Unsupervised detection of words – questioning the relevance of segmentation

Louis ten Bosch [1], *Hugo Van hamme* [2], *Lou Boves* [1]

[1] Language and Speech, Radboud University Nijmegen, the Netherlands
[2] ESAT, Katholieke Universiteit Leuven, Belgium

l.tenbosch@let.ru.nl

Abstract

In this paper, we discuss a computational model of language acquisition which focuses on the detection of words and that is able to detect and build word-like representations on the basis of multimodal input data. Experiments carried out on three European languages (Finnish, Swedish, and Dutch) show that internal word representations can be learned without a predefined lexicon. The computational model is inspired by a memory structure that is assumed to be functional for human cognitive processing. The model does not use any prior segmentation, nor does it use the concept of segmentation later in the processing. This calls into question the importance that is conventionally attributed to the segmentation of the speech signal in terms of symbolic units for the purpose of detecting structure in speech.

Index Terms: cross-modal learning, language acquisition, computational modelling, word detection.

1. Introduction

In nearly all disciplines dealing with speech, e.g., phonetics, speech technology, and psycholinguistics, it is assumed that the speech signal can be adequately represented in terms of sequences of discrete symbols. This idea of 'beads-on-a-string' ([7]) is manifest at several different layers: a speech signal can be described in terms of e.g. phone- or phoneme-like units or words, and multi-tier representations can exploit the relations between different tiers thanks to the symbolic representations in these tiers. Symbolic representations are also used in mainstream approaches in Automatic Speech Recognition and in models of human speech processing such as Shortlist ([6]).

When using symbolic sequences, one tacitly assumes that each symbol in the sequence corresponds with a particular stretch of speech in the input signal, and that these stretches can be concatenated (with hard or soft boundaries) to reproduce the speech signal. This assumption, however, has a problematic character. For example, it is well known that human transcribers may substantially disagree about the transcription and the segmentation of speech at the phone-level. Nevertheless, in many cases symbolic representations are extremely useful, be it in the form of a single symbolic sequence (first-best) or as a graph.

Segmentation and the availability of *symbolic* representations go hand in hand. In speech corpora, segmentation information might be explicitly provided, such as the segmentation on phone-level in TIMIT, but in most speech corpora information about segmentation is not available. While the segmentation is not provided explicitly, it can often be hypothesised by aligning trained acoustic models with the speech signal, as done in automatic segmentation by ASR, or by bottom-up segmentation approaches ([13]). These former approaches show that segmentation can be seen as a side effect of speech decoding, rather than as an essential ingredient that needs to be available a priori for speech processing.

The issue of symbolic representation and segmentation is reflected in many approaches in modelling speech *processing* and language *acquisition*. For example, there is now considerable evidence from psycholinguistic and phonetic research that sub-segmental (i.e. sub-symbolic, fine-grained, acoustic-phonetic) and supra-segmental (i.e. prosodic) detail in the speech signal help the (adult) listener segment a speech signal into syllables and words (e.g. [19, 20]. And with respect to acquisition, infants face the task to detect word-like units in speech without any prior knowledge about lexical identities or information about the segmentation of the continuous speech signal ([10, 11, 15, 16, 18]). Newborns are not completely blank - they possess an auditory system that has been exposed in the pre-birth period to band-limited sounds with the same type of rhythm and variation as speech. It appears that infants can identify their native language based on stress patterns very soon after birth. A few months old, they are able to segment words and distinguish between familiar and unfamiliar words based on stress patterns (whether or not the word actually means anything to the infant, e.g. [15]). An infant of six months old can distinguish native and non-native vowels ([16]). And within 8 months, infants can segment words based on the statistical patterns in the observed phonotactics (e.g. [18]). After 2 minutes of exposure, infants can then use the statistical properties of the co-occurrence of syllable-sized units to segment novel words. These studies show that young infants are sensitive to the (statistical) structure in the speech signal on the level of phoneme or syllable-sized units. The model that we present here does not directly address the topic about to what extent this implies that phoneme-like units are used to represent words in the internal (mental) lexicon. Episodic theories of speech perception assume that listeners store multiple entries, in the form of detailed perceptual traces ('episodes'). In contrast, experimental data on e.g. perceptual learning in speech recognition are difficult to explain without hypothesizing more abstract phonological representations (features, phonemes or syllables) (for a discussion see [17]).

An interesting result of our endeavours in building computational models of word discovery is that the role of segmentation in general may be overestimated. The model for unsupervised word detection presented here does not require any information about segmentation in the input and does not even use the concept of segmentation in the entire processing. We will show that the emerging representations that result from word detection can be used to segment a speech signal, but the segmentation procedure differs from the segmentation procedure as performed by e.g. HMM-models in many aspects.

In the next section, we discuss this word detection algorithm. The third section presents word decoding results in various

conditions, showing that the model is capable of building and updating internal representations of speech fragments. This part clarifies the status of segmentation in the entire model. The fourth section deals with segmentation in the decoding, while the last section concludes with a summary and outlook for future research.

2. The word detection algorithm

Our computational model of language acquisition is designed in particular to account for the processes involved in word discovery from 'raw' multimodal data ([1]). The model has similarities with the Cross-channel Early Lexical Learning (CELL) model ([8]), but differs from CELL in an essential aspect. CELL makes the assumption that babies represent speech signals in the form of a lattice of pre-defined phonemes. From the perspective of human language acquisition that assumption is questionable. The model presented here avoids the use of any pre-existing representation for decoding the information in the input signals. Instead, the representations in the model *emerge* from the multimodal stimuli that are presented to the model. This is in line with growing evidence that speech and language skills are *emergent* capabilities of a developing communicative system ([3]). The way in which linguistic patterns are stored and used during language acquisition change constantly as these patterns become more numerous and fine-grained, and as the methods needed for processing these patterns become more complex ([10]).

The computational model comprises two interacting modules: a carer module that presents multimodal stimuli to the learner module, and the learner module that simulates the young language learner. The learner operates as follows:

a) the learner receives from the carer a multimodal stimulus, consisting of an utterance (presented as sampled data, without segmentation) in combination with an abstract tag that represents an interpretation of a corresponding visual stimulus. For example, an utterance 'look at this nice ball' is associated with a tag 'ball'. This tag does not necessarily refer to the *word* 'ball', nor does it give any clue about the phonetic realisation of the speech fragment associated to this tag or its position in the utterance.

b) Next, the waveform is converted into a sequence of feature vectors (currently MFCC, but this choice is not relevant for the discussion here).

c) Acoustic events in an utterance are represented by the unigram and bigram counts on the indices of a codebook. A mapping M is defined such that each utterance is mapped to a vector of a fixed (preset) dimension in a vector space S. This vector consists of the unigram and bigram counts. See [12].

d) M has the essential property that the *joint presence* of events in an utterance is translated into an *additive* property in S. For example, an utterance with the events 'A B A' translates into a vector equal to twice the vector associated to 'A' plus one time the vector associated to 'B', that is: $M(\text{'ABA'}) = 2M(\text{'A'})+M(\text{'B'})$ (see sect. 4).

e) The resulting vectors of all utterances observed so far are collected in one (big) data matrix X. An increasingly popular technique, Non-negative Matrix Factorization ([2, 9, 12]) is applied to factorise the matrix X into two smaller matrices W and H. In combination with the tags in the input, the columns in W code speech fragments in the input that relate to specific tags. If the fragments correspond to 'words', this is equivalent to building internal representations of words.

f) Test: For an unknown utterance, the corresponding vector in S is decomposed in terms of the columns of W, thereby providing the decoding of the utterance in terms of the set of already learned speech fragments.

Segmentation is not used at the input side, nor in the process that results in the emerging word representations.

3. Data and results

This word detection algorithm was applied to three databases (Dutch, Finnish and Swedish). For each language the database contains 2000 utterances from 2 male and 2 female speakers. Each utterance comes with (exactly) one tag. In total, there are 13 different tags per language that are provided by the carer model to the learner model and must be learned by the learner. Figure 1 presents an example of the accuracy of the learner in recognizing tags presented by the carer. In the beginning, the learner does not have any word representation, but after a few hundred utterances the learner is able to bootstrap its internal word representations. The figure displays the accuracy of the learner's performance (y-axis) over time during one particular training session, in which the 8000 utterances in Dutch are presented to the learner in random order. Along the x-axis, the number of tokens (utterances) is displayed.

Figure 2 also shows the accuracy of the learner as a function of the number of observed tokens, but now in the case where utterances are presented in speaker-blocked ordering. The tokens 1-2000 are from the first speaker (NL, female), tokens 2001-4000, 4001-6000 and 6001-8000 are from a male, different female and male speaker, respectively. One observes that the learner is able to learn the words from speaker 1, every new speaker leads to a drop in performance, but the learner is able to 'catch up'. A close analysis of the performance drop reveals that it this drop is mainly due to the speaker-dependency of the learned word representations, while the remainder of the effect is due to the introduction of one single new word per speaker.

Figure 3 shows the performance of the learner in a multilingual acquisition experiment. The four speakers are a Dutch female, Dutch male, Swedish female, and Swedish male, respectively. As in figure 2, each new speaker is associated with a drop in performance. The language change (at x = 4000) shows a more dramatic drop in performance. The two curves relate to two different evaluation measures: the solid line corresponds to the case in which the tags are language independent, while the dashed line corresponds to the case were the learner is confronted with language-dependent tags.

These performance results have been obtained without using any information about segmentation (or even word ordering) in the entire processing. The results presented here and similar results for Swedish and Finnish in various training conditions show that, for a small number of words, word detection is possible without using the concept of segmentation. Even more, the models (i.e. the set of acquired and updated word representations) are powerful enough to allow a form of segmentation of the input signal. This is discussed in more detail in the next section.

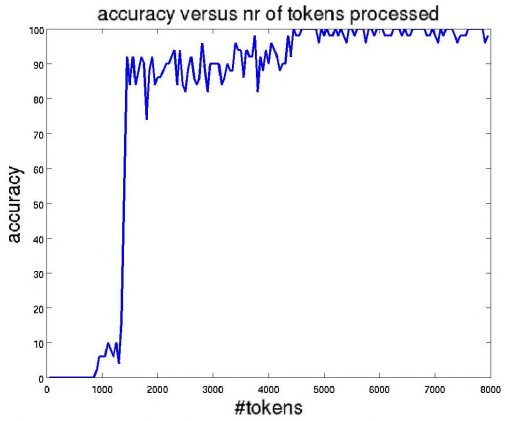

Figure 1. This figure shows the accuracy of the learner as a function of the number of observed tokens (utterances). The 8000 (Dutch) stimuli are presented in random order.

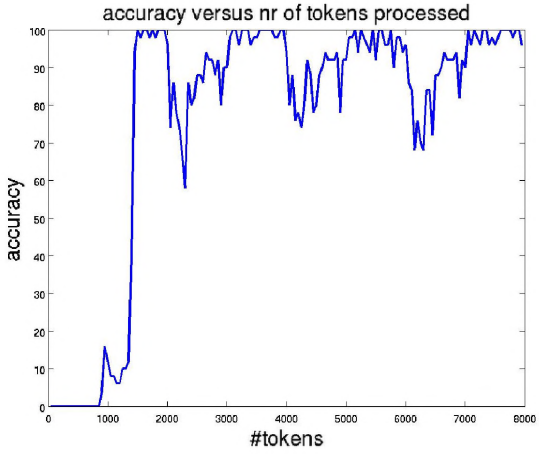

Figure 2. As in figure 1, this figure shows the accuracy of the learner as a function of the number of observed tokens (utterances). The (Dutch) stimuli are now presented in speaker-blocked order. New speakers occur at x=2000, 4000 and 6000.

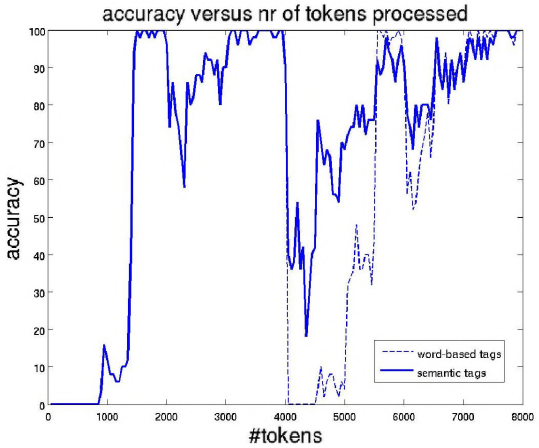

Figure 3. This figure shows the performance of the learner in a multilingual acquisition experiment. For a discussion see the text.

4. Segmentation

The basic property of the current learning algorithm is to detect the presence of words (and word-like fragments) without looking which stretch of the signal matches best with one of its internal word representations. Instead, the learner analyses the entire utterance as a *whole* and tries to decompose, based on the additive property of the mapping M explained in section 2, the *entire* utterance using all its internal representations. This way of analysis resembles the way in which an emission spectrum of an unknown chemical compound is analysed to unravel the compound into its basic constituents. In the same way as the emission spectrum of the compound is an overlay of more elementary spectra, the word detection algorithm analyses the utterance by assuming it is a temporal composition of words that are associated to its internal representations. It follows that segmentation is not used in the input and in the learning process, nor is it directly available in the decoding.

Although it is not directly part of the output, segmentation can still be obtained. The learned word representations allow a segmentation of the input. This is possible due to the additive property of M. Each time a certain word 'A' appears in the input, the corresponding vector in S has an additional term M(A) in its decomposition. That means that the location of a word can be found by presenting gated versions of the input to the decoding algorithm, by monitoring the coefficient of M(A) in the resulting decomposition, and by investigating the evolution over time of this particular coefficient. Figure 4 presents the result for the Dutch word 'luier' ('diaper') that appears in one of the Dutch utterances. Along the x-axis, the last frame number of the gated input is displayed. A value of 100 means that exactly one second of the utterance (measured from the beginning of the waveform) is presented to the decoding algorithm.

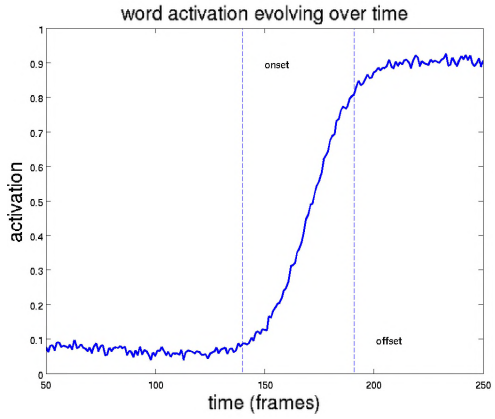

Figure 4. The location of the word 'luier' visualised by plotting the activation of the 'luier' representation in the decoding of gated utterances. Along the x-axis, the last frame number is given of the gated utterance. For example, x=160 corresponds to a gated signal of length 160 frames (1.60 seconds). The actual start and end of the word 'luier' in the input are represented by the vertical bars.

The y-axis shows the weight ('activation') of the vector corresponding to 'luier' in the decoding of the gated input. Until x=140, the word 'luier' does not appear in the signal and the corresponding decoding result is in accordance with this. Between x=140 and x=200, more and more information from the acoustic token 'luier' is available in the input, and the decoding shows an increasing activation for the vector associated with the internal representation 'luier' between these instants.

5. Discussion and outlook

The experiments with the word detection model show that the learner is able to do the following:

- The learner learns to relate acoustic word forms and references in the form of tags. The learner needs a minimal number of acoustic tokens before it can make a reliable word representation.
- The learner rapidly adjusts to a new speaker. When the role of the carer is fulfilled by one speaker, the learner's internal representations will be speaker-dependent. As soon as a new speaker starts interacting with the learner, the existing internal representations will be adapted to accommodate the characteristics of the new speaker.
- In order to detect words, the learner does not rely on segmentation information. The internal representations, however, are powerful enough to support segmentation.

It is clear from figure 4 that it is not easy to define the exact start and end of the target word (here the word 'luier' '*diaper*'). The figure shows that the algorithm profits from information available in the context of the word 'luier'. We have applied the segmentation algorithm on more tokens of several words, and a preliminary analysis of the results shows that the segmentation algorithm as described here is able to locate onset and offset of words with an temporal accuracy of σ = 10 frames (approximately). Broadly speaking, this corresponds to 1 a 1.5 times the average duration of a phone in read-aloud speech.

It is difficult (and close to unfair) to compare the learner's performance to the performance of human listeners. E.g. human transcribers perform better when segmenting at the phone or word level, but they have the advantage of using context information and top-down knowledge in their segmentation decision. Moreover, most of them are adults – exposed to much more speech material than the learner algorithm in these experiments.

The main question currently under investigation is to what extent and how the cognitive plausibility of the model can be improved.

6. Acknowledgement

This research was funded in part by the European Commission, under contract number FP6-034362, in the ACORNS project (www.acorns-project.org).

7. References

[1] Boves, L., ten Bosch, L., Moore, R. K. (2007). ACORNS – towards computational modeling of communication and recognition skills, Proc. ICCI-2007.

[2] Hoyer, P.O. (2004) Non-negative matrix factorization with sparseness constraints, Journal of Machine Learning Research, 5, 1457–1469.

[3] Johnson, S. (2002) Emergence. New York: Scribner.

[4] Kuhl, P.K. (2004) Early language acquisition: cracking the speech code. Nat. Rev. Neuroscience, 5: 831-843.

[5] Maloof, M.A., Michalski, R.S. (2004). Incremental learning with partial instance memory. Artificial intelligence 154, 95-126.

[6] Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, Vol. 52, 1994, pp. 189-234.

[7] Ostendorf, M (1999) Moving beyond the 'beads-on-a-string' model of speech, in Proc. IEEE ASRU-99, Keystone, Colorado, USA. Dec 12-15.

[8] Roy, D.K. & Pentland, A.P. (2002) Learning words from sights and sounds: a computational model. Cognitive Science, 26: 113-146.

[9] Stouten, V., Demuynck, K., Van hamme, H. (2007). Automatically Learning the Units of Speech by Non-negative Matrix Factorisation. Interspeech 2007, Antwerp, Belgium.

[10] Werker, J.F. and Curtis, S. (2005) PRIMIR: a developmental framework for of infant speech processing. Language Learning and Development, 1: 197-234.

[11] Werker, J.F. and Yeung, H.H. (2005) Infant speech perception bootstraps word learning. TRENDS in Cognitive Science, 9: 519-527.

[12] Stouten, V., Demuynck, K, Van hamme, H. (2007). Discovering phone patterns in spoken utterances by non-negative matrix factorisation. Signal Processing Letters IEEE vol 15, p.131-134.

[13] Ten Bosch, L., and Cranen, B. (2007). A computational model for unsupervised word discovery. Proceedings Interspeech 2007, Antwerp, Belgium, p. 1481-1484.

[14] Graf Estes, K., Evans, J.L., Alibali, M.W., and Saffran, J.R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. Psychological Science. 18(3): 254--260

[15] Jusczyk, P.W. (1999) How infants begin to extract words from speech. TRENDS in Cognitive Science, 3: 323--328.

[16] Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show facilitation for native language phonetic perception between 6 and 12 months. Developmental Science, 9, 13--21.

[17] James M. McQueen, Anne Cutler, Dennis Norris (2006). Phonological Abstraction in the Mental Lexicon. Cognitive Science 30 (2006), 1113--1126.

[18] Saffran J.R., Newport E.L., Aslin R.N. (1996). Word segmentation: the role of distributional cues. J Mem Lang 35:606–621.

[19] Davis, M.H., Marslen-Wilson, W.D., Gaskell, M.G., 2002. Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance, 28, 218-244.

[20] Salverda, A.P., Dahan, D., McQueen, J.M., 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition, 90, 51-89.