Cañamero, L., Oudeyer, P.-Y., and Balkenius, C. (Eds) Proceedings of the Ninth International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems. Lund University Cognitive Studies, 146.

An Unsupervised Model of Infant Acoustic Speech Segmentation

Matthew Miller and Alexander Stoytchev
Developmental Robotics Laboratory
Iowa State University
{mamille|alexs}@iastate.edu

Abstract

There is a long standing hypothesis in Developmental Psychology that children use statistical information to segment acoustic speech streams into words. Additionally, several experiments have demonstrated that infants are able to find word breaks using distributional cues. In this paper we propose an algorithm for the unsupervised segmentation of audio speech, based on the Voting Experts (*VE*) algorithm. We show that this algorithm can reproduce results obtained from segmentation experiments performed with 8-month-old infants.

1. Introduction

Spoken human language contains no analogue to the spaces placed between written words. The pauses that do exist in audio speech appear between phrases, when the speaker takes a breath, or when the airflow is stopped in the pronunciation of certain consonants. The sounds that are separated by these pauses are rarely composed of a single word, and there are no universal markers to indicate where those single words might be (Klatt, 1979). However, when we hear our native language, we hear discrete words. We unconsciously break the stream into its constituents, rendering it comprehensible. This is possible because we know the language, and are familiar with the large lexicon of words we might expect to hear. When confronted with a novel word, we need only segment the words before and after it to identify it as a brand new token.

Infants, however, do not share this luxury. They must learn to segment their mother's tongue from scratch. Every word is a novel word, and their lexicon starts off empty. Fortunately, human beings have an apparently innate ability to use statistical information to segment continuous spoken speech into words, and that ability is present in infants as young as 8 months old. Apparently, they can perform this task without any feedback or other salient cues as to the locations of word breaks (Saffran et al., 1996) (Saffran et al., 1999).

An accurate characterization of this ability would presumably be theoretically and practically advantageous. Along those lines, this paper proposes a method for the unsupervised segmentation of spoken speech, based on an algorithm designed to segment discrete time series into meaningful episodes. We suggest that our model may capture part of the human process of speech segmentation. To substantiate our claim, we replicate an experiment that was performed on 8-month-old infants, and show that our algorithm performs similarly to the children.

2. Related Work

There are two main fields that are related to this topic. The first is the study of the speech segmentation methods that are used by infants. This constitutes a very broad area of research, with many subfields. This work is most related to the study of statistical learning in developmental psychology, which focuses on infants' ability to use statistical cues to segment language streams. These studies are the direct inspiration for this line of research, but they suggest no practical algorithm for replicating the results they have observed. The second related field of research pertains to algorithms for the segmentation of time series data. These studies suffer from the opposite problem. That is, there are many strategies by which to segment data, but not many that serve as a plausible model of infant segmentation.

2.1 Statistical Learning

The idea that infants use statistical cues to segment speech streams is very old (Harris, 1955). Specifically, the canonical theory is that they use the transitional probabilities between syllables as an indicator of word boundaries. Suppose that α and β are syllables in some language. Then the transitional probability $TP(\alpha \rightarrow \beta)$ is the probability that β follows

α when α appears in the speech stream. It stands to reason that syllables that appear together inside of a word would have a higher transitional probability than those that do not. Therefore, the argument goes, the transitional probabilities between syllables inside of words should be high, but the TP between syllables that cross a word boundary should be low. Hence, a child might easily segment a sequence of syllables by noting whenever the transitional probability dips down low.

This is precisely the strategy suggested in a series of experiments performed by Saffran *et. al.* (Saffran et al., 1996) (Saffran et al., 1997) (Saffran et al., 1999). The first of these experiments demonstrated that 8-month-old infants can, in fact, segment words based solely on statistical information. The children were played an artificially generated acoustic stream composed of the words *tupiro*, *golabu*, *bidaku* and *padoti* repeated in random order. After two minutes they were played a second stream consisting of a single word repeated over and over. Half of the time the word was from the original language, and the other half of the time it was a novel word, generated from the same syllables. The stimulus streams had no audible breaks between the words, no variation in pitch or meter, and no other cues as to the word breaks. The only clue was the transitional probability between the syllables. Inside of words it was always 100%, but between words it dropped to 25%. The stimulus stream was constructed specifically to be segmentable by the TP strategy. And the amazing result was that, after only two minutes, the infants were able to tell the novel words from the old.

The results of these experiments were taken as evidence that human infants really do pay attention to the transitional probabilities between syllables, and that they use them to segment audio speech. However, that's not really what these experiments showed. They showed that infants can segment audio speech using *some kind* of statistical model, and that it is powerful enough to work on the stimulus stream they were presented. Dips in inter-syllable transition probability were the simplest cue that they could have used to segment the sequence, but virtually any sophisticated model should have picked up this very simple pattern. And there is significant evidence to suggest that infants, in fact, are not using TPs to do this.

Most dramatically, multiple studies have showed that the direct application of the TP strategy performs poorly when used to segment phonetic transcripts of speech (Cairns and Shillcock, 1997) (Gambell and Yang, 2008). This exposes several of the weaknesses of the traditional statistical learning approach. First of all, a very high percentage of common words contain only one syllable. It is therefore impossible for there to be a TP valley on both sides of the word. Moreover, the original conclusion that word-internal transitions should have higher probabilities than word-external ones is not always true in practice. Often, the last syllable of one word and the first syllable of the next happen to form a perfectly common pair. Similarly, many words contain syllable combinations that are, in general, rare (perhaps only appearing in a handful of words). The difference in single-syllable TP inside of and between words is more of a trend than a reliable rule.

2.2 Segmentation Algorithms

Most of the algorithms mentioned in this section are used to segment discrete token sequences (*i.e.*, they segment text - or text based phonemic transcripts of speech). This paper describes an algorithm that runs on real audio, and is able to perform the unsupervised segmentation of individual words from acoustic speech streams. So, in some sense, we are comparing apples and oranges. However, this previous work is certainly related, since it is also inspired by developmental psychology, and intends to accomplish roughly the same task.

There exist a wide variety of algorithms capable of segmenting discrete time series into meaningful "chunks." For instance, compression algorithms that find minimum description lengths can often be coerced into segmentation by using whatever encoding they perform (Nevill-Manning and Witten, 1997) (Cohen et al., 2007). Several studies have attempted to train Neural Nets to predict the subsequent phoneme given the last few, and induce breaks whenever the prediction is uncertain (Elman, 1990) (Cairns and Shillcock, 1997). Gambell and Yang suggested a method of segmenting speech by assuming that every word contains a single stressed syllable (Gambell and Yang, 2008). They reported very good results on the CHILDES dataset, transcribed to phonemes and then concatenated into syllables. Michael Brent published a thorough survey of many different strategies for attacking this problem (Brent, 1999b). In fact, his own algorithm has set the bar for the unsupervised segmentation of phonemic transcripts of infant directed speech (Brent, 1999a). It incrementally builds a lexicon and induces maximum likelihood parses in short phrases. Using this strategy, Brent was able to segment phonemic transcripts of child directed speech with precision and recall above 80%. This remains the best performing algorithm on this type of data.

However, Brent's algorithm pays no attention to statistical regularities in phoneme sequences, and typically builds very large lexicons with many wrong words. For instance, this algorithm would be incapable of segmenting the stimulus streams used in the statistical learning experiments, since they contained

no phrase boundaries. This demonstrates that, while some kind of bootstrapping, lexicon-based segmentation method might be useful, it does not completely model the human system. Perhaps infants use a similar process as part of their strategy, but they are also sensitive to statistical cues.

Recently the ACORNS project has been created to investigate human language acquisition (Boves et al., 2007). This research is unique, in that it attempts to learn the grounded meaning of words in an unsupervised way. However, the automatic segmentation of speech into words is a secondary goal to the extraction of semantic meaning. These two strategies are certainly complimentary, and children must perform both of these tasks in order to acquire language. In this paper, we do not address word learning, but instead focus entirely on unsupervised segmentation. Our goal is to introduce a unique unsupervised method for segmenting continuous data streams, to apply the method to speech, and suggest that such a model might characterize the statistical segmentation abilities of human infants.

2.3 Voting Experts

Voting Experts (*VE*) is an algorithm for the unsupervised segmentation of discrete time series into meaningful episodes (Cohen et al., 2007). It is a purely distributional algorithm, in that it relies solely on statistics calculated from the time series itself. *VE* has demonstrated an ability to accurately segment text, phonetic transcripts, vertical pixel columns scanned from text, discrete robot sensor data and even a text transcript of the acoustic streams used in this paper (Miller and Stoytchev, 2008a) (Cohen et al., 2007). It's based on the hypothesis that natural breaks in a sequence are usually accompanied by two information theoretic signatures. These are low *internal entropy* of chunks, and high *boundary entropy* between chunks.

In this context, the internal entropy of a chunk is simply its Shannon information, or the negative log of its probability (Shannon, 1951). So the higher the probability of a chunk, the lower its internal entropy. We can calculate the probability of a short sequence of tokens by observing how often that sequence appears in a longer time series. So, essentially, this marker picks out short sequences of tokens that appear often.

Boundary entropy is the uncertainty at the boundary of a chunk. Given a sequence of tokens, the boundary entropy is the expected information gain of being told the next token in the time series. This is calculated as

$$H_B(c) = -\sum_{h=1}^{m} P(h \mid c) log(P(h \mid c))$$

where c is this given sequence of tokens, $P(h \mid c)$ is the conditional probability of symbol h following c, and m is the number of tokens in the alphabet. Well formed chunks are groups of tokens that are found together in many different circumstances, so they are somewhat unrelated to the surrounding elements. If the boundary entropy of a subsequence is high it means that there is no particular token that is very likely to follow that subsequence. In other words, the next token is unpredictable.

In order to segment a discrete time series, *VE* preprocesses the series to build an n-gram trie, which represents all its subsequences of length less than or equal to n. It then passes a sliding window of length n over the series. At each window location, two "experts" use the trie to vote on how they would break the contents of the window. One expert votes to minimize the internal entropy of the induced chunks, and the other votes to maximize the entropy at the break. After all the votes have been cast, the sequence is broken at the "peaks" - locations that received more votes than their neighbors, so long as the total votes at the location exceeded a threshold V_t. For all of our experiments we chose $n = 7$, and we varied V_t over a range of values. The effect of this variation will be discussed later, and evident in the results of our experiments. The choice of n roughly approximates the expected length of an individual "chunk." This algorithm runs in linear time with respect to the length of the sequence, and can therefore be used to segment very long sequences. For further technical details of *VE*, or a more in-depth discussion of the roles of V_t and n, see (Cohen et al., 2007).

This model bears a strong resemblance to the statistical learning approach mentioned before. If the conditional probability between each syllable within a word is high, then by definition the internal entropy of the word is low. But instead of evaluating each transitional probability in isolation, *VE* looks for short sequences of tokens where all of the TPs are high. Similarly, the boundary entropy of a sequence is high precisely when there is no particular token that is very likely to come next. However, instead of focusing on the transition probability between two syllables that happened to be adjacent, *VE* looks at whether the TP is *expected* to be low. This is an important difference, and it solves one of the major problems with the transitional probability approach. When the last syllable of one word and the first syllable of the next happen to form a likely pair, the TP based approach fails. But *VE* isn't affected when the TP at the word boundary is high, as long as the next token is unpredictable based on several previous tokens. This extra power is afforded by the use of the more sophisticated entropy metrics. Moreover, the model should still be extremely sensitive to the transitional probability cues, since the entropy cues must be present wherever the TP cues are.

In this paper we extend *VE* to work on audio data. We then use this algorithm to reproduce Saffran *et al.'s* original experiments. *VE* might not be the best possible unsupervised distributional segmentation algorithm, but it is certainly a powerful one. Additionally, the complexity of its metrics seems close to the horizon of biological plausibility. It is not unrealistic to think that humans naturally pick out commonly recurring sequences of sounds, and tend to place breaks at moments of unpredictability. Accordingly, we suggest that *VE* is a strong candidate for a usable model of the human distributional segmentation mechanism.

3. Datasets

We obtained two stimulus streams from the original infant speech segmentation experiments (Saffran et al., 1996). Each audio stream is about 60 seconds long and contains roughly 90 "words." The first stream (stream A) was composed, as described above, of randomly ordered instances of the four words *tupiro*, *golabu*, *bidaku* and *padoti*. The second stream (stream B) was composed of random instances of the words *tilado*, *dapiku*, *pagotu* and *burobi*. The second language is composed of the same syllables as the first, but arranged so that the concatenation of words in either language cannot produce a word from the other. So in some sense these two audio streams are disjoint.

In the original experiment, the infants were played a stream created in the same way as stream A, and then tested on a single word repeated over and over. This method is useful when evaluating infants because it is simple. However, we can perform a more thorough evaluation of our model since it produces explicit break locations. We found it more informative to test our model by training it on one stimulus stream and then testing it on the other. This provides more information on the performance of the model, but the results can clearly be compared to those of the infant experiments.

In order to evaluate the segmentations induced by our algorithm, we manually recorded the timestamps of all of the word boundaries in the two stimulus streams. It is impossible for this process to be absolutely precise, since spoken audio is not actually composed of distinct phonemes, and word breaks are not always marked by silence. The sound morphs from one allophone to the next, providing few clear boundaries. However, the speech in the streams used by Saffran *et al.* is very regular, which allowed us to consistently place breaks at the same location in each word. The waveform in between each word pair was identical every time it appeared, since it was generated artificially. The beginning and ending of each word was verified acoustically once, and then the boundaries could be placed in exactly the same location for each instance of each word. The resulting "answer keys" were therefore consistent, and as close to the ground truth as possible.

4. Audio Segmentation Algorithm

The raw audio of both stimulus streams was converted into a sequence of Mel-cepstral feature vectors, along with their first and second order time derivatives and their log energy (Davis and Mermelstein, 1980). The standard 13 cepstral features were used, so that each time slice of the audio was represented by a 42-dimensional real valued feature vector. That's 13 cepstral features, 13 first order and 13 second order time derivatives, and the log energy of each. This is a standard method of feature extraction for speech processing, and it was performed using the Matlab package "Voicebox."

Since *VE* is designed to work on a sequence of tokens, these feature vectors must be quantized into a manageable alphabet. A common technique in automatic speech recognition is to use Hidden Markov Models with continuous observation densities to recognize phonemes (Rabiner, 1990). We will draw inspiration from these models, however we cannot apply the techniques exactly. In the infant experiments the children learned to segment novel language streams in a completely unsupervised way. Therefore, any model of this process must also be entirely unsupervised. These HMMs are typically trained on labeled data, disqualifying them as plausible models. Specifically, a separate Markov chain is typically trained to represent each phoneme in the language. The models are built using a large set of hand-labeled instances of each phoneme, and then their parameters are improved by bootstrapping over a large audio corpus. Instead, we will suggest an unsupervised model that can convert an audio stream into a state sequence suitable for segmentation, but one that does not necessarily correspond to the phoneme sequence as a human would label it.

4.1 Unsupervised Acoustic Model

The critical observation is that we don't necessarily need a sequence that corresponds to the true phonemes of the language. All that's needed is a model that decomposes an audio stream into a sequence of its most salient acoustic features. These may or may not correspond to the "phonemes" as a human might label them. But that is irrelevant, at least as far as *VE* is concerned.

Just such a model was suggested by (Iwahashi, 2006), and implemented by (Brandl et al., 2008). We used a version of that model in this work. Each phoneme was represented using a 3-node Markov chain with Bakis-topology, with the observation probability density of each state represented by a mixture of Gaussian functions

(Rabiner, 1990). In order to train these models without labeled data, we first trained a completely connected Markov network containing 10 Gaussian mixture states on the acoustic stream. The parameters of the network were initialized using k-means, and then optimized using EM, so no labeled data was required. Then, we stochastically sampled paths of length 3 through that network based on the learned transition probabilities. The m most common paths were used to initialize m 3-node Markov chains. The last state of each chain was connected to the first state of every other chain, including itself, initialized with uniform transition probability. The parameters of this larger Markov model were then optimized over the corpus using EM.

In one implementation, m was set using the Akaike information criterion (Brandl et al., 2008). Instead we used $m = 10$ to build the models used in this paper. We varied this parameter, and found that it did not have a strong effect on the performance of the model on this task. The results of that evaluation are not included for space considerations. However, if this model were to be applied to a larger or more complex dataset, such an evaluation would certainly be necessary.

4.2 Segmentation

Given a model as described above and an acoustic stream for segmentation, we converted the stream into a state sequence using Viterbi decoding. The state sequence was simplified by assuming that all nodes from the same Markov chain were equivalent. So instead of a sequence of nodes in the HMM, the stream was represented as a sequence of 3-node Markov chain labels. However, this created sequences with long stretches of the same label repeated over and over. These repeated labels were collapsed into a single token. So the final token sequence represented the order in which these chains were visited in the decoding of the stimulus stream, with no information about how long the sound stayed in the same chain. If the chains corresponded to the phonemes of the language, as they do in more typical acoustic models, the result would be a transcription of the spoken phonemes of the stream. The idea is that the unsupervised model approximates the phoneme sequence, but perhaps extracts a slightly different set of fundamental sounds.

We ran *VE* on the resulting label sequence. *VE* placed breaks at locations of low internal entropy and high boundary entropy. Then, after accounting for the collapsed (*i.e.*, repeated) states, it produced the time stamps of all of the induced break locations in each audio stream. These breaks were then checked against the answer keys that had been manually created for each stimulus stream (See Figure 1).

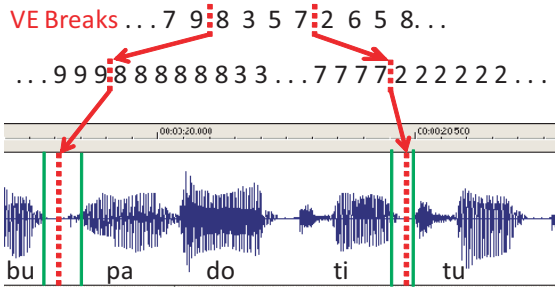

Figure 1: Evaluation of the breaks induced by *VE*. Each break is mapped to its location in the expanded state sequence, which corresponds to a timestamp in the audio stream. The break counts as correct if it falls within the marked boundary between two words. The states are represented by their numeric index in the Markov model.

5. Evaluation Methodology

In order for an induced break to count as a correct break, it had to be placed between the specified end of the previous word and the beginning of the next one, within an error of one time slice. The feature vectors that composed the audio stream were calculated using a window that was 0.016 seconds wide with a 50% overlap. This means that the additional time slice allowed at each boundary increased the break window by 0.008 seconds. This leeway was provided to compensate for labeling errors or other boundary conditions.

An induced break was counted as breaking two words if it was placed anywhere in the window between them. Both stimulus streams were 61.2 seconds long. Stimulus stream A contained approximately 7.7 seconds of "break" time, and stream B contained 7.2 seconds. The reason for the discrepancy is that the different pronunciations of the first and last syllables of the words in each stream led to slightly different amounts of time between them. It should be noted that these "breaks" are not perceivable when listening to the stream, and are no longer than the space between the phonemes within words.

Unfortunately these boundaries make it easier for the algorithm to accidentally induce a break between two words. Thus, even random breaks will be counted as correct some of the time. Accordingly, we used a Monte Carlo method to simulate random segmentations for each experiment. Each reported result is accompanied by the results of inducing a large number of random segmentations, each one having the same number of induced breaks as the algorithm produced. The random breaks were induced in the same compressed state sequence used by *VE*, and were evaluated in the same manner. These random trials are averaged and provide a baseline from which to evaluate the algorithm.

The quality of the segmentation is evaluated based

on the accuracy, hit-rate and f-measure of the induced breaks. In this case, accuracy is the percentage of induced breaks that are correct, hit-rate is the percentage of true breaks found by the algorithm, and the f-measure is the harmonic mean of the two, given by

$$\text{f-measure} = \frac{2*accuracy*hitrate}{accuracy+hitrate}$$

The f-measure is treated as most important, since it strikes a balance between the other two. It's possible to increase the accuracy of the segmentation by inducing fewer breaks, but being more confident about those that are induced. However, this will lower the hit-rate. Similarly we can raise the hit-rate by inducing more breaks, but this will lower the accuracy. The Voting Experts algorithm lets us explicitly make this tradeoff by adjusting the threshold V_t for the minimum number of votes required to induce a break at a location. All three of these metrics will be reported for each of our experiments. Additionally, the experiments will be repeated for a range of thresholds V_t, and the sensitivity of these metrics to variation in that threshold will be demonstrated.

6. Experimental Results

We have outlined a general, unsupervised algorithm for the segmentation of an audio stream. First, convert the stream into an appropriate sequence of feature vectors - in our case the Mel-cepstrum. Then train an unsupervised Gaussian Mixture HMM (GMHMM) on the sequence as described above. Use this model to produce a sequence of Markov chain labels based on the audio stream. Finally, collapse the repeated labels in this sequence and run *VE* on the result.

This algorithm constitutes a very basic application of the *VE* model to a real audio stream. The first question is whether this can induce an accurate segmentation. The second question is whether we can use this system to model the human segmentation mechanism. The following experiments were designed to answer both of these questions.

Experiment 1: We ran the segmentation process described above separately on each stimulus stream (A and B). We then compared the induced breaks to the true breaks for each stimulus stream. The results are shown in Figure 2.

The segmentation induced on both audio streams was significantly more accurate than chance. Clearly this model is capable of segmenting the given stimulus streams. These results are even more surprising when considering that these models were each trained on only one minute of audio. Presumably infants might be better equipped to perform this task since they have the advantage of a previously trained acoustic model. They do not have to learn it from scratch in just one minute as we have done here. But even with that limitation, *VE* performs very well.

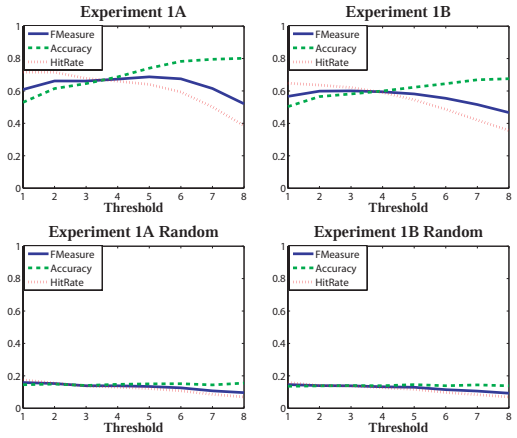

Figure 2: The F-measure, accuracy and hit-rate of the segmentation of both stimulus streams in Experiment 1, along with the performance of random segmentations on both datasets.

It should be noted that the initialization of the acoustic models is a stochastic process, and leads to a unique model every time. The EM algorithm does not necessarily find a global optimum for the model parameters, but only a local maximum. Therefore, the model should not be evaluated based on a single instantiation, but rather based on several trials. Accordingly, we trained 10 different acoustic models on each of the two stimulus streams. All three experiments were performed 10 different times with 10 different pairs of models. The results were averaged to produce the results reported.

Additionally, the segmentation step, where *VE* was run on the token sequence, was repeated for different threshold values ranging from 1 to 8 for each experiment. Notice the tradeoff between accuracy and hit-rate as V_t varies. The f-measure, accuracy and hit-rate are reported both for the aggregate over all 10 models, as well as for the random trials over the same data. For each trial that was done with a single model, 10 random trials were performed. So, overall, 100 random trials were performed in each experiment for each stimulus stream.

Experiment 2: The point of this experiment is to demonstrate that an acoustic model trained on stimulus stream A can still be used to segment the audio from stream B, and vice versa. The two streams are composed of the same set of syllables. The only difference is the order in which the syllables are spoken, which may produce some interaction effects that the GMHMM cannot model. However, most of the sounds are the same. So, for instance, the tokenization of stream B by an acoustic model trained on stream A should still be useful for inducing a segmentation on B.

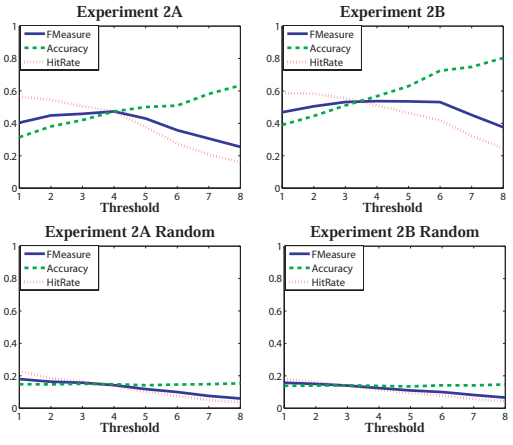

Figure 3: The F-measure, accuracy and hit-rate of the segmentation of both stimulus streams in Experiment 2. Once again, the performance of random segmentation is also shown.

To demonstrate this, we trained an acoustic model on each stream to obtain GMHMM_A and GMHMM_B. Then we used GMHMM_A to tokenize the feature vectors from stimulus stream B and GMHMM_B to tokenize stream A. Then we trained a *VE* model on each of the token sequences and induced a segmentation. Once again we used the true breaks to evaluate the results (see Figure 3).

There is a slight drop in both the accuracy and hit rate of each segmentation in this experiment. However, in each case the algorithm still performed much better than chance. There is not a tremendous loss due to the unmodeled interaction of the diphones in the stimulus streams. This fact is important in understanding the results of experiment 3.

Experiment 3: This experiment is intended to replicate the results of the infant studies. In those experiments, the children listened to one stimulus stream, and were then presented a novel token from the second stream. Similarly, in this experiment, our model is trained on one stimulus stream, and then used to segment the other. That is, the GMHMM and the statistical model of *VE* (the experts) are trained on stream A, and then that model is used to segment stream B and vice versa.

Figure 4 shows that the algorithm is almost completely unable to induce a segmentation. It performs only slightly better than chance, and this is most likely due to its ability to pick out syllables. From the results of experiment 2 we can conclude that the poor performance is not the fault of the acoustic model. Instead, the language model trained on one language is insufficient to induce a segmentation in another.

As the threshold increases, the algorithm induces very few breaks. When V_t is higher than 5, almost no breaks are induced (*e.g.*, no breaks were induced at all when $V_t = 8$). This explains why the accuracy becomes erratic at higher threshold levels, and the hit-rate drops very low. The random segmentations only contained as many breaks as the algorithm induced, so the random hit-rate drops as well. The fact that not very many breaks were induced indicates that the experts did not vote for the same break locations very often. They could not agree on suitable breaking points, and therefore did not create many breaks. Essentially, the algorithm was confused.

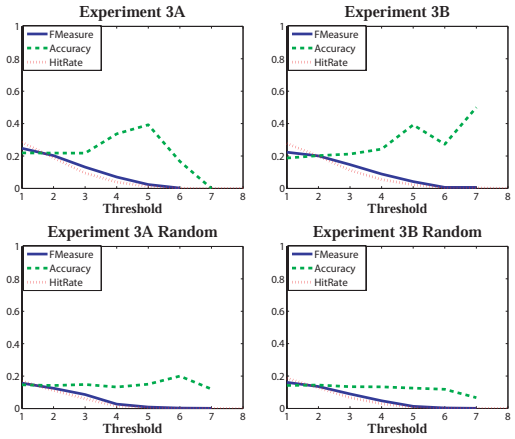

Figure 4: The F-measure, accuracy and hit-rate of the segmentation of both stimulus streams in Experiment 3, along with the results of the random segmentation.

This corresponds precisely with the situation of the 8-month-old who listens to stimulus stream A, and then hears a novel word from stream B. The child has learned the sounds present in the stream, and has learned a statistical model that characterizes it. Then, suddenly, that model is violated. The child is initially unable to use the old model to "understand" the novel word, and therefore becomes confused.

7. Conclusions and Future Work

We have described an unsupervised technique for transforming spoken audio into a discrete sequence of tokens suitable for segmentation by the Voting Experts algorithm. This algorithm is novel in its application to real audio, and its reliance on simple but powerful information theoretic cues. We have shown that the *VE* model is capable of inducing an accurate segmentation on an audio stimulus stream with very limited training data. Finally, we have shown that the behavior of this model mimics the behavior of 8-month-old infants. This should be counted as a small victory for *VE* as a model of human segmentation. It also demonstrates that distributional cues can be used to segment audio streams. Specifically, the low internal entropy and high boundary entropy of chunks provide sufficient markers to do so.

The psychological studies that have explored infant statistical learning have used stimulus streams

that could be segmented using transitional probabilities. Infants can segment these simple streams, but the full extent of their capabilities remains unknown. *VE* can segment the same stimulus streams, and therefore is not disqualified as a possible model of the human distributional speech segmentation mechanism. If an algorithm can pass that test, it's at least a plausible candidate. However, this may be an easier task than children face with natural language.

It is simply unknown how important a role distributional segmentation really plays in the acquisition of language, and how sophisticated that mechanism is. Presumably it is significantly useful, or else children wouldn't demonstrate this ability at such a young age. Since some studies have shown that the simple statistical learning approaches are not sufficient to segment natural language, we should conclude that infants have a more sophisticated strategy. *VE* has the advantage of being able to segment many different kinds of speech, including natural language phoneme sequences (Miller and Stoytchev, 2008a). This makes it a much more attractive candidate for modeling human segmentation, since the approaches based on transitional probabilities have not done the same. The next logical step is to use this model on a natural language corpus to see how effective it can really be.

Acknowledgments

An earlier version of this paper, with a much simpler acoustic model and less detailed analysis, was accepted into the NIPS 2008 Workshop on Speech and Language (Miller and Stoytchev, 2008b). We would like to thank the organizers and participants for the suggestions and feedback that helped improve our work. We would also like to thank Richard Aslin from the University of Rochester for providing us with the stimulus streams used in the original Saffran *et al.* experiments. Finally, we would like to thank Paul Cohen from the University of Arizona for generously providing the source code for the original Voting Experts algorithm.

References

Boves, L., ten Bosch, L., and Moore, R. (2007). ACORNS – towards computational modeling of communication and recognition skills. In *Proceedings of ICCI*.

Brandl, H., Wrede, B., Joublina, F., and Goerick, C. (2008). A self-referential childlike model to acquire phones, syllables and words from acoustic speech. In *Proceedings of the 7th IEEE International Conference on Development and Learning*, pages 31–36.

Brent, M. R. (1999a). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105.

Brent, M. R. (1999b). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Science*, 8(3):294–301.

Cairns, P. and Shillcock, R. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, 33:111–153.

Cohen, P., Adams, N., and Heeringa, B. (2007). Voting Experts: An unsupervised algorithm for segmenting sequences. *Journal of Intelligent Data Analysis*, 11(6):607–625.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):357–366.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Gambell, T. and Yang, C. (2008). Mechanisms and constraints in word segmentation. Manuscript, Yale University.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, 31.

Iwahashi, N. (2006). *Symbol Grounding and Beyond*, volume 4211/2006 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7:279–285.

Miller, M. and Stoytchev, A. (2008a). Hierarchical Voting Experts: An unsupervised algorithm for hierarchical sequence segmentation. In *Proceedings of the 7th IEEE International Conference on Development and Learning (ICDL)*.

Miller, M. and Stoytchev, A. (2008b). Unsupervised audio speech segmentation using the Voting Experts algorithm. In *NIPS Workshop on Speech and Language: Learning-based Methods and Systems*.

Nevill-Manning, C. and Witten, I. (1997). Identifying hierarchical structure in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, pages 7:67–82.

Rabiner, L. R. (1990). A tutorial on hidden markov models and selected applications in speech recognition. *Readings in speech recognition*, pages 267–296.

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70:27–52.

Saffran, J. R., Newport, E. L., Aslin, R. N., and Tunick, R. A. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, 8(2):101–105.

Shannon, C. (1951). Prediction and the entropy of printed english. Technical report, Bell System Technical Journal.