The Role of Articulatory Feature Representation Quality in a Computational Model of Human Spoken-Word Recognition

Odette Scharenborg[1,2] *and Danny Merkx*[2]

[1] Multimedia Computing Group, Delft University of Technology, Delft, the Netherlands

[2] Centre for Language Studies, Radboud University, Nijmegen, the Netherlands

o.e.scharenborg@tudelft.nl, d.merkx@let.ru.nl

Abstract

Fine-Tracker is a speech-based model of human speech recognition. While previous work has shown that Fine-Tracker is successful at modelling aspects of human spoken-word recognition, its speech recognition performance is not comparable to that of human performance, possibly due to suboptimal intermediate articulatory feature (AF) representations. This study investigates the effect of improved AF representations, obtained using a state-of-the-art deep convolutional network, on Fine-Tracker's simulation and recognition performance: Although the improved AF quality resulted in improved speech recognition; it, surprisingly, did not lead to an improvement in Fine-Tracker's simulation power.

Index Terms: convolutional networks, spoken-word recognition, computational modelling, articulatory features

1. Introduction

There is ample evidence that listeners use subtle acoustic information in the speech signal to help resolve temporary ambiguity due to words being embedded in other words (e.g., [1, 2, 3, 4]). For instance, [2] showed that listeners are able to disambiguate between words like 'ham' and 'hamster' before the offset of the first syllable, an effect that the authors attribute to the fact that monosyllabic words are on average slightly longer than these same syllables embedded in a longer word.

Fine-Tracker is a speech-based computational model of human spoken-word recognition which was specifically designed to investigate the role of durational information in spoken-word recognition [5]. Fine-Tracker is based on the abstractionist theory underlying [6], which assumes that speech recognition is a two-staged process. At the prelexical, first level, the acoustic signal is mapped to a set of limited 'abstract' representations. These prelexical units are then mapped to words at the lexical, second level. In Fine-Tracker, the prelexical representations are extracted from the speech signal using neural network classifiers, and consist of articulatory features (AFs), which are acoustic correlates of articulatory properties of the speech signal. The prelexical representations are then passed on to the lexical level for word recognition.

Fine-Tracker has been successful in simulating the results of human experiments on the use of durational information in spoken-word recognition [1, 2] and as such provided important evidence for the theoretical assumptions on the role of durational information in human speech processing [5]. While Fine-Tracker's simulation power is strong, it is a fairly poor speech recognition system. The word recognition process, as it takes the AF representations as input, is dependent on the quality of the extracted AFs. Potentially, Fine-Tracker's word recognition and simulation performance could benefit from improved AF classification. The goal of this study is to investigate the effects of improved AF quality on Fine-Tracker's word recognition and simulation power. The original implementation of Fine-Tracker used multi-layer perceptrons (MLPs) with a single hidden layer to map the speech signal to AFs. Deep convolutional neural networks (CNNs) have been applied to automatic speech recognition (ASR) with much success (e.g. [7, 8, 9]). Here, we investigate the use of deep CNNs for the extraction of the AFs on Fine-Tracker's recognition performance and simulation power, and compare the results to the human data [2] and the modelling results [5].

2. CNN and Fine-Tracker models

Fine-Tracker's prelexical level creates a multi-dimensional AF vector for every 5 ms of speech of the speech signal. Each AF vector has a continuous value between 0 and 1 (which is the posterior probability output by the networks) for each of seven AF types: *manner* and *place of articulation*, *voice*, *backness*, *height*, *lip rounding*, *vowel duration/diphthong* (identical to [5]), resulting in AF vectors of size 33 for each input frame. A separate network is trained for each AF.

In [10], we created and compared three different DNNs on the task of AF classification. The best-performing DNN, a CNN, is used in the current study as Fine-Tracker's prelexical level. The CNNs consisted of an input layer, 5 blocks each consisting of 2 convolutional layers followed by a max pooling layer ending with 4 fully connected layers and a softmax output layer. The CNN architecture was trained using Mel Filterbank features consisting of 64 filters, which were computed using 25 ms analysis windows with 5 ms shift. The CNNs were trained using the read speech material from the Corpus Spoken Dutch (CGN, Corpus Gesproken Nederlands) [11]. The material consisted of 64 hours of read speech by 324 unique speakers. The training data was split into a training (80% of the full data set), validation (10%) and test set (10%) with no overlap in speakers. AF labels were derived by first forced aligning the speech data with the phonemic transcriptions using a GMM-HMM system implemented in Kaldi [12]. Next, for each frame, the phonemic CGN label was replaced with the canonical AF types using a look-up table. These newly created AF CNN-based classifiers showed relative improvements of up to 18.61% for each AF compared to the original Fine-Tracker MLPs, with AF accuracies of *manner*: 86.9%, *place*: 86.3%, *voicing*: 93.5%, *backness*: 89.2%, *height*: 89.2%, *rounding*: 90.6%, *duration/diphthong*: 88.2%. This improvement could not be explained from an increase in training material (see [10] for more details). These AFs were passed on to Fine-Tracker's lexical level.

Fine-Tracker's lexical layer is kept as in the original model [5]. The lexicon is represented as a tree where each node is a canonical AF vector and branches are words. The search consists of a probabilistic, breadth-first word search which maps the prelexical AF vectors onto the canonical AF vectors of the words. The output consists of a ranked N-best list (N=50) of predictions for every input frame. This allows for the evaluation of the word activations over time.

3. Experimental set-up

The goal of the modelling study is to investigate the role of durational information in human speech processing which is done by comparing two models: one with and one without the ability to use durational information. To investigate the improved AFs effect on the modelling and recognition power of Fine-Tracker, we similarly compared the original MLP-based and the new-CNN-based versions of Fine-Tracker with and without the ability to use durational information. We used the (read-speech) stimulus materials, like was done in [5], from the eye-tracking study [2]: 28 multi-syllabic target words of which the first syllable was also an embedded monosyllabic word, such as 'ham' in 'hamster'. There were two conditions for every word: 1) MONO condition: the first syllable of the target word 'hamster' was replace by a recording of the monosyllabic word 'ham'; 2) CARRIER condition: the first syllable of 'hamster' came from another recording of 'hamster'. During the original experiment, participants' eye movements were monitored while they were listening to the target words embedded in sentences. Analysis of the eye movements showed that there were significantly more transitory fixations to pictures representing monosyllabic words (e.g., 'ham') in the MONO condition than in the CARRIER condition.

Following [5], durational information is hard-coded in Fine-Tracker's lexicon. To investigate the effect of durational information, there were two lexicons. In the canonical lexicon (= no durational information), the lexical AF representations for the monosyllabic words and the first syllable of the target words were identical. To accommodate the use of durational information, in the duration lexicon, each phoneme of the monosyllabic words was (arbitrarily) represented by two identical AF vectors, while each phoneme of the first syllable of a multi-syllabic word was represented using a single AF vector. The lexicon used in this study contained only the target words and the embedded words, i.e., a lexicon size of 56.

Fine-Tracker's simulation performance is evaluated by comparing the word activations of the embedded words over time in the MONO and CARRIER conditions. Following [5], a correct simulation is when, before the end of the first syllable of the target word (e.g., 'hamster'), the word activation for the embedded word (e.g., 'ham') is higher in the MONO condition than in the CARRIER condition. If durational information is indeed important for the disambiguation of the embedded and target words, the duration lexicon condition should result in more correct simulations than the canonical lexicon condition.

Table 1. The number of times (max = 28) the target word was found in the final 50-best list per lexicon type. Between brackets:#target word was top prediction.

Condition	**MLPs-[5]**		**CNNs**	
	can	dur	can	dur
MONO	27 (20)	25 (15)	28 (23)	23 (19)
CARRIER	23 (21)	22 (16)	28 (23)	21 (18)

4. Results

Word recognition performance was higher for the CNN architecture compared to the MLP architecture from [5]: 23 times (in both lexicon conditions) the target word was the top prediction (see Table 1 for the full results). However, for the duration lexicon, the number of target words appearing in the 50-best list is slightly lower than for MLP-[5], although the CNN-based AF vectors do outperform MLP-[5] in terms of words that were correctly recognised (= the top prediction).

Comparing the modelling ability of Fine-Tracker with the MLP-based and CNN-based architectures showed that both architectures correctly simulated the effect of the role of durational information in embedded word disambiguation: the duration lexicon outperformed the canonical lexicon. The number of times the embedded word (e.g., 'ham') had a higher activation over time in the MONO condition than in the CARRIER condition, i.e., where the 'winning' condition had the highest activation over the largest part of the stimulus, was 9 for both architectures for the lexicon without durational information. When the duration lexicon was used, these numbers increased to 17 for the MLPs and 15 for the CNNs. A one-tailed McNemar for paired samples (without continuity correction) showed that the difference between the canonical and duration lexicons was significant for both architectures (see [5] for the MLP results; CNN: $\chi^2 = 3.6$, $p = .029$).

In the human experiment, the correct image attracted on average more eye fixation only for 18 out of the 28 stimuli. Comparisons of the model performance and human behaviour on a stimulus-by-stimulus level shows that the Fine-Tracker results agree with the human data on 10 out of 18 cases for the CNNs and 8 out of 18 times for the results reported in [5].

5. Concluding remarks

As expected, the higher quality of the AF vectors from the CNN-based system resulted in improved word recognition performance. For the canonical lexicon, all of the target words appeared in Fine-Tracker's final predictions with 23 (out of 28) words appearing as the top prediction which is the best recognition performance reported so far, while also the duration lexicon condition outperformed the original model's recognition performance.

The results suggest that Fine-Tracker's modelling performance is to some degree dependent on the AF vectors' quality. However, even though the CNNs showed large AF classification improvements and the best word recognition results, these improved AFs did not increase Fine-Tracker's ability to model human behavioural data compared to the results in [5], although on a stimulus-by-stimulus basis they did outperform the MLP-version of the model. It is not clear why these improvements did not result in better modelling of the use of durational information. In order to make a fair comparison to previous Fine-Tracker results the simulation settings were the same as those in [5]. Perhaps these parameters need to be tuned to the new AF vectors in order to allow Fine-Tracker to make better use of the increased quality of the AF vectors.

6. Acknowledgements

O.S. was supported by a Vidi-grant from The Netherlands Organization for Scientific Research (NWO; grant number: 276-89-003). This work was carried out by D.M. as part of a project under the supervision of O.S.

7. References

[1] K. B. Shatzman and J. M. McQueen, “Segment duration as a cue to word boundaries in spoken-word recognition,” *Perception & Psychophysic*s, vol. 68, no. 1, pp. 1–16, 2006.

[2] A. P. Salverda, D. Dahan, and J. M. McQueen, “The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension,” *Cognition*, vol. 90, no. 1, pp. 51–89, 2003.

[3] M. H. Davis,W. D. Marslen-Wilson, and M. G. Gaskell, “Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition,” *Journal of Experimental Psychology*, vol. 28, no. 1, pp. 218–244, 2002.

[4] R. Kemps, M. Ernestus, R. Schreuder, and R. H. Baayen, “Prosodic cues for morphological complexity: The case of Dutch plural nouns,” *Memory & Cognition*, vol. 33, no. 1, pp. 430–446, 2005.

[5] O. Scharenborg, “Modeling the use of durational information in human spoken-word recognition,” *Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3758–3770, 2010.

[6] D. Norris, “Shortlist: A connectionist model of continuous speech recognition,” *Cognition*, vol. 52, no. 1, pp. 189–234, 1994.

[7] S. M. Siniscalchi, D. Yu, L. Deng, and H. Lee, “Exploiting deep neural networks for detection-based speech recognition,” *Neurocomputing*, vol. 106, pp. 148–157, 2012.

[8] Y. Qian and P. Woodland, “Very deep convolutional neural networks for robust speech recognition,” *arXiv:1610.00277*, 2016.

[9] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, “Modelling fine-phonetic detail in a computational model of word recognition,” *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Japan, pp. 4277–4280, 2012.

[10] D. Merkx and O. Scharenborg, “Articulatory feature classification using convolutional neural networks,” *Proceedings of Interspeech*, India, 2018.

[11] N. H. J. Oostdijk, W. Goedertier, F. van Eynde, L. Boves, J. P. Martens, M. Moortgat, and H. Baayen, “Experiences from the spoken Dutch corpus project,” *Proceedings of LREC – Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, pp. 340–347, 2002.

[12] D. Povey, A. Ghoshal, G. Bouilanne, L. Burget, O. Glembek, N. Goel, M. Hanneman, P.Motlek, Y. Qian, P.Schwarz, J. Silovsk, G. Stemmer, and K. Vesel, “The kaldi speech recognition toolkit,” *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, U.S.A., 2011.