapter 2, pages 81-104. MIT Press, Cambridge, MA, 1991.

De Marcken, C. *Unsupervised Language Acquisition*, Ph.D. thesis, Massachusetts Institute of Technology (1996).

Friederici, A.D. and Wessels, J.M.I. Phonotactic knowledge and its use in infant speech perception, *Perception and Psychophysics*, 54, 287-295 (1993).

Huttenlocher, J. and Smiley, P. Early word meanings: the case of object names, in *Language Acquisition: Core Readings*, P. Bloom (ed.), MIT Press, Cambridge, MA, 222-24 (1994).

Julesz, B. *Foundations of Cyclopean Perception*, University of Chicago Press, Chicago, (1971).

Kuhl, P.K., Williams, K.A., Lacerda, F, Stevens, K.N., and Lindblom, B. Linguistic experiences later phonetic perception in infants by 6 months of age, *Science*, 255, 606-608 (1992).

Miller, G.A. The magical number seven plus or minus two: Some limits on our capacity for processing information, Psych. Rev., 63, 81-9 (1956).

Robinson, T. An Application of Recurrent Nets to Phone Probability Estimation, IEEE Trans. on Neural Networks. 5, (1994).

Roy, D., M. Hlavac, M. Umaschi, T. Jebara, J. Cassell, and A. Pentland. (1997). Toco the toucan: A synthetic character guided by perception, emotion, and story. In *Visual Proceedings of Siggraph*, Los Angeles, CA. ACM Siggraph.

Roy, D. *Learning Words from Sights and Sounds: A Computational Model*. Ph.D. thesis, Massachusetts Institute of Technology (1999).

Roy, D. Integration of speech and vision using mutual information. To appear in *Proc. Int. Conf. Speech and Signal Processing*, Istanbul, Turkey, 2000.

Saffran, J. and Aslin, R. and Newport, E. Statistical learning by 8-month-old infants, *Science*, 274, 1926-1928 (1996).

Schiele, B. and Crowley, J.L. Probabilistic Object Recognition using Multidimensional Receptive Field Histograms, *Proc. of the 13th Int. Conf. on Pat. Recognition*, (1996).

Seneff, S. and Zue, V. Transcription and Alignment of the TIMIT Database*, Proc. of the Second Symposium on Advanced Man-Machine Interface through Spoken Language*, 1988.

Snow, C.E. Mothers' speech research: from input to interaction. In C.E. Snow and C.A. Ferguson (eds.), *Talking to children: language input and acquisition*. Cambridge University Press, 1977.

Werker, J.F. and Tees, R.C. Developmental changes across childhood in the perception of non-native speech sounds, *Canadian J. Psych*., 37, 278-286 (1983).