Automatic self-supervised learning of associations between speech and text

Juho Knuuttila, Okko Räsänen, Unto K. Laine

Department of Signal Processing and Acoustics, School of Electrical Engineering,
Aalto University, Espoo, Finland
{juho.knuuttila , okko.rasanen, unto.laine} at aalto.fi

Abstract

Discovery of statistically significant patterns from data and learning of associative links between qualitatively different data streams is becoming increasingly important in dealing with the so-called Big Data problem of the modern society. In this work, a methodological framework for automatic discovery of statistical associations between a high bit-rate and noisy sensory signal (speech) and temporally discrete categorical data with different temporal granularity (text) is presented. The proposed approach does not utilize any phonetic or linguistic knowledge in the analysis, but simply learns the meaningful units of text and speech and their mutual mappings in an unsupervised manner. The first experiments with a limited vocabulary of child-directed speech show that, after a period of learning, the method is successful in the generation of a textual representation of continuous speech.

Index Terms: statistical learning, associative learning, multimodal processing, unsupervised learning, self-supervised learning

1. Introduction

Conceptual learning in the human cognitive system never occurs inside a single modality, but in terms of associations between representations in multiple perceptual modalities and motor outputs. As the events in our environment often provide information through multiple modalities, the learning can also occur through co-occurrences of structured activities at different modal dimensions. In this context, pattern discovery in one modality is basically only data segmentation or clustering and the created clusters are meaningless without grounding through *multimodal associations*.

The main purpose of this study is to develop general methodology to analyze, discover, and model associations between two qualitatively different data streams. We demonstrate how an unsupervised pattern discovery problem can be turned into a self-supervised learning process where automatically derived representations in one modality can be used to aid discovery of patterns in another modality. The two "modalities" used in the current work are spoken English utterances and textual representations corresponding to the utterance contents but with all white spaces and special characters removed. Speech and text are chosen as the data types for this study due to the fact that despite their strong mutual interdependency, they are clearly qualitatively different, have different temporal characteristics, but the results of the associative pattern discovery are still easy to evaluate. Most importantly, the basic units in text (i.e., letters) do not have direct correspondence to any units of speech that could be defined purely based on the raw acoustic signal and the pronunciation of the letters depends on the lexical context so that each letter becomes realized in various acoustic forms. In general, one can hypothesize that the statistical connection between structure of speech and text is most significant at the level of word-like patterns of speech and text instead of the low-level feature/letter representations of the modalities. In order to learn the dependencies between the modalities, one has also to learn these temporally spanning patterns first. Although only text and speech are used in this study, the same proposed concept should be applicable also to other pairs of sequential data streams that model the same phenomena or are hypothesized to have high correlation for some other reason.

Note that the current approach is in contrast to the automatic speech recognition systems (ASR) that rely heavily on linguistics and on recognition of phonemes or their combinations. These linguistically motivated units are recognized mostly without any semantic component or associations to other modalities or information sources. To make a clear difference to such systems, it should be emphasized that the present approach is based merely on the signal statistics within and between the two different data streams, and not based on any prior phonetic or linguistic expert knowledge. Similarly, a child learning multimodal association does not have this kind of expert knowledge and is still able to learn to speak and to understand spoken messages.

In the following sections, a self-supervised method for building associations between two sequential data representations is proposed. Its performance is demonstrated with experiments using orthographically annotated speech corpus with limited vocabulary. In the experiments, the labeling for the patterns in the speech signal is created from the annotation text using three different variants of greedy grammar inference. The three variants are compared by their respective speech-to-text transcription capabilities.

2. Cross-modal statistical learning

Written English is not based on systematic coding of its sounds to orthography, therefore, most of the time it is not possible to transform temporal structures of continuous speech directly into a sequence of characters. Continuous speech can be transformed to a sequence of vector quantized (VQ)-indices each representing a momentary spectral representation. However, it is difficult to find any relevant associations between this index sequence and the corresponding orthography directly. Both streams (VQ-indices and sequence of characters) must be first pre-processed to discover patterns spanning larger intervals inside both streams. In continuous text these patterns are subsegments of utterances, or word-like units, resembling syllables, words, or phrases. These statistically discovered word-like units are then used in labeling of the speech signal in a weakly supervised learning process.

The proposed learner for cross-modal associations operates offline. The goal of the learning is to maximize the predictabil-

ity of the text stream given the speech stream. First, a context-free grammar (CFG) is inferred from the text stream. After the CFG is inferred from a collection of utterances, it is used to extract contextual label(s) for each individual utterance. The labels are the root nodes of CFG bottom up parse trees (Fig. 1). Then, these labels, each corresponding a non-terminal symbol, can be used also as indices and are needed for concept matrix concept matrix (CM) algorithm [1], a weakly supervised pattern discovery and recognition algorithm for sequential data. The CM attempts to find the relationship between the text patterns and their acoustic counterparts. The CM algorithm is used in this study because it is already proved to be effective in the weakly supervised pattern discovery with weakly aligned contextual labeling. The grammatical inference for text stream is selected for the generative property of CFG and its usefulness in the study of discrete sequences.

The prediction capability, the ability to build hypotheses, is acquired only to one direction (speech-to-text) because CM algorithm is not generative. This leads to the definition of the patterns in the both streams: the patterns in the text representation are exact segments of the stream itself and in speech stream they are statistical models of element-to-element transitions in VQ representations. The grammatical inference and the CM algorithm are presented in the following subsections.

2.1. Greedy grammatical inference of patterns from text

In order to first discover patterns inside the text modality, a CFG in Chomsky normal form is inferred from the textual representations of utterances in an unsupervised manner. The production rules are selected along agglomerative compression of the textual representations: during each iteration, a symbol pair $a_k a_l$ is selected and every instance of the pair in the data is replaced by a new non-terminal symbol a_n. The corresponding production rule $a_n \rightarrow a_k a_l$ is appended to the CFG. The idea of this type of agglomerative compression was first presented by Solomonoff [2], who proposed that the pair with the highest frequency of occurrence (Freq) should be always selected, resulting in maximal data compression. However, always selecting the highest frequency pair for the agglomeration does not necessarily produce a grammar that is optimal for pattern recognition purposes. In addition to using frequency of the pairs as the criterion, other possibilities include, e.g., mutual information (MI) information gain (IG), or increase in entropy-rate [3] as the criteria for selection of the coded pair. It has been pointed out in [4] that MI is susceptible to estimation errors especially when the frequency of occurrences both symbols are rare. The IG, as formulated in [5], could alleviate the problem since it measures the amount of information in a symbol (in bits) when both left- and right contexts of the symbol are taken into account.

$$
\begin{aligned}
&IG(a_k, a_l) = \\
&P(a_k, a_l)\log\frac{P(a_k, a_l)}{P(a_k)P(a_l)} + P(a_k, \bar{a}_l)\log\frac{P(a_k, \bar{a}_l)}{P(a_k)P(\bar{a}_l)} + \\
&P(\bar{a}_k, a_l)\log\frac{P(\bar{a}_k, a_l)}{P(\bar{a}_k)P(a_l)} + P(\bar{a}_k, \bar{a}_l)\log\frac{P(\bar{a}_k, \bar{a}_l)}{P(\bar{a}_k)P(\bar{a}_l)}, \quad (1)
\end{aligned}
$$

where $P(a_k, \bar{a}_l) = \frac{Freq(a_k) - Freq(a_k, a_l)}{N-1}$ and $P(\bar{a}_k, \bar{a}_l) = 1 - P(a_k, \bar{a}_l) - P(\bar{a}_k, a_l)$.

As the CFG estimation proceeds, the sequence is gradually compressed. Along the compression and the increase of the total amount of different symbols in the grammar, the reliability of statistical estimator for probability of a symbol pair $\hat{P}(a_k, a_l)$ deteriorates. The low reliability indicates that the new production rules included in the grammar are no longer based on reliable structural information in the data, and therefore the goodness of statistics can be used as an automatic stopping criterion for the inference. Lesne et al. [6] have defined the reliability of pair statistics in the following form :

$$\max P(a_k, a_l) N_{eff} \gg 1, \quad (2)$$

where N_{eff} is the effective length of the compressed utterance:

$$N_{eff} = \frac{Nh}{\ln K} \quad (3)$$

where N is the actual length of the compressed utterance, h the entropy rate and K number of the unique symbols in the compressed sequence. In this work, this criterion is used to stop the inference process when Eq. (2) is no longer satisfied (N was required to have a value of > 20 in order to continue inference).

2.2. Associative learning between text patterns and audio patterns

The raw speech data is continuous valued by nature and is not suitable for the CM algorithm, which requires sequential data with a finite alphabet. For the CM algorithm, the speech samples are preprocessed. First, the original speech samples are downsampled to 16 kHz. Second, 12-dimensional Mel-frequency cepstral coefficient (MFCC) vectors are extracted from 32 ms long Hamming widowed sub-segments of the speech signal using a window shifts of 10 ms. Then, the MFCC-vectors are clustered to $N_A = 128$ clusters with the k-means clustering with Euclidean distance measure. The initial clustering was performed on a random subset of 15,000 MFCC vectors from the training set and then all of the vectors from all utterances were assigned to a their closest centroids. Finally, every vector is replaced by their cluster id's. This results the VQ representation of the speech signal at 100 Hz symbol rate and alphabet size of N_A.

In order to discover the mapping between the text and the audio, the CM algorithm [1] is trained with VQ sequences and categorical contextual information related to the VQ sequences. The algorithm learns to extract important segments of the VQ sequences by accumulating element-to-element transition probability statistics for different contexts. The learned models for each context are normalized transition probabilities from a VQ element to another at different lags k in the context of different labels c. For example, a lag of two means that the analyzed elements are separated by one undefined element in between. The alignment between a context and a VQ sequence does not need to be exact. It is enough that the consecutive VQ elements related to the context are somewhere in the sequence. The relevant element-to-element transition probabilities start to stand out in the model of the context as more evidence is accumulated.

In the recognition, for a given VQ sequence, the CM model provides for each element-to-element transition a conditional probability that the actual transition belongs to certain context. The probabilities are given for each trained lags (within each context) and combined to produce model activation $A(c, t)$ for each trained label c at each time frame t for the final recognition/classification [1].

In this study the contextual labels $\mathbf{c}$ for individual spoken utterance are the root nodes (non-terminals) of the CFG bottom up parse trees (Fig. 1). These contexts could be used as either unordered or ordered set in the input for the CM algorithm [1].

Since the two streams (speech and text) are different representations of the same phenomena, we use some a priori knowledge in that the patterns are in the same order in both representations. VQ representations can be segmented in order to discard the VQ elements that are almost surely not related to the text pattern.

It is hypothesized that the pattern boundaries are approximately in the same relative positions in the both representations. Due to the nature of English language and the existence of silence periods in speech, it cannot be guaranteed that all the relevant VQ elements are between the boundaries suggested by the corresponding text pattern (the dashed vertical lines in Fig 1). For this reasons the length of VQ sequence used in the training together with a text pattern is extended by experimental factor of 1.4 . For example if a VQ segment suggested by the text pattern is $v_i v_{i+1} \ldots v_{i+n-1}$, where v is a VQ element and n is the length of the segment, then the extended segment is $v_{i-0.7n} \ldots v_{i+1.7n-1}$ (the highlighted portion of the sliced bar representing VQ sequence in Fig. 1).

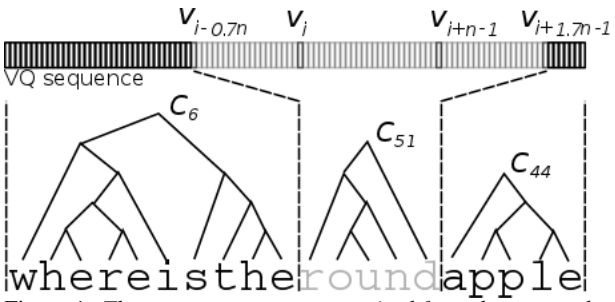

Figure 1: *The text patterns are recognized from the root nodes (the C's) of CFG bottom up parse trees. The CM is trained with longer segments of VQ representation than text the pattern suggests. The segment is extended 70 percents from both ends.*

2.3. Hypothesizing the context labels from a VQ sequence

The recognition process for a VQ sequence consists of first selecting an ordered set of labels based on model activations provided by the CM algorithm. Second, the labels are transcribed to the corresponding text patterns and final hypothesis is formed by concatenation of the patterns.

First, in the process of selecting the labels, instantaneous model activation values are median filtered with decay factor $\gamma = 6$ to smoothen the results. Then, a short-term sliding window is used to sum the filtered activation value outputs $A(c,t)$ into local pattern probability estimates. The integral of activation values for pattern c at time t with window length T is:

$$A_{tot}(c,t) = \Sigma_{t-T/2}^{t+T/2} A(c,t) \quad (4)$$

The label that has the highest integral across the window is selected for the hypothesis for the short time interval. A sequence of label hypothesis is acquired by selecting a hypothesis every Ts frames for duration samples. Removing consecutive duplicates from the sequence of label hypotheses forms the final ordered set of labels. The sliding window size T and the step size Ts are user defined parameter values and they are estimated from the training set of utterances.

3. Experiments

In the current experiments, the learning algorithm is given a training set of utterances with the textual and VQ representations in order to build the statistical association across the two representations. The measure of the success of pattern discovery and associative learning is the fidelity of the textual representations that are derived from the given audio utterances of a disjoint test set.

3.1. Used material

The material used in the experiment is taken from the CAREGIVER corpus [7]. The corpus contains infant directed spoken utterances in multiple languages and speakers. Along the speech signals there are corresponding orthographic transcripts of each utterance. From the Y2-version of the corpus 2397 utterances of a single English speaker are randomly divided to disjoint training and testing sets of sizes 2000 and 397 utterances respectively. The orthographic annotation of each utterance was modified into *a continuous text string* by removing all the special characters, including the whitespaces.

3.2. Extraction of label - VQ sequence pairs for the CM algorithm

The textual representations of utterances in the training set are concatenated, and the grammatical inference, described in the subsection 2.1, run until the stopping criterion in the Eq. (2) is met. The resulting CFG is used in bottom up parsing of individual utterances. The root nodes representing the word-like units of text are then used as labels in the CM algorithm. The information about the portions that the recognized patterns occupy in the textual representations is used in segment extraction from the corresponding VQ representations. The label recognition and segmentation of VQ representation is illustrated in Fig. 1. The extraction is performed to all training utterances and the CM matrices are trained with resulting label-VQ segment pairs using lags: $k \in \{1, \ldots, 13\}$.

3.3. The measure of predictability

A variant of Levenshtein distance [8] between hypothesized textual representation of utterance and corresponding reference is used as the measure of predictability. The Levenshtein distance is the minimum amount of single character edits (insertion, deletion, and substitution) needed to convert a string to another string. Here, a variant referred as *edit distance* is used where the weight of insertion and deletion is one and the weight of substitution is two. The measure of predictability of a set of utterances given the corresponding speech signals is the *sum of the edit distances* (SED) of individual hypotheses in the set.

3.4. Estimation of optimal parameter values

The optimal parameter values for sliding window length and for the step size are evaluated from the training set with 5-fold evaluation. In the each fold a disjoint set of utterances are used as the evaluation set. The SEDs of the evaluation sets are computed with different parameter values and averaged over the 5 folds. The estimates for optimum parameters values are the ones that give the minimum average SED.

3.5. The results

The experiments were run with three different variants of CFG. In each variant the selection criteria for the symbol pairs in the grammatical inference was varied. The best results, given by IG, are presented in detail, and the results given by MI and Freq are just briefly summarized. The MI measure is the first term of the IG in the Eq. (1).

The averaged SEDs over the 5-fold evaluation with different parameter values are presented in Fig. 2. The minimum SED in the evaluation is reached with a sliding window of size 48 samples and a step size of 17 samples. The SED across the test of 397 utterances with the estimated parameter values is 1868. The success of the estimation is evaluated by comparing the SED given by the estimated parameter values to the optimal SED given by parameter values that are optimized for the test set. The optimal SED for the test set is 1826 given by 46 and 17 samples long window and step size respectively. The heat map for parameter optimization in the test set is left out, but it largely reminds Fig 2. That implies the robustness of the parameter evaluation.

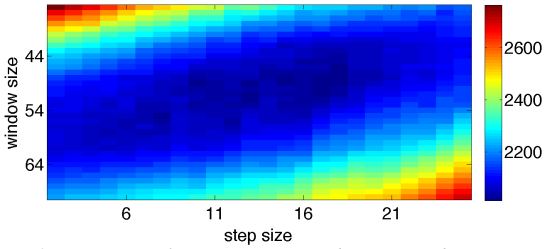

Figure 2: *Heat map for parameter evaluation in the IG variant. Each tile's color represents the average SED in the evaluation sets of 400 utterances with corresponding step and sliding window sizes. The estimates for parameter values are the ones that give minimum SED. The optimum is reached with sliding widow size 48 and step size 17.*

The results of each CFG variant are summarized in the table 1. All the estimated parameter values and the corresponding SEDs are compared to the corresponding optimized values. In all variants the parameter values are quite reliably estimated since the SEDs given by those estimated values differs only 2.4%-5.1% from the optima. In terms of SED the IG variant outperforms the MI variant by 8.7% and the Freq variant by 40.6%. More comprehensive results are presented in [9].

Table 1: *Summarized results with the three variants. The table includes the parameter values estimated from the training set and the corresponding SED in the test set. For comparison the optimized parameter values for the test set and corresponding SED are also presented*

		window size	step size	SED of the test set
IG	estim	48	17	1868
IG	optim	46	17	1826
MI	estim	48	14	2045
MI	optim	47	17	1942
Freq	estim	47	1	3144
Freq	optim	48	7	3044

The individual edit distances with the estimated parameter values of IG variant are examined in more detail. 26,7% of the hypotheses produced by the variant were correct. In order to compare edit distances of utterances with different lengths, individual edit distances are normalized with the corresponding reference utterance length. The resulting measure is referred as *relative edit distance* (RED). In the Fig. 3 the REDs are ordered to ascending order by the reference utterance length. The variation in the recognition of single word utterances is large. Many of the hypotheses are correct and some are way off. In general the length of the multi-word utterances does not have a remarkable effect to the prediction accuracy. This is illustrated in the

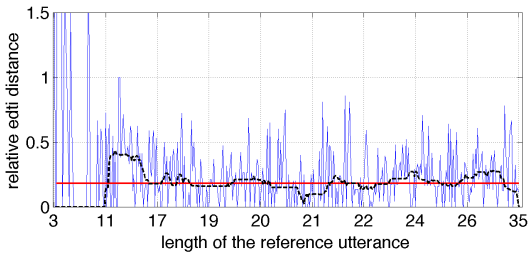

Figure 3: *relative edit distances ordered by the corresponding reference utterance lengths. The individual edit distances are plotted with thin blue line. The bold dashed line is median filtered smoothing with 30 samples long window. The straight line is the median of relative edit distances in the test set. Note that x-axis is not linear, the tick labels denote the length of the reference utterance in that index.*

figure by moving median of 30 samples (bold dashed line) and the median over the whole test set (0.182, solid straight line). This means that on the average over 80% of the characters in the hypothesized utterances are correct.

4. Discussion and summary

The results of these first experiments are promising, although there is still some room for improvement. Optimizing the segmentation of VQ representations in the training phase and trying different filtering methods of the CM output in the recognition could improve the results.

If the presented methodology would have been treated like ASR a statistical language model could have been created for the recognition. The model would include for example estimated conditional probabilities of the text patterns given the previous pattern. The usage of a such model would allow the creation of multiple hypothesized text pattern sequences and selecting the most probable. For example there are some cases where two text patterns, that both exist only in the beginning of utterances in training, are hypothesized to the beginning of an utterance: *"hereihereisa..."*. At least these kind of errors would be eliminated.Creation of a language model was not tried because the domain expertise was kept to minimum and the methodology as general as possible.

The nature of the corpus was potentially beneficial. The vocabulary is statistically balanced. The utterances are grammatically correct but not necessarily sensible or logical. Since there are less word-to-word dependencies, it is more likely that the grammatical inference first discovers the text patterns matching real words and then joins them, rather than discovering text patterns spanning over two incomplete words.

A novel method for discovering associations between two co-occurring and qualitatively different sequential data streams was presented. The self-supervised method is based on first discovering patterns with grammatical inference from textual representation and then using them as labels in weakly supervised learning in the vector quantized speech representation. The method was shown to acquire promising predictability of the textual representation when given the corresponding speech signal.

5. Acknowledgements

This work is a part of the TOMU4 project funded by Nokia Research Center Tampere.

6. References

[1] O. Räsänen and U. K. Laine, "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences," *Pattern Recognition*, vol. 45, no. 1, pp. 606–616, Jan. 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0031320311002044

[2] R. J. Solomonoff, "A formal theory of inductive inference. part II," *Information and Control*, vol. 7, no. 2, pp. 224–254, Jun. 1964.

[3] U. Laine, "Entropy-Rate Driven Inference of Stochastic Grammars," *Twelfth Annual Conference of the International Speech*, no. August, pp. 2489–2492, 2011.

[4] M. Wu and K. Su, "Corpus-based Automatic Compound Extraction with Mutual Information and Relative Frequency Count," *Proceedings of ROCLING VI*, 1993. [Online]. Available: http://www.aclclp.org.tw/rocling/1993/M09.pdf

[5] C.-C. Wong, H. M. Meng, and K.-C. Siu, "Learning strategies in a grammar induction framework," in *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, November 27-30, 2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan*, 2001, pp. 153–157.

[6] A. Lesne, J.-L. Blanc, and L. Pezard, "Entropy estimation of very short symbolic sequences," *Physical Review E*, vol. 79, no. 4, pp. 1–10, Apr. 2009. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.79.046208

[7] T. Altosaar, L. ten Bosch, G. Aimetti, C. Koniaris, K. Demuynck, and H. van den Heuvel, "A speech corpus for modeling language acquisition: Caregiver," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010.

[8] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966.

[9] J. Knuuttila, "untitled," Master's Thesis, Aalto University, To be published 2013.