CONTEXT INDUCED MERGING OF SYNONYMOUS WORD MODELS IN COMPUTATIONAL MODELING OF EARLY LANGUAGE ACQUISITION

Okko Räsänen

Department of Signal Processing and Acoustics, Aalto University, Finland

ABSTRACT

It has been shown that both infants and machines are able to discover recurring word-like patterns from continuous speech in the absence of supervision. However, these early models for words do not always generalize well across different acoustic variants of the same words. Instead, several parallel models for words or multiple fragments of a word are initially learned. In this work, we study a two-stage computational framework for refining the initially acquired representations of acoustic word patterns. In the first stage, the automatically discovered word patterns are studied in the context of visual word referents, enabling grounding of the word forms to the systematically co-occurring objects and actions in the environment. In the second stage, synonymy of the words is measured in terms of the similarity of their environmental contexts. The word models that share similar external context are merged together, producing a lexicon with a smaller number of parallel models for each word and with a greater generalization capability from each model towards new realizations of the word. The experimental results show that the context-based equivalence and merging of parallel models leads to a more compact and higher quality lexicon than a learning process based purely on acoustic similarities.

Index Terms— language acquisition, pattern discovery, latent learning, random indexing

1. INTRODUCTION

The manner how infants acquire their native language seems almost effortless. Towards the end of their first year, they are already sensitive to typical phonetic and prosodic aspects of their native language, recognize a number of spoken words, and are on the brink of producing words by themselves. However, closer inspection on early language development reveals that young infants are not always successful in understanding previously learned words when they are spoken in new prosodic contexts or by new speakers with novel acoustic characteristics. Instead, the infants may treat the same words spoken by different speakers as totally different acoustic patterns (e.g., [1]; see also [2] and references therein). Only later, through extensive exposure to the words spoken in various situations with non-random referential context (events and objects that the words are referring to), the infant acquires knowledge that acoustically different patterns can map to a same external concept.

The above behavioral findings are in line with the research in automatic speech recognition (ASR) and computational modeling of language acquisition (CMLA). In ASR, it is well known that acoustic models trained on only one speaker and one speaking style generalize poorly to new speakers. This means that training data from numerous speakers are required in order to build speaker-independent systems. In CMLA, models investigating the emergence of early speech perception skills face the same problem of acoustic variability.

For example, in the work of [3] and [4], the distributional learning of vowel categories from speech was investigated. The authors applied unsupervised statistical methods to estimate proper vowel categories from formant frequencies ([3]) or MFCC features ([4]). When the obtained distributions were evaluated in speaker independent case, the categorization performance of vowel tokens was notably lower than in the speaker-dependent case. In [5], a computational model for fully unsupervised acquisition of ungrounded word patterns from continuous speech was presented. When the performance of the model was studied in detail, it was found that there were typically several parallel internal representations that had been learned for each annotated word even for speech material spoken by only one speaker. When several speakers with varying voice qualities were used to train the model, the number of parallel models increased even more [5]. Although the learned word models were still responsive to numerous varying tokens of the same word, the overall results seem to suggest that the acoustic variability, and on the other hand the acoustic overlap between different words, is too high in order to obtain perfectly selective and sensitive word models in purely bottom-up manner. In the absence of any additional source of constraints to the learning problem, and in order to maintain distinctiveness of different lexical items, the variability inevitably leads to a situation where there are initially more acoustic word pattern models than that there are actual words; the system has no way of knowing which aspects of speech signals are relevant for differentiating phonetic content from the acoustic carriers and suprasegmental details (but see also [6]). This is similar to the effects reported in the study of infant speech perception [2].

In this work, we extend the work of [5] and use computational simulations to explore grounding of automatically discovered word forms into external word referents. Furthermore, we present a mechanism for merging of functionally equivalent (synonymous) word models together in order to obtain a more compact lexicon than what can be possibly obtained in the case of purely bottom-up acoustic clustering. We propose that the referential contexts in which words occur play an essential role in the development of early vocabulary, providing the necessary constraint for mapping of acoustically distant speech tokens under the same linguistic categories. The proposed learning process is closely connected to the definition of word synonymy, i.e., the degree of similarity of contexts in which two or more words typically occur. Here we simply expand the definition of synonymy to the level of acoustic patterns, studying synonymy of acoustic patterns in the presence of events and objects that the words refer to.

Also note that the present approach is different from word learning models such as [7-9] in that it does not assume that words are always learned directly in the context of more or less definite contextual referents. Instead, the system first learns recurring acoustic patterns from speech and only later attempts to ground them into their referents through cross-situational learning.

We will first describe the speech material used in the experiments, followed by description of the computational methods. The third section is dedicated to the experiments with word model merging, whereas the final section discusses the findings and conclusions from the experiments.

2. MATERIAL

The speech material used in the experiments was taken from the Y2 UK section of CAREGIVER corpus [10]. In one talker case, the entire material from one female speaker (*Speaker-02*) was used so that 2000 utterances were used for training and the remaining 397 novel utterances were used for evaluation. For two talker case, data from one male and one female was used (*Speaker-01* and *Speaker-02*), with a total of 4000 utterances for training in randomized order and 794 utterances for evaluation.

In the CAREGIVER Y2 corpus, each utterance contains 1-4 target keywords surrounded by carrier sentences (mean 5.96 words including function words; e.g., "Where is the **happy horse**?", keywords emphasized). There are a total of 50 different keywords in the material and the overall vocabulary size is 80. Unlike real speech, the presence of keywords is statistically balanced over the corpus in order to remove any word-to-word dependencies of the keywords. This is required in order to avoid over-simplification of the learning problem. Each of the 50 keywords is associated with a unique tag that denotes the presence of a keyword in an utterance, simulating a situation where the learner can simultaneously hear the speech and see the salient word referents that are being discussed about. The idea is that the tags enable grounding of the learned word patterns to their visual referents. Since are typically multiple keywords and multiple referents for each utterance, the grounding is essentially a cross-situational learning problem. During the discovery of initial word patterns, no visual tags were utilized.

3. METHODS

3.1. Unsupervised word pattern discovery

The algorithm used to discover word patterns from speech is based on transitional probabilities (TPs) of atomic acoustic events [5]. On a conceptual level, the pattern discovery process can be considered as a spectrotemporal clustering process in which temporally distributed patterns are assigned into a non-predefined number of clusters (models) based on their temporal and spectral similarity.

The atomic acoustic events are vector quantized (VQ) speech frames: features are obtained from standard MFCC extraction (12 coefficients, 32 ms window, 10 ms frame shift). A subset of MFCCs from the training data is then passed to k-means clustering in order to obtain a codebook of size N = 150, and all vectors are then quantized using the codebook. The pattern discovery algorithm analyses TPs between the VQ-indices at several lags (temporal distances), and builds a non-predefined number of TP-based models for speech patterns. Creation of new patterns is based on the similarity of TPs between the contents of the current window of analysis, and the TPs characteristic to previously learned models. If the contents of the current window of analysis are sufficiently similar to the previously learned best matching model, the model is updated with the new data. Otherwise a new model is created from the contents of the window. Then the window is shifted forward. Once a set of models $m_1,m_2,\ldots,m_n \in M$ has been learned, it can be used to recognize similar patterns from novel input. The models also automatically segment novel input into a sequence of auditory patterns. When compared to underlying annotation, these patterns typically correspond to words, part-words or often co-occurring combinations of short words [5].

From the perspective of the work reported here, it is most important to note that there is a novelty threshold parameter ϕ in the algorithm that defines how good match is required between a previously learned model and the current signal content under analysis. The higher the threshold ϕ, the more selective the models will be, and the more there will be unique models in order to cover the entire speech material with the high selectivity models. An interested reader is suggested to see [5] for a more detailed description of the algorithm and the related results on unsupervised discovery of word patterns from speech.

3.2. Contextual analysis of words

Once the patterns, or word-like units, *M* have been learned, their occurrences in the context of possible word referents are studied. The referents simulate visual input to the learner and the assumption is that the learner can perform categorical perception of the visual world so that the referents can be represented as a set of unique discrete tags $c_1, c_2, \ldots, c_j \in C$. The tags correspond to the keywords annotated in the CAREGIVER corpus.

In order to study the context of each auditory pattern, we apply the principles of a technique called random indexing[1] (RI; [11,12]). More specifically, each word referent c_i is assigned with a random and unique sparse vector $\mathbf{v}_i$ of length L that contains mainly of zeros, but has a small number of elements with +1 and -1 values at random dimensions. Since the vectors are long and sparse, each randomly generated vector $\mathbf{v}$ is approximately orthogonal to all other vectors corresponding to other word referents. Also, a zero matrix $\mathbf{G}$ of size MxL is initialized so that each row in the matrix corresponds to a unique auditory pattern discovered in the earlier stage. This will be called the context matrix.

During learning, the speech material is fed to the system utterance by utterance. For each utterance, the simultaneously present word referents $\mathbf{c} = \{c_1,c_2,..,c_n\}$ are converted into corresponding context vectors $\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$, and the vectors are summed into an overall visual context $\mathbf{v}_{\text{cont}} = \mathbf{v}_1+\mathbf{v}_2+\ldots+\mathbf{v}_n$. Then for each auditory pattern m_i present in the utterance, the visual context vector is added to the corresponding row i in the context matrix $\mathbf{G}$. After the entire training data is preprocessed, the rows of $\mathbf{G}$ denote typical visual contexts in which the auditory patterns M occur. This information serves two purposes: 1) The strength of association between auditory pattern m_i and visual referent c_i can be computed by simply computing $\boldsymbol{h}_i\boldsymbol{v}_j^T$, where $\boldsymbol{h}_i$ is the i:th row of $\mathbf{G}$. 2) The degree of synonymy (with respect to visual context) between different models M can be computed by normalizing the rows $\mathbf{G}$ to unit vectors and then computing $\mathbf{S} = \mathbf{G}\mathbf{G}^T$. In this representation, elements $\mathbf{S}(i, j)$ obtain a value between -1 and 1 depending on the similarity of the contexts in which i and j typically occur, indicating their functional equivalence.

3.3. Merging of synonymous words

Once the degrees of synonymy between word models are known, the word models sharing sufficiently similar contexts can be merged. In the experiments of this paper, the synonymy matrix $\mathbf{S}$ was analyzed incrementally row by row. For each row i, the

[1] Note that RI is not necessary to achieve grounding of word forms, but simple joint distribution $p(c,w)$ would suffice. However, RI allows flexible representation of more complicated contexts of multiple items and the memory requirements do not scale exponentially with the vocabulary size.

maximally synonymous column j ($i \neq j$) of **S** was searched for. If the degree of synonymy satisfied $\mathbf{S}(i,j) > \delta$, where δ was a user defined parameter, the models m_i and m_j were merged together. Then the models i and j were excluded from further merges, and the analysis proceeded to the next row i+1 of **S** (naturally only if i+1 ≠ j). The process was repeated for each row of **S**, merging all models to the most synonymous one if their mutual synonymy exceeded the threshold δ. Note that if the merging would be performed based on the acoustic similarity instead of contextual similarity, the result would correspond to smaller novelty threshold ϕ in bottom-up word learning, leading to less-constrained clustering of the patterns.

The actual merging of models was performed as follows: the word models of the algorithm are in practice sets of matrices **P** explaining TPs between sequence elements at different temporal lags, and the TPs are always derived from the corresponding frequency matrices **F** for transitions $f_i(a_y|a_x,k)$ at lag k [5,13]. Therefore the model combination is achieved directly by summing frequency matrices of m_i and m_j so that $f_n(a_x|a_y,k) = f_i(a_y|a_x,k) + f_j(a_y|a_x,k)$ for all x,y and k, and then normalizing them according to the normal procedure in Räsänen [5] in order to obtain model specific transition probabilities.

3.4. Evaluation

The quality of the word models and the accuracy of the grounding was measured using a previously unseen test set of speech data. For each frame n in an utterance, the most likely word pattern m_i was determined. Then the corresponding association strengths $A(n,c)$ between the sparse representation of m_i and all visual referents C were computed from **G** (see section 3.2). The cumulative activation $A(c)$ over the entire utterance was obtained by summing over all frames n, and the K most activated referent hypotheses were compared to the ground truth with K true referents. The word association performance (WAP) was defined as the proportion of correct hypotheses over the entire test set.

The merging process was evaluated iteratively using the following scheme: 1) The entire test data was recognized using the learned models M_t and the RI-based context matrix $\mathbf{G}_t$ that had been computed based on co-occurring visual referents. The word association performance was evaluated using the grounding information derived from $\mathbf{G}_t$. 2) Synonymous pattern models were merged together based on $\mathbf{S}_t$ derived from $\mathbf{G}_t$ in order to form a smaller set of models M_{t+1}. 3) The entire corpus was analyzed again using the new models M_{t+1} in order to learn a new context matrix $\mathbf{G}_{t+1}$. Then the test set was recognized again with the new representation. The process in 1)-3) was iteratively repeated as long as there were at least one model pair m_i and m_j that exceeded δ in synonymy. It is acknowledged that this was somewhat unrealistic learning situation because the same speech tokens were perceived by system after each iteration. However, it was a necessary simplification due to the finite amount of data that was available for the experiments. Naturally, the signals in the test were never used in the training but were simply used to probe the performance.

As an outcome of the evaluation process, the WAP was obtained as a function of the total number of word models. The hypothesis was that the total number of word models can be decreased from the originally discovered set using the proposed merging scheme, and that the merging could be done without essential loss in word association performance. This would lead to a more compact set of pattern detectors that would respond and generalize selectively to specific words in the audio. In other words, the proposed model combination should yield higher WAPs for models that are learned from bottom-up statistics and refined by additional contextual constraints than in the case where the same number of word models are learned purely based on acoustic similarity.

4. RESULTS

The word association performance was first measured in a purely bottom-up approach (no merging) by varying the novelty threshold $\phi \in [0.04,\ 0.059]$ of the pattern discovery algorithm, leading to the discovery of varying number of word patterns. The analysis window length of PD was 480 ms with 240 ms window shift between the frames (see [5]). For random indexing, randomly generated hyperdimensional vectors of length L = 1500 with 15% of non-zero components were used.

Blue lines with squares in Fig. 1 show the results from the single speaker experiment. As can be observed, the grounding is relatively successful, yielding a WAP of 67% correct associations between the audio and the referents. Note that, the optimal number of word models is notably higher than the true number of words in the material (455 vs. 80). This suggests that even in the presence of a small number of non-selective "trash" models, the overall number of parallel models for acoustic variants of each lexical entry is quite large.

In the merged condition, the original patterns from the highest threshold (N = 455 patterns, ϕ = 0.059) case were used a starting point for merging. The blue dashed line in Fig. 1 shows the WAP on the single speaker test set, probed while the merging proceeds with the synonymy threshold δ = 0.8. As can be observed, the association strength between audio patterns and the visual referents does not decrease very notably while the total number of models is decreased to 12% of the original number (N = 56). When the performance is compared to bottom-up learning result with an equal number of word models, the difference is almost 10 percentage units in the favor of the merged models.

Red lines with circles in Fig. 2 show the WAP for speech from one male and female talker. Again, the quality of word form to word referent associations increases as the number of learned patterns is increased. Also, the optimal number of acoustic pattern models is notably larger than the true size of the lexicon of the material. Again starting from the maximum number of word models obtained in bottom-up pattern discovery (N = 737), the merging was performed to combine models systematically sharing similar visual context. The result shows that the number of models reduces to less than fourth of the original, leading only to small degradation in performance (from 63% to 56%). When compared to the performance with the same number of purely bottom-up models, the difference is 16 percentage units in favor of top-down merging, indicating notable enhancement in model selectivity and generalization.

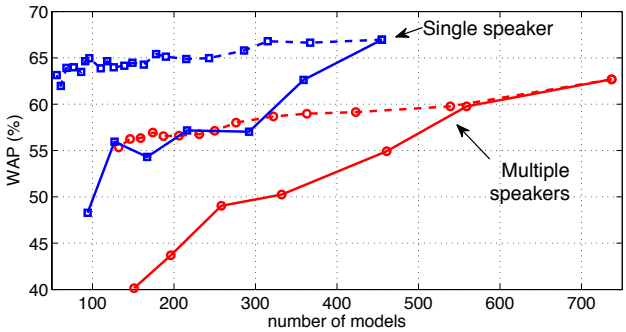

Fig. 1: Word association performance (WAP) for single speaker (blue line with squares) and multiple speakers (red line with circles). Results from bottom-up learning (solid line) and after top-down merging (dashed line) are shown separately.

5. DISCUSSION AND CONCLUSIONS

The focus of this work was to study the quality of word models when they are first learned in a purely unsupervised manner from speech, and only later grounded to external referents such as objects and events in the environment. It was also studied how the selectivity and generalization capability of the word models change when parallel models representing same lexical items are merged together based on their synonymy (or functional equivalence).

The obtained results suggest that the bottom-up acoustic learning can lead to notably above chance-level in word to referent associations. However, the number of word models is much higher than the true number of lexical items in the data. When contextual information regarding the visual referents corresponding to the spoken utterances is provided to the learner, the learning algorithm is able to estimate the degree of synonymy of the learned models. By merging the models with a high degree of synonymy, the overall number of internal representations for words can be decreased notably with only a minor impact on the ability of the lexicon to account for previously unseen word tokens. On the other hand, generalization ability of each single model increases notably in the process. This can seen as increased word association performance of the contextually merged word models when compared to the same number of models learned purely on the basis of acoustic similarity.

The current work also provides the first transitional probability based word learning framework in which words are first discovered from continuous speech based on their acoustic similarity and only later associated to contextual referents through cross-situational learning. As can be observed, the word association performance is not perfect. Actually, it is notably worse than in a learning process where the acoustic patterns are directly learned in the context of relevant grounding information. For example, Räsänen and Laine report a keyword recognition rate of above 92% in the same task using the TP analysis algorithm in a weakly-supervised training mode [13], whereas the current unsupervised approach achieves only 67% on the same performance scale. The same experiments have also been conducted on a simpler 10 keyword material of CAREGIVER Y1 corpus, leading to approximately 96% WAP with indirect grounding using the currently discussed methodology (unpublished results), whereas WAP of 100% is achieved in weakly-supervised learning mode [14].

However, the difference between indirect and direct grounding is not surprising because the statistical constraints available in the in the weakly-supervised situation are much stronger. For example, in discovery of a word, a weakly-supervised algorithm can take into account only those utterances that are present concurrently with the respective visual referent. Also, the number of unique lexical items can be derived from the number of visual referents in the training set (i.e., the algorithm is able to listen to speech in the context of each specific referent). When the patterns are discovered in an unsupervised manner, the only constraints are provided by the statistics of the auditory stream itself. This necessarily leads to sometimes vague or inaccurate models that do not precisely correspond to true words (see [5] for detailed analysis). However, this is also what is observed in young infants when they are learning their first words (see [2] and references therein), and also allows learning of representations for lexical items that do not have directly perceivable referents.

Given the current result and the ones obtained earlier in [5], two main hypotheses can be formulated that should be studied in more detail in future:

1) Given a distributional framework for unsupervised word learning from purely acoustic signals, it seems that the acoustic variability across different realizations of words is too high for a learner to directly achieve high-quality speaker-independent word models for each lexical item in the familiarization data. Instead, multiple parallel representations for words are discovered, corresponding to realizations of the words in different linguistic contexts or spoken by different talkers.

2) The quality and generality of the original proto-lexical items becomes refined as the learner accumulates experience of situated spoken language with caregivers and other people. The role of contextual referents is important in this process, allowing the discovery of functional equivalence between acoustically different patterns (or distinctiveness between acoustically similar patterns) that would be otherwise impossible to derive purely on the basis of speech signal properties.

6. ACKNOWLEDGEMENTS

This research was funded by the Finnish Graduate School for Language Studies (Langnet) and by Nokia Foundation.

7. REFERENCES

[1] Houston D. and Jusczyk P., “Infants’ long-term memory for the sound patterns of words and voices,” *J. Exp. Psychology: Human Perception and Performance*, Vol. 29, pp. 1143-1154, 2003.

[2] Werker J. and Curtin S., “PRIMIR: A Developmental Framework of Infant speech Processing,” *Lang. Learning and Development,* Vol. 1, pp. 197-234, 2005.

[3] Vallabha G., McLelland J., Pons F., Werker J. and Amano S., “Unsupervised learning of vowel categories from infant-directed speech,” *Proceedings of National Academy of Sciences*, Vol. 104, pp. 13273-13278, 2007.

[4] Kouki M., Kikuchi H. and Mazuka R., “Unsupervised Learning of Vowels from Continuous Speech Based on Self-Organized Phoneme Acquisition Model,” *Proc. Interspeech’2010*, pp. 2914-2917, 2010.

[5] Räsänen O., “A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events,” *Cognition*, Vol. 120, pp. 149-176, 2011.

[6] Minematsu N., Qiao Y., Asakawa S. and Suzuki M., "Speech structure and its application to robust speech processing," *Journal of New Generation Computing*, Vol. 28, No. 3, pp. 299-319, 2010.

[7] ten Bosch L., Van hamme H., Boves L. and Moore R.K., “A computational model of language acquisition: the emergence of words,” *Fundamenta Informaticae*, Vol. 90, pp. 229-249, 2009.

[8] Räsänen O., Laine U.K. and Altosaar T., “A noise robust method for pattern discovery in quantized time series: the concept matrix approach,” *Proc. Interspeech '09, Brighton*, England, pp. 3035-3038, 2009.

[9] Aimetti G., “Modelling early language acquisition skills: Towards a general statistical learning mechanism,” In Proc. EACL-2009-SRWS, Athens, Greece, pp. 1-9, 2009.

[10] Altosaar T., ten Bosch L., Aimetti G., Koniaris C., Demuynck K. and van den Heuvel H., “A Speech Corpus for Modeling Language Acquisition: CAREGIVER,” *Proc. Int. Conf. on Language Resources and Evaluation*, Malta, 1062-1068, 2010.

[11] Kanerva, P., Kristoferson J. and Holst A., “Random Indexing of Text Samples for Latent Semantic Analysis,” *Proc. 22nd Annual Conference of the Cognitive Science Society*, 2000.

[12] Sahlgren M., “An introduction to random indexing”, *Methods and Applications of Semantic Indexing Workshop, 7th Int. Conf. on Terminology and Knowledge Engineering*, 2005.

[13] Räsänen O. and Laine U., "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences,” *Pattern Recognition*, Vol. 45, pp. 606-616, 2012.

[14] ten Bosch L., Räsänen O., Driesen J., Aimetti G., Altosaar T., Boves L.: "Do Multiple Caregivers Speed up Language Acquisition,” *Interspeech '09*, Brighton, England, pp. 704-707, 2009.