One-shot learning of generative speech concepts

Brenden M. Lake* Brain and Cognitive Sciences MIT

Chia-ying Lee* CSAIL MIT

James R. Glass CSAIL MIT

Joshua B. Tenenbaum Brain and Cognitive Sciences MIT

Abstract

One-shot learning – the human ability to learn a new concept from just one or a few examples – poses a challenge to traditional learning algorithms, although approaches based on Hierarchical Bayesian models and compositional representations have been making headway. This paper investigates how children and adults readily learn the spoken form of new words from one example – recognizing arbitrary instances of a novel phonological sequence, and excluding non-instances, regardless of speaker identity and acoustic variability. This is an essential step on the way to learning a word's meaning and learning to use it, and we develop a Hierarchical Bayesian acoustic model that can learn spoken words from one example, utilizing compositions of phoneme-like units that are the product of unsupervised learning. We compare people and computational models on one-shot classification and generation tasks with novel Japanese words, finding that the learned units play an important role in achieving good performance.

Keywords: one-shot learning; speech recognition; category learning; exemplar generation

Introduction

People can learn a new concept from just one or a few examples, making meaningful generalizations that go far beyond the observed data. Replicating this ability in machines has been challenging, since standard learning algorithms require tens, hundreds, or thousands of examples before reaching a high level of classification performance. Nonetheless, recent interest from cognitive science and machine learning has advanced our computational understanding of "one-shot learning," and several key themes have emerged. Probabilistic *generative models* can predict how people generalize from just one or a few examples, as shown for data lying in a low-dimensional space (Shepard, 1987; Tenenbaum & Griffiths, 2001). Another theme has developed around *learning-to-learn*, the idea that one-shot learning itself develops from previous learning with related concepts, and Hierarchical Bayesian (HB) models can learn-to-learn by highlighting the dimensions or features that are most important for generalization (Fei-Fei, Fergus, & Perona, 2006; Kemp, Perfors, & Tenenbaum, 2007; Salakhutdinov, Tenenbaum, & Torralba, 2012).

In this paper, we study the problem of learning new spoken words, an essential ingredient for language development. By one estimate, children learn an average of ten new words per day from the age of one to the end of high school (Bloom, 2000). For learning to proceed at such an astounding rate, children must be learning new words from very little data. Previous computational work has focused on the problem of learning the meaning of words from a few examples; for instance, upon hearing the word "elephant" paired with an exemplar, the child must decide which objects belong to the set of "elephants" and which do not (e.g., Xu & Tenenbaum, 2007). Related computational work has investigated other factors that contribute to learning word meaning, including learning-to-learn which features are important (Colunga & Smith, 2005; Kemp et al., 2007) and cross-situational word learning (Smith & Yu, 2008; Frank, Goodman, & Tenenbaum, 2009). But by any account, the acquisition of meaning is only possible because the child can also learn the *spoken word as a category*, mapping all instances (and excluding non-instances) of a word like "elephant" to the same phonological representation, regardless of speaker identify and other sources of acoustic variability. This is the focus of the current paper. Previous work has shown that children can do one-shot spoken word learning (Carey & Bartlett, 1978). When children (ages 3-4) were asked to bring over a "chromium" colored object, they seemed to flag the sound as a new word; some even later produced their own approximation of the word "chromium." Furthermore, acquiring new spoken words remains an important problem well into adulthood whether its learning a second language, a new name, or a new vocabulary word.

The goal of our work is twofold: to develop one-shot learning tasks that can compare people and models side-by-side, and to develop a computational model that performs well on these tasks. Since the tasks must contain novel words for both people and algorithms, we tested English speakers on their ability to learn Japanese words. This language pairing also offers an interesting test case for learning-to-learn through the transfer of phonetic structure, since the Japanese analogs to English phonemes fall roughly within a subset of English phonemes (Ohata, 2004).

Can the recent progress on models of one-shot learning be leveraged for learning new spoken words from raw speech? How could a generative model of a word be learned from just one example? Recent behavioral and computational work suggests that *compositionality*, combined with Hierarchical Bayesian modeling, can be a powerful way to build a "generative model for generative models" that supports one-shot learning (Lake, Salakhutdinov, & Tenenbaum, 2012; Lake et al., 2013). This idea was applied to the one-shot learning of handwritten characters, a similarly high-dimensional domain of natural concepts, using an "analysis-by-synthesis" approach. Given a raw image of a novel character, the model learns to represent it by a latent dynamic causal process, composed of pen strokes and their spatial relations (Fig. 1a). The sharing of stochastic motor primitives across concepts (Fig. 1a-i) provides a means of synthesizing new generative models out of pieces of existing ones (Fig. 1a-iii).

Compositional generative models are well-suited for the problem of spoken word acquisition, as they relate to classic

* The first two authors contributed equally to this work.

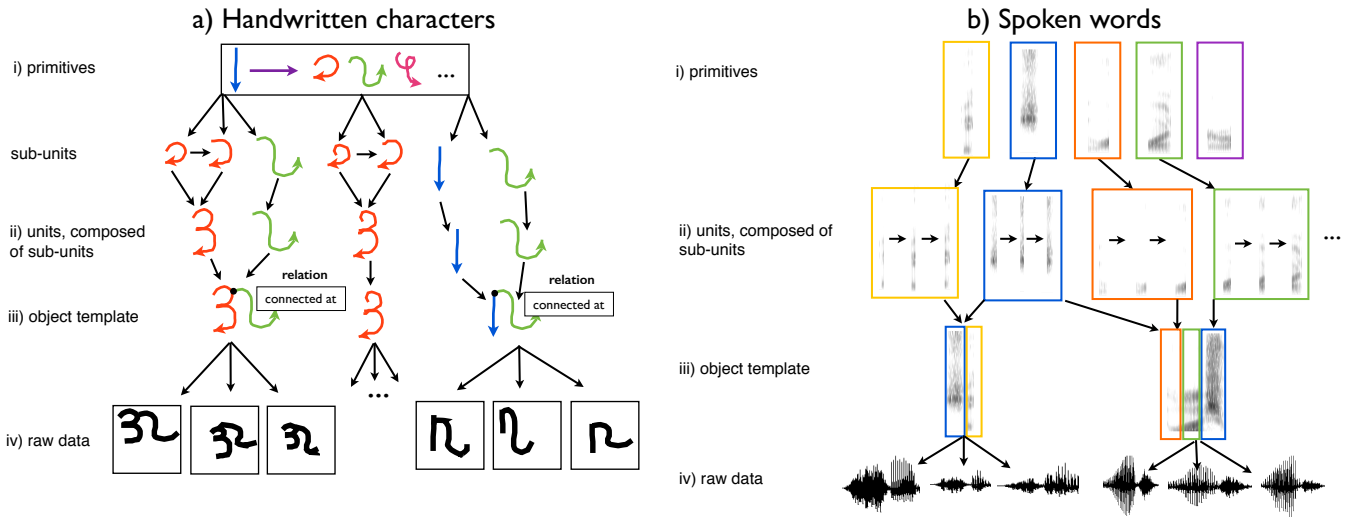

Figure 1: Hierarchical Bayesian modeling as applied to handwritten characters (Lake et al., 2013) and speech (this paper). Color coding highlights the re-use of primitive structure across different objects. The speech primitives are shown as spectrograms.

analysis-by-synthesis theories of speech recognition (Halle & Stevens, 1962; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) and more standard Hidden Markov models (HMMs) for Automatic Speech Recognition (ASR) (Juang & Rabiner, 1991). We extend the model of Lee and Glass (2012), which uses completely unsupervised learning to acquire a sequence of "phone-like" units (Fig. 1b), and test it on one-shot learning. Compared to the standard supervised training procedures in ASR, this more closely resembles the problem faced by an infant learning the speech sounds of their native language from raw speech, without any segmentation or phonetic labels. Once the units are learned, they can be combined together in new ways to define a generative model for a new word (Fig. 1b-iii). We compare people and the model on both the one-shot classification and one-shot generation of new Japanese words.

Model

Modern ASR systems usually consist of three components: 1) the language model, which specifies the distribution of word sequences, 2) the pronunciation lexicon, which bridges the gap between the written form and the spoken form, and 3) the acoustic model, which captures the acoustic realization of each phonetic unit in the feature space (Juang & Rabiner, 1991). The acoustic model is the only relevant component for this paper, and we represent it as a Hierarchical Hidden Markov model (HHMM) with two levels of compositional structure (Fine, Singer, & Tishby, 1998). At the top level, the phonetic units in a language (primitives in Fig. 1b-i) are the states of a Hidden Markov Model (HMM), where the state transition probabilities correspond to the bigram statistics of the units. At the lower level, each phonetic unit is further modeled as a 3-state HMM (Fig. 1b-ii), where the 3 sub-units (or sub-states) correspond to the beginning, middle, and end of a phonetic unit (Jelinek, 1976). These 3-state HMMs can be concatenated recursively to form a larger HMM that represents a word (Fig. 1b-iii).

Our HHMM model induces the set of phone-like acoustic units directly from the raw unsegmented speech data in a completely unsupervised manner, like an infant trying to learn the speech units of his or her native language. This contrasts with the standard supervised training procedure in ASR, requiring a parallel corpus of raw speech with word or phone transcripts. Similarly, existing cognitive models of unsupervised phoneme acquisition typically require known phonetic boundaries, where the speech sounds are represented in a low-dimensional space such as the first and second formant (Vallabha, McClelland, Pons, Werker, & Amano, 2007; Feldman, Griffiths, Goldwater, & Morgan, 2013).

Our model only receives raw unsegmented speech data, and as illustrated in Fig. 2, it must solve a joint inference problem that involves dividing the raw speech x into segments (vertical red lines in Fig. 2), identifying segments that should be clustered together with inferred labels z_s (color coded horizontal bars), and, most importantly, learning a set of phone-like acoustic units θ_i for that language, where the inferred labels z_s assign segments to acoustic units. Some of the other learned HHMM parameters are shown in Fig. 2, including the probability π_i of using any unit i as the initial state and the probability $\phi_{i,j}$ of transitioning from the i^{th} to the j^{th} acoustic unit. As is standard for acoustic models in ASR, each phone-like acoustic unit $1 \leq i \leq K$ is modeled as a 3-state HMM with parameters θ_i. The emission distribution of each sub-state is modeled by a 16-component Gaussian Mixture Model (GMM). These 3-state HMMs then generate the observed speech features $x_{s,1} \ldots x_{s,d_s}$ in each variable length segment, which are the standard Mel-Frequency Cepstral Coefficients (MFCCs) (Davis & Mermelstein, 1980).[1] The duration of each segment d_s is determined by the number of steps needed to traverse from the beginning to the end of the 3-state HMM that the segment is assigned to.

The full generative model for a stream of raw speech can

[1] Speech data are converted to 25 ms 13-dimensional MFCCs and their first and second order time derivatives at a 10 ms analysis rate.

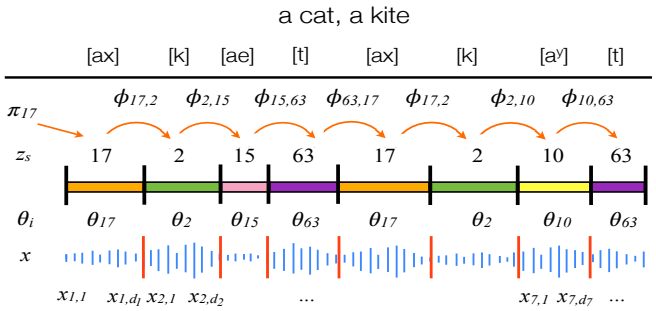

Figure 2: The model jointly segments the speech, clusters the segments (z_s), discovers a set of units (θ_i), and learns the transition probability between the units ($\phi_{i,j}$). Note that only speech data (x) was given to the model; the text *a cat, a kite* and the pronunciation are only for illustration.

be written as follows. For ease of explanation, we assume that the number of segments N is known. However, during learning, its value is unknown and can be learned by the inference method described below. The generative model is

$$\begin{aligned} \pi &\sim Dir(\eta) \\ \beta &\sim Dir(\gamma) \\ \phi_i &\sim Dir(\alpha\beta) \quad i = 1, \ldots, K \\ \theta_i &\sim H \\ z_1 &\sim \pi \\ z_s &\sim \phi_{z_{s-1}} \quad s = 2, \ldots, N \\ x_{s,1}, \ldots x_{s,d_s} &\sim \theta_{z_s}, \end{aligned} \tag{1}$$

where η, γ, and α are fixed hyper-parameters and variables π, β, and ϕ_i are all K-dim vectors with Dirichlet priors. The variable β can be viewed as the overall probability of observing each acoustic unit in the data, and it ties all the priors on transition probability vectors ϕ_i together. We impose a generic prior H on θ_i, where the details can be found in Sec. 5 and Sec. 6 of Lee and Glass (2012).

Inference has two main stages. First, the set of acoustic units is learned from a corpus by performing inference in the full generative model described above. Second, the learned model $(\pi, \beta, \phi, \theta)$ is fixed, and then individual word representations can be inferred as described in Experiment 1. Here we describe how the acoustic units are learned using Gibbs sampling. To sample from the posterior on units z_s for the corpus, we need to integrate over the unknown segmentation, which includes the number of segments N and their locations. We employ the message-passing algorithm described in Johnson and Willsky (2013) to achieve this.[2] Once the samples of z_s are obtained, the conditional posterior distribution of ϕ_i, β and π can be derived based on the counts of z_s. Also, we can then block-sample the state and Gaussian mixture assignment for each feature vector within a speech segment given the associated 3-state HMM. With the state and mixture assignment of each feature vector, we can update the parameters of the unit HMMs θ_i. Finally, we ran the Gibbs sampler for 10,000 iterations to learn the models reported in Experiment 1 and 2.

[2]We slightly modify the algorithm by ignoring the duration distribution of the hidden semi-Markov model.

Our model is an extension of the unsupervised acoustic unit discovery model presented in Lee and Glass (2012). However, unlike Lee and Glass (2012), which only captures the unigram distribution of the acoustic units, our model also learns bigram transition probabilities between units through a hierarchical Bayesian prior. We fixed the number of units, or states, at $K = 100$; however, we can easily extend the model to be non-parametric by imposing a hierarchical Dirichlet process prior on the states representing the phonetic units.

Experiment 1: Classification

Human subjects and several algorithms tried to classify novel Japanese words from just one example. Evaluation consisted of a set of tasks, where each task used 20 new Japanese words matched for word length in Japanese characters. Tasks required that the human or algorithm listen to 20 words (training) and then match a new word (test), spoken by a different talker, to one of the 20. Each task had 20 test trials, with one for each word. Since generalizing to speakers of different genders can be challenging in ASR, we had two conditions, where one required generalizing across genders while the other did not.

Stimuli. Japanese speech was extracted from the Japanese News Article Sentences (JNAS) corpus of speakers reading news articles (Itou et al., 1999). There were ten same-gender tasks, five with male talkers (word lengths 3 to 7) and five with female talkers (same word lengths). There were also ten different-gender tasks with word lengths from 3 to 12.

Humans. In this paper, all participants were recruited via Amazon's Mechanical Turk from adults in the USA. Analyses were restricted to native English speakers that do not know any Japanese. Before the experiment, participants passed an instructions quiz (Crump, McDonnell, & Gureckis, 2013), and there was a practice trial with English words for clarity.

Fifty-nine participants classified new Japanese words in a sequence of displays designed to minimize memory demands. Pressing a button played a sound clip, so words could be heard more than once. Participants were assigned to one of two conditions with same (5 trials) or different (10 trials) gender generalizations. To ensure that learning was indeed one-shot, participants never heard the same word twice and completed only one randomly selected test trial from each task. Responses were not accepted until all buttons had been tried. Corrective feedback was shown after each response. Eight participants were removed for technical difficulties, knowing Japanese, or selecting a language other than English as their native language.

Hierarchical Bayesian models. Two HHMMs were trained for the classification task. One model was trained on a 10-hour subset of the Wall Street Journal corpus (WSJ) (Garofalo, Graff, Paul, & Pallett, 1993) to simulate an English talker. The other model was trained on a 10-hour subset of the JNAS corpus with all occurrences of the training and test words excluded. The second model can be viewed as a

Japanese speaking child learning words from his/her parents; therefore, we allowed the talkers of the training and test words to overlap those in the 10-hours of Japanese speech.

As in the human experiment, for every trial, the model selects one of the 20 training words that best matches the test word. The Bayesian classification rule is approximated as

$$\underset{c=1\ldots20}{\text{argmax}}\ P(X^{(t)}|X^{(c)}) = \underset{c=1\ldots20}{\text{argmax}} \int_Z P(X^{(t)}|Z)P(Z|X^{(c)})\ dZ$$
$$\approx \underset{c=1\ldots20}{\text{argmax}} \sum_{l=1}^{L} P(X^{(t)}|Z^{(c)[l]}) \frac{P(X^{(c)}|Z^{(c)[l]})P(Z^{(c)[l]})}{\sum_{j=1}^{L} P(X^{(c)}|Z^{(c)[j]})P(Z^{(c)[j]})}, \quad (2)$$

where $X^{(t)}$ and $X^{(c)}$ are sequences of features that denote the test word and training words respectively. Words are defined by a unique sequence of acoustic units, such that $Z^{(c)} = \{z_1^{(c)}, \ldots, z_s^{(c)}\}$ are the units the model uses to parse $X^{(c)}$. Since it is computationally expensive to compute the integral, we approximate it with just the $L = 10$ most likely acoustic unit sequences $Z^{(c)[1]}, \ldots, Z^{(c)[L]}$ that the model generates for $X^{(c)}$ (Eq. 2). It is straightforward to apply the inferred model parameters π and ϕ_i to compute $P(Z^{(c)[l]})$. To compute $P(X^{(c)}|Z^{(c)[l]})$, we form the concatenated HMM for $Z^{(c)[l]}$ and use the forward-backward algorithm to sum over all possible unit boundaries and hidden sub-state labels.

Following Lake et al. (2013), we find marginally better performance by using the classification rule in Eq. 3 instead of Eq. 2,

$$\underset{c=1\ldots20}{\text{argmax}}\ P(X^{(t)}|X^{(c)}) = \underset{c=1\ldots20}{\text{argmax}}\ P(X^{(t)}|X^{(c)}) \frac{P(X^{(c)}|X^{(t)})}{P(X^{(c)})}, \quad (3)$$

where $P(X^{(t)}|X^{(c)})$ and $P(X^{(c)}|X^{(t)})$ are approximated as in Eq. 2, and, specifically, $P(X^{(c)}|X^{(t)})$ is computed by swapping the roles of $X^{(c)}$ and $X^{(t)}$. Both sides of Eq. 3 are equivalent if inference is exact, but due to the approximations, we include the similarity terms (conditional probabilities) in both directions. We also use the approximation $P(X^{(c)}) \approx \sum_{l=1}^{L} P(X^{(c)}|Z^{(c)[l]})P(Z^{(c)[l]})$.

Lesioned models. To more directly study the role of the learned units, we included three kinds of lesioned HHMMs. Two "unit-replacement" models, at the 25% or 50% levels, took the inferred units Z and perturbed them by randomly replacing a subset with other units. After the first unit was replaced, additional units were also replaced until 25% or 50% of the speech frames $x_{i,j}$ now belonged to a different unit. Both the English and Japanese trained models were lesioned in these ways. An additional "one-unit" HHMM model was trained on Japanese with only one acoustic unit, providing a rather limited notion of compositionality.

Dynamic Time Warping (DTW) We compare against the classic Dynamic Time Warp (DTW) algorithm that measures similarity between two sequences of speech features, requiring no learning (Sakoe & Chiba, 1978). The DTW distance between two sequences is defined as the average distance between features of the aligned sequences, after computing an optimal non-linear alignment.

Results and discussion. The one-shot classification results are shown in Table 1. Human subjects made fewer than 3% errors. For the same gender task, the HHMM trained on Japanese achieved an error rate of 7.5%, beating both the same model trained on English and DTW. All models performed worse on the different gender task, which was expected given the simple MFCC feature representation that was used. The gap between human and machine performance is much larger for the HHMM trained on English than the model trained on Japanese. This difference could be the product of many factors, including differences in the languages, speakers, and recording conditions. While the English-trained model may be more representative of the human participants, the Japanese-trained model is more representative of everyday word learning scenarios, like a child learning words spoken by a familiar figure.

The superior performance of the HHMM over DTW supports the hypothesis that learning-to-learn and compositionality are an important facilitator of one-shot learning. The dismal performance of the lesioned HHMM models, which never achieved did better than 88% errors regardless of training language, further suggests that learning-to-learn alone, without a rich notion of compositionality, is not powerful enough to achieve good results.

Table 1: One-shot classification error rates

Learner	Same gender	Different gender
Humans	2.6%	2.9%
HHMM (Japanese)	7.5%	21.8%
HHMM (English)	16.8%	34.5%
DTW	19.8%	43%
Lesioned HHMM	$\geq$ 88.5%	$\geq$ 88.8%

Experiment 2: Generation

Humans generalize in many other ways beyond classification. Can English talkers generate compelling new examples of Japanese words? Here we test human subjects and several models on one-shot generation. Performance was measured by asking other humans (judges) to classify the generated examples into the intended class, which is an indicator of exemplar quality. This test is not as strong as the "auditory Turing test" (Lake et al., 2013), but the HHMM cannot yet produce computerized voices that are confusable with human voices.

Humans. Ten participants spoke Japanese words after listening to a recording from a male voice. Each participant was assigned a different word length (3 to 12) and then completed twenty trials of recording using a computer microphone. Participants could re-record until they were satisfied with the quality. This procedure collected one sample per stimulus used in the previous experiment's different gender condition.

Hierarchical Bayesian models. All of the full and lesioned HHMM models from Experiment 1 listened to the same new Japanese words as participants and then synthesized new ex-

amples. To generate speech, the models first parsed each word into a sequence of acoustic units, Z, and generated MFCC features from the associated 3-state HMMs. While it is easy to forward sample new features, we adopted the procedure used by most HMM-based speech synthesis systems (Tokuda et al., 2013) and generated the mean vector of the most weighted Gaussian mixture for each HMM state. Furthermore, HMM-based synthesis systems have an explicit duration model for each acoustic unit in addition to the transition probability (Yoshimura, Tokuda, Masuko, Kobayashi, & Kitamura, 1998). Since this information is missing from our model, we forced the generated speech to have the same duration as the given Japanese word. More specifically, for each inferred acoustic unit z_i in $Z = \{z_1, \ldots, z_s\}$, we count the number of frames d_i in the given word sample that are mapped to z_i and generate d_i feature vectors evenly from the 3 sub-states of θ_{z_i}. Finally, to improve the quality of the speech, we extracted the fundamental frequency information from the given word sample by using *Speech Signal Processing Toolkit (SPTK)* (2013). This was combined with the generated MFCCs and the features were then inverted to audio (Ellis, 2005).

Evaluation procedure. Using a within-subjects design, 30 participants classified a mix of synthesized examples from both people and the comparison models. The trials appeared as they did in Experiment 1, where instead of an original Japanese recording, the top button played a synthesized test example instead. The 20 training clips played original Japanese recordings, matched for word length within a trial as in Experiment 1. Since the synthesized examples were based on male clips, only the female clips were used as training examples. There was one practice trial (in English) followed by 50 trials with the synthesized example drawn uniformly from the set of all synthesized samples across conditions. Since the example sounds vary in quality and some are hardly speech-like, participants were warned that the sound quality varies, may be very poor, or may sound machine generated. Also, the instructions and practice trial were changed from Experiment 1 to include a degraded rather than a clear target word clip. All clips were normalized for volume.

Results and discussion. Samples of the machine generated speech are available online.[3] Several participants commented that the task was too long or too difficult, and two participants were removed for guessing.[4] The results are shown in Fig. 3, where a higher "score" (classification accuracy from the judges) suggests that generated examples were more compelling. English speakers achieved an average score of 76.8%, and the best HHMM was trained on Japanese and achieved a score of 57.6%. The one-unit model set the baseline at 17%, and performance in the HHMM models decreased towards this baseline as more units were randomly replaced. As with Experiment 1, the Japanese training was superior to English training for the HHMM.

The high performance from human participants suggests that even naive learners can generate compelling new examples of a foreign word successfully, at least in the case of Japanese where the phoneme structure is related. The full HHMMs did not perform as well as humans. However, given the fact that the one-unit and unit-replacement models only differed from the full HHMMs by their impoverished unit structure, the better results achieved by the full HHMM models still highlight the importance of learning-to-learn and compositionality in the one-shot generation task.

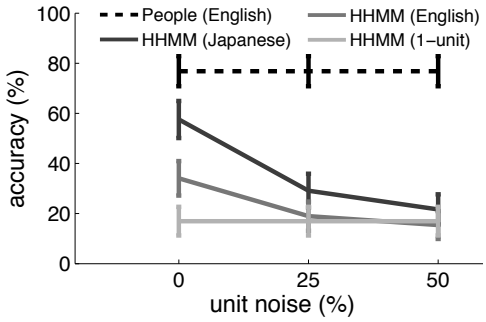

Figure 3: Percent of synthesized examples that human judges classified into the correct spoken word category. Parentheses indicate the language the model was trained on. Error bars are 95% binomial proportion confidence intervals based on the normal approximation.

Replication. As mentioned, a number of participants commented on the task difficulty. Since human and machine voices were intermixed, it is possible that some participants gave up on trying to interpret any of the machine speech. We investigated this possibility by running a related between subjects design without the degraded models. Forty-five participants were assigned to one of three conditions: speech generated by humans, by the HHMM trained on Japanese, or by the HHMM trained on English. Three participants were removed for knowing some Japanese, and three more were removed by the earlier guessing criterion. The results largely confirmed the previous numbers. The human-generated speech scored 80.8% on average. The HHMM trained on Japanese and on English scored 60% and 27.3%. All pair-wise t-tests between these groups were statistically significant ($p<.001$). The previous numbers were 76.8%, 57.6%, and 34.1%, respectively.

General Discussion

We compared humans and a HHMM model on one-shot learning of new Japanese words, evaluating both classification and exemplar generation. Humans were very accurate classifiers, and they produced acceptable examples of Japanese words even with no experience speaking the language. These successes are consistent with the rapid rate in

[3] http://web.mit.edu/brenden/www/speech.html

[4] Participants spent from 19 to 87 minutes on the task, and there was correlation between accuracy and time (R=0.58, $p<0.001$). In a conservative attempt to eliminate guessing, two participants were removed for listening to the "target word" fewer than twice on average per trial (6 times was the experiment average). This made little difference to the pattern of results.

which children acquire new vocabulary, and our model aimed to provide insight into how this is possible. The HHMM trained on Japanese, when acquiring new words in its "native" language, comes within 5% of human performance on classification. The lower performance of the HHMM trained on English could have resulted from many factors, and in the future, we plan to investigate whether the trouble lies in generalizing across speakers, across data sets, or across languages. The lesioned models and Dynamic Time Warp demonstrated inferior performance on the classification and generation tasks, adding to previous evidence that compositionality and learning-to-learn are important for one-shot learning (Kemp et al., 2007; Salakhutdinov et al., 2012; Lake et al., 2013).

Far from the final word, we consider our investigation to be a first step towards understanding how adults and children learn new phonological sequences from just one exposure. We see a more realistic analysis-by-synthesis approach as a promising avenue for further research (Bever & Poeppel, 2010). Influential theories of speech perception have argued for explicit modeling of the articulatory process (Halle & Stevens, 1962; Liberman et al., 1967), and in our model, aspects of production are only implicitly represented through the learned acoustic units. Despite the inherent challenges in representing and inverting a complex generative process, could one-shot learning be improved by more faithfully following the analysis-by-synthesis program, and could this lead to general improvements in automatic speech recognition?

Acknowledgements. We would like to thank Peter Graff, Tim O'Donnell, Stefanie Shattuck-Hufnagel, Elizabeth Choi, and Max Kleiman-Weiner for helpful discussions, and Ann Lee for providing the DTW scoring tool. This work was supported by the Center for Minds, Brains and Machines (CBMM) funded by NSF STC award CCF-1231216, ARO MURI contract W911NF-08-1-0242, and a NSF Graduate Research Fellowship held by Brenden Lake.

References

Bever, T. G., & Poeppel, D. (2010). Analysis by synthesis: a (re-) emerging program of research for language and vision. *Biolinguistics*, *4*, 174–200.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Colunga, E., & Smith, L. B. (2005, April). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, *112*(2), 347–82.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013, March). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3).

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *28*(4), 357–366.

Ellis, D. P. W. (2005). *RASTA/PLP/MFCC feature calculation and inversion.* Retrieved from `www.ee.columbia.edu/~dpwe/resources`

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–78.

Fine, S., Singer, Y., & Tishby, N. (1998). The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine learning*, *62*, 41–62.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009, May). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–85.

Garofalo, J., Graff, D., Paul, D., & Pallett, D. (1993). *CSR-I (WSJ0) Other.* Linguistic Data Consortium, Philadelphia.

Halle, M., & Stevens, K. (1962). Speech Recognition: A Model and a Program for Research. *IRE Transactions on Information Theory*, *8*(2), 155–159.

Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., & Shikano, K. (1999). JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of Acoustical Society of Japan*, *20*, 199–206.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*, 532 – 556.

Johnson, M., & Willsky, A. (2013). Bayesian nonparametric hidden semi-Markov models. *Journal of Machine Learning Research*, *14*, 673–701.

Juang, B. H., & Rabiner, L. (1991). Hidden Markov models for speech recognition. *Technometrics*, *33*(3), 251–272.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2012). Concept learning as motor program induction: A large-scale empirical study. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2013). One-shot learning by inverting a compositional causal process. In *Advances in Neural Information Processing Systems 26.*

Lee, C.-y., & Glass, J. (2012). A Nonparametric Bayesian Approach to Acoustic Model Discovery. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 40–49.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Ohata, K. (2004). Phonological differences between Japanese and English: Several Potentially Problematic Areas of Pronunciation for Japanese ESL/EFL Learners. *Asian EFL Journal*, *6*(4), 1–19.

Sakoe, H., & Chiba, S. (1978, February). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49.

Salakhutdinov, R., Tenenbaum, J., & Torralba, A. (2012). One-Shot Learning with a Hierarchical Nonparametric Bayesian Model. *JMLR WC&P Unsupervised and Transfer Learning*, *27*, 195–207.

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323.

Smith, L., & Yu, C. (2008, March). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–68.

Speech Signal Processing Toolkit (SPTK). (2013). Retrieved from `http://sp-tk.sourceforge.net/`

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–40.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013). Speech synthesis based on Hidden Markov Models. *Proceedings of the IEEE*, *101*(5), 1234–1252.

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

Xu, F., & Tenenbaum, J. B. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*(2), 245–272.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1998). Duration modeling for HMM-based speech synthesis. In *ICSLP* (Vol. 98, pp. 29–31).