Learning words and speech units through natural interactions

Jonas Hörnstein[1], *José Santos-Victor*[1]

[1]Institute for System and Robotics (ISR), Instituto Superior Técnico, Lisbon, Portugal

jhornstein@isr.ist.utl.pt, jasv@isr.ist.utl.pt

Abstract

This work provides an ecological approach to learning words and speech units through natural interactions, without the need for preprogrammed linguistic knowledge in form of phonemes. Interactions such as imitation games and multimodal word learning create an initial set of words and speech units. These sets are then used to train statistical models in an unsupervised way.

Index Terms: multimodal learning, ecological approach, motor learning, interactions

1. Introduction

Infants are able to acquire impressive language skills from very little speech exposure, and are doing this in a natural way through the interaction with their caregivers. Computer based systems for automatic speech recognition (ASR) not only require much more data to achieve comparable word error rates [1], but also require much of the data to be hand labeled. As a result, traditional ASR-system are still far from human capabilities when it comes to flexibility, and it is therefore interesting to try to mimic the way infants learn their language. Unfortunately it is not completely understood how infants do this, nor to which extent linguistic knowledge is already preprogrammed in the human brain. Here we follow the ecological and emergent approach [2] and assume that linguistic structures such as phonemes are learned through the interaction with the environment rather than innate.

The infants' relatively lack of speech exposure compared to ASR-systems, is compensated by the richness of the data directed to them. Infant directed speech (IDS) is highly structured and characterized by what seems like physically motivated tricks to maintain the communicative connection to the infant, actions that at the same time also may enhance linguistically relevant important aspects of the signal. For example, target words are typically highlighted using focal stress and utterance-final position [3] [4]. Also, whereas communication between adults usually is about exchanging information, speech directed to infants is of a more referential nature. The adult refers to objects, people and events in the world surrounding the infant [5]. Because of this, the sound sequences that the infant hears are very likely to co-occur with actual objects or events in the infant's visual field.

This type of information has also be used in computer-based systems such as CELL [6], Cross-channel Early Lexical Learning. There, an architecture for processing multisensory data is developed and implemented in a robot. The robot is able to acquire words from untranscribed acoustic and video input and represent them in terms of associations between acoustic and visual sensory experience. Compared to training conventional ASR systems that maps the speech signal to human specified labels, this is an important step towards creating more ecological models. However, significant shortcuts are still taken, such as the use of a predefined phoneme-model where a set of 40 phonemes are used and the transition probabilities are trained off-line with a large scale database. An unsupervised model for learning words and speech units from multisensory data is described in [7]. However, the method requires words to be readily segmented and does not work for natural interactions. An alternative method is to completely avoid the phonemes and directly look for word-like segments using simple methods for pattern matching or Dynamical Time Warp (DTW) [8] [9]. According to the ecological approch to language acquisition, these simple models may very well describe how infants are able to learn their first words. Underlying concepts like phonemes may instead be seen as emergent consequences imposed by increasing representation needs [10] [11].

While the exact phonemes differ among languages some phonemes, such as the corner vowels [i], [a] and [u], are widely used. This is natural when looking at how these are produced, as these are defined as the extreme points in our articulatory vowel space. Also other phonemes may be better understood when looking at how they are produced. Since infants do not only learn to recognize speech sounds, but also to produce them, they may take advantage of this also when learning their speech units. The parallel development of speech production and recognition during an infant's first year has been described in [12]. At birth infants are able to discriminate phonetic contrasts of all languages, but later develops a "phonetic magnet" that forces sound to be perceived as one of the phonemes that are used in the particular language. Interestingly, infants seem to begin producing such sounds shortly before they show a preference for perceiving these same sound. This relationship between sounds that we can produce and those that we perceive leads to believe that the motor area in the brain is involved not only in the task of production, but also in that of recognition. This was first suggested in the Motor Theory of speech perception [13] and is supported by more recent work in neuroscience [14]. While we do not take a pure motor-based approach, we augment the traditional acoustic features with additional motor-based information. Also, by being able to produce speech sounds it becomes possible for the computer-based system to take part in imitation games that may have an important role in human language development. Several works have shown that motor-learning can be used for finding speech units [15] [16]. This work shows how interactions can be used to create an initial set of words and speech units, and how to use these to build more advanced statistical models of the language.

The rest of paper is organized as follows. In section 2 we describe the initial word learning, based on pattern matching and multimodal information. In section 3 we describe how motor learning and imitation games can be used to bootstrap the

learning of speech units. The statistical models are described in section 4, and experimental results are presented in section 5. Finally conclusions are given in section 6.

2. Multimodal word learning

In order to create an initial word model the robot looks for recurring acoustic events and associate those to visual objects in its environment. A short term memory (10-20 s length) is used to restrict the search space and increase the possibility that the recurring acoustic patterns that are found refer to the same object. Recurring patterns in the short-term memory are paired with the visual object and send to a long-term memory where they are organized in hierarchical tries. Finally, the mutual information criterion is used to find which words are consistently used for a certain object.

2.1. Finding recurring events

In order to find recurring patterns the sound stream is first sequenced into utterances. This is done automatically when the sound level is under a certain threshold value for at least 200 ms. Each utterance within the short term memory at a given time is compared pair-wise with all other utterances in the memory in order to find recurring patterns. The utterances are aligned in time and we calculate the sum of differences between their mel coefficients creating a vector with the acoustic distance between the two utterances at each window. The second utterance is then shifted forward and backward in time and for each step a new distance vector is calculated. These are then combined into a distance matrix. Word candidates are found by looking for minima in the distance matrix and then moving left and right in the matrix as long as the distance metric is below a certain threshold.

In order to take advantage of the structure of infant directed speech and to mimic infants' bias towards target words in utterance-final position and focal stress, we also check for these features. Focal stress is found by looking for the F0-peak. While there are many ways for adults to stress words (e.g. pitch, intensity, length) it has been found that F0-peaks are mainly used in infant directed speech [3]. If the F0-peak of the utterance as a whole is within the boundaries of the word candidate, the word candidate is considered to be stressed. If a word candidate is not stressed and in utterance-final position we may reject it with a specified probability.

The same pattern matching technique is also used to compare visual objects. Using the silhouette of the object we create a representation of its shape by taking the distance between the center of mass and the perimeter of the silhouette. When comparing two object representations with each other we first normalize the vectors and then perform a pattern matching, much in the same way as for the auditory representations, by shifting the vectors one step at a time. By doing this we get a measurement of the visual similarity between objects that is invariant to both scale and rotation. For details please refer to [8].

2.2. Hierarchical clustering

When both a word candidate and a visual object are found, their representations are paired and send them to a long term memory. To organize the information we use an hierarchical clustering algorithm. Word candidates and visual objects are organized independently into two different tree clusters. The algorithm starts by creating one cluster for each item. It then iteratively joins the two clusters that have the smallest average distance between their items until only one cluster remains.

While the algorithm is the same for both trees, the distance measure varies slightly between them. The distance between the visual objects is measured directly through the pattern matching explained above. For the acoustic similarity we use Dynamic Time Warp (DTW) to measure the distance between different word candidates. The reason to use DTW instead of directly applying the pattern matching described earlier is to be less sensitive to how fast the word candidate is pronounced.

2.3. Multimodal integration

When we have interconnected multimodal representations, which is the case for the word candidates and visual objects that assumingly refers to the same object we can make use of these connections, not only to create associations, but also to find where we should cut the trees in order to get a good representations of the words and the objects. In order to find which branch in the word candidate tree that should be associated with which branch in the object tree we use the mutual information criterion. In the general form this can be written as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \left(\frac{p(x,y)}{p_1(x)p_2(y)} \right) \quad (1)$$

Where $p(x,y)$ is the joint probability distribution function of X and Y, and $p_1(x)$ and $p_2(y)$ are the marginal probability distribution functions of X and Y respectively.

We want to calculate $I(X;Y)$ for all combinations of clusters and objects in order to find the best word representations. For a specific word cluster A and visual cluster V we define the binary variables X and Y as

$$X = \{1\ if\ observation \in A;\ 0\ otherwise\}$$
$$Y = \{1\ if\ observation \in V;\ 0\ otherwise\}$$

The probability functions are estimated using the relative frequencies of all observations in the long-term memory, i.e. $p_1(x)$ is estimated by taking the number of observations within the cluster A and dividing with the total number of observations in the long-term memory. In the same way $p_2(y)$ is estimated by taking the number of observations in the cluster V and again dividing with the total number of observations. The joint probability is found by counting how many of the observations in cluster A that is paired with an observation in cluster V and dividing by the total number of observations.

3. Bootstrapping speech units

The robot can learn an initial set of speech unit by imitating its caregiver. To do this the robot needs to be able to produce sounds. It has therefore been equipped with a simulated vocal tract [18], and a synthesizer. It also has a neural network that is used as an audiomotor map, which maps speech data to vocal tract positions. This audiomotor map must be learnt before the robot can participate in the imitation games.

3.1. Learning the audiomotor map

To learn the audiomotor map the system first makes use of babbling where it makes random articulations and listens to the sound produced. If the sound is over a threshold level the system tries to map the sound back to the sensorimotor map. The

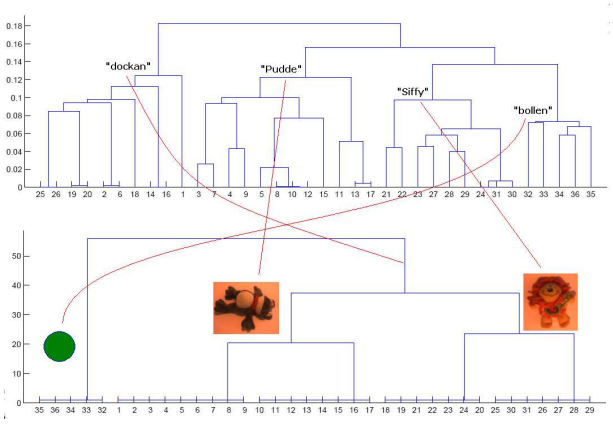

Figure 1: Clusters of the extracted word candidates (above) and visual objects (below), and the four word-object connections with highest mutual information.

mapped positions are then compared to the original ones and the error is used to update the sensorimotor map with back-propagation. By repeating this the error will gradually decrease.

However, in the same way as the sound produced by an infant is different from a similar sound produced by an adult, the sound produced by the robot is different from that of its caregiver. This can be overcome if the caregiver imitates the robot and thereby allows the robot to use not only its own utterance, but also that of the caregiver, to update the map. We have previously shown that prosodic features can help the system to decide if there is an imitation or not [17].

3.2. Learning speech units

Once the map is learned, the system can use a parroting behaviour where it tries to imitate the caregiver. Even if the map is not perfect, the caregiver can often get the system to repeat the desired speech sound by slightly changing the voice. For most of vowels it is not necessary to adapt the voice too much. Typically between one and ten attempts were enough to obtain a satisfying result. When the caregiver is happy with the sound produced by the system it gives positive feedback which causes the system to store the current articulator positions in its speech motor vocabulary. Using this method we have been able to teach the system vocal tract positions for nine Swedish vowels (a, o, u, å, e, i ,y, ä, ö).

4. Creating statistical models

While the initial word learning works well for creating a small vocabulary from a limited number of demonstrations, the hierarchical trees continuously grow as the number of demonstrations increase. To avoid having to store every single example of a word or a speech unit, it becomes necessary to create statistical models. In ASR-system, words are typically modelled using Hidden Markov Models (HMM). Instead of directly creating a HMM for each word, it is more efficient to create a model for each speech unit and then concatenate those into words. Since we don't know the number of speech units we start with a small number and then increase the number iteratively as long as it improves our word model. The speech units are found by clustering speech data in an unsupervised way using K-means. As an initial guess for the number of clusters, and their center positions, we can use the speech units that were learnt in the imitation games. This bootstrapping is an important step since K-means is relatively sensible to the initial guess.

For the statistical modelling, the speech signal is represented both by the MFCC (including their first and second derivatives) and the mapped vocal-tract positions, resulting in a vector of 46 features for each window.

Next, a Gaussian model is calculated for each speech unit, and all speech units are then connected to themselves and all others in a HMM. The same speech data that was used to create the clusters is now used to estimate the transition probabilities and updating the Gaussian models using the Baum-Welch EM algorithm. This results both in updated models for our speech units and the creation of a phonotactic language model.

The statistical word models are estimated by selecting the speech example in the center of each cluster in the initial word model, and calculating the most likely path in the phonotactic model for those observations. This is done with the Viterbi algorithm, and the resulting path is then used as a HMM for the word.

Finally the word models are evaluated by calculating the word recognition rate for all remaining examples in the initial word cluster. A new cluster is inserted and the process is repeated for as long as the recognition rate improves. The complete process can be summarized in the following steps:

1. K-means is used to cluster the speech data into a specified number of speech units in an unsupervised manner. A Gaussian model is then created for each speech unit.
2. A phonotactic model is created by estimating the transition probabilities between all speech units using Baum-Welch algorithm. At the same time the Gaussian models are reestimated for each speech unit.
3. A HMM is created for each word in the initial word model by choosing the speech example in the center of each cluster and calculating the most likely sequence of speech units with the Viterbi algorithm.
4. The word recognition rate is calculated on a test set containing the remaining examples in each cluster from the initial word model.
5. Add a speech unit and reiterate as long as the recognition rate increases.

5. Experimental results

The multimodal word learning has been implemented in a humanoid robot. In a previous experiment the robot was able to learn the names of a number of toys that the caregiver placed in front of the robot Figure 1. In this experiment we were mainly interested in testing the statistical model. The robot was therefore taught a number of additional words. Like in the previous experiments only full sentances and no single words were given to the robot. However, this time no images were used. Instead the utterances were labeled with a number representing the object. The pattern matching resulted in 88 word candidates that were divided into 8 different clusters by using hierarchical clustering and the mutual information criterion.

For each cluster we then created a statistical model using the method described above. This was done both with and without bootstrapping. With bootstrapping, the vowels learnt by imitation were used as initial guesses for the positions of the speech units. Without bootstrapping, random samples from the speech data was used for initializing the K-means algorithm.

We started with only 5 speech units and iteratively increased the number until 12 speech units when there was no longer any improvements in the recognition rate. The results are shown in Figure 2.

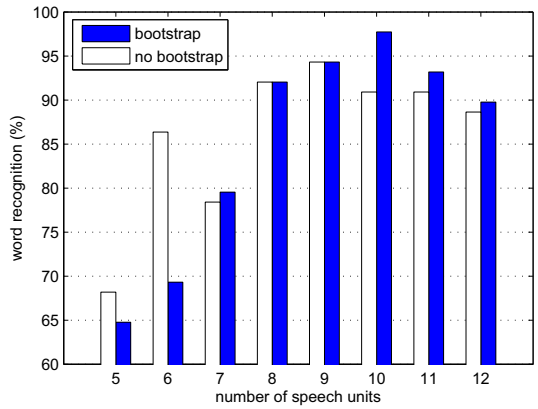

Figure 2: Word recognition rates for different number of speech units.

The best result, 98% recognition rate, was obtained when using 10 speech units and bootstrapping. The resulting word models are shown in Table 1. Note that some of the speech units have a close to one-to-one relation with real phonemes, such as 1=a and 3=m.

Table 1: *Statistical word models for 10 speech units with bootstrapping*

word	representation
siffy	5 7 5 7
pudde	9 6 10 5 8
docka	6 10 6 10 5 8
pappa	6 1 6 10 6 1
mamma	3 1 3 1
lampa	7 9 1 2 3 6 10 6 1
pippi	7 5 10 6 7
vovve	4 2 6 2 6 5 2 9

6. Conclusions and future work

This work has presented an ecological approach where words and speech units are learnt through natural interactions without any preprogrammed linguistic knowledge in the form of phonemes. Initial word models are found using pattern matching and multimodal information. These can then be used to create statistical models. We have also show that the statistical models can be improved by teaching the robot a number of initial speech units that can be used to bootstrap the statistical learning. This can be done through the use of interactions in the form of imitation games.

Near future work includes testing this model on larger data sets. The model can also be extended to learn actions and events and higher level information such as grammar.

7. Acknowledgements

This work was partially supported by EU Project HANDLE and by the Fundação para a Ciência e a Tecnologia (ISR/IST pluriannual funding) through the POS Conhecimento Program that includes FEDER funds.

8. References

[1] Moore, R. K., "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners", Proc. EUROSPEECH'03, Geneva, pp. 2582-2584, 1-4 September, 2003

[2] Lacerda, F., Klintfors, E., Gustavsson, L., Lagerkvist, L., Marklund, E., Sundberg, U., "Ecological Theory of Language Acquisition", In Genova: Epirob 2004, 2004

[3] Fernald, A. and Mazzie, C., "Prosody and focus in speech to infants and adults", Developmental Psychology, 27, pp 209-221, 1991

[4] Albin, D. D., and Echols, C. H., "Stressed and word-final syllables in infant-directed speech", Infant Behavior and Development, 19, pp 401-418, 1996

[5] Lacerda, F., Marklund, E., Lagerkvist, L., Gustavsson, L., Klintfors, E., Sundberg, U., "On the linguistic implications of context-bound adult-infant interactions", In Genova: Epirob 2004, 2004

[6] Roy, D. and Pentland, A., "Learning words from sights and sounds: A computational model", Cognitive Science, 2002, vol 26, pp 113-146

[7] Iwahashi, N., "Language acquisition through a human-robot interface by combining speech, visual, and behavioral information", Information Sciences 153, pp 109-121, 2003

[8] Hörnstein, J., Gustavsson, L., Lacerda, F., Santos-Victor, J., "Multimodal Word Learning from Infant Directed Speech", IEEE/RSJ International Conference on Intelligent Robotic Systems, St. Louis, USA, October 2009

[9] Aimetti, G., "Modelling Early Language Acquisition Skills: Towards a General Statistical Learning Mechanism", Proceedings of the EACL 2009 Student Research Workshop, pp 1-9, Athens, Greece, 2 April, 2009

[10] Nowak, M. A., Plotkin, J. B., Jansen, V. A. A., "The evolution of syntactic communication", Nature, 404, pp 495-498, 2000

[11] Lacerda, F., "Phonology: An emergent consequence of memory constraints and sensory input", Reading and Writing: An Interdisciplinary Journal, 16, pp 41-59, 2003

[12] Kuhl, P., Andruski, J. E., Christovich, I. A., Christovich, L. A., Kozhevnikova, E. V., Ryskina, V. L. et al., "Cross-language analysis of Phonetic units in language addressed to infants", Science, 277, pp. 684-686, 1997

[13] Liberman, A. and Mattingly, I., "The motor theory of speech perception revisited", Cognition, 21:1-36, 1985

[14] Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G., "Speech listening specifically modulates the excitability of tongue muscles: a TMS study", European Journal of Neuroscience, Vol 15, pp. 399-402, 2002

[15] Hörnstein, J. and Santos-Victor, J., "A Unified Approach to Speech Production and Recognition Based on Articulatory Motor Representations", 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, USA, October 2007

[16] Kanda, H. and Ogata, T., "Vocal imitation using physical vocal tract model", 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, USA, October 2007, pp. 1846Ű1851

[17] Hörnstein, J., Gustavsson, L., Santos-Victor, J., Lacerda, F., "Modeling Speech imitation", IROS-2008 Workshop - From motor to interaction learning in robots, Nice, France, September 2008.

[18] Maeda, S., "Compensatory articulation during speech: evidence from the analysis and synthesis of vocat-tract shapes using an articulatory model", in Speech production and speech modelling (W. J. Hardcastle and A. Marchal, esd.), pp. 131-149. Boston: Kluwer Academic Publishers