AUTOMATIC TEMPORAL SEGMENTATION OF HAND MOVEMENTS FOR HAND POSITIONS RECOGNITION IN FRENCH CUED SPEECH

Li Liu, Gang Feng, and Denis Beautemps

Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes

ABSTRACT

In the context of Cued Speech (CS) recognition, the recognition of lips and hand movements is a key task. As we know, a good temporal segmentation is necessary for the supervised recognition system. However, lips and hand streams cannot share the same temporal segmentation since they are not synchronized. In this work, we propose a hand preceding model to predict temporal segmentations of hand movements automatically by exploring the relationship between hand preceding time and the vowel positions in sentences. To evaluate the performance of the proposed method, we apply the hand preceding model to a multi-speakers database. Hand positions recognition is realized with the multi-Gaussian and Long-Short Term Memory (LSTM). The results show that using the predicted temporal segmentation significantly improves the recognition performance compared with that using the audio based segmentation. To the best of our knowledge, this is the first automatic method to predict the temporal segmentation for hand movements only from the audio based segmentation in CS.

Index Terms— Cued Speech, hand preceding model, temporal segmentations of hand position movements, Hand positions recognition, LSTM.

1. INTRODUCTION

To overcome the problems of lip-reading [1] and to improve the reading ability of deaf children, in 1967, Cornett [2] developed the Cued Speech (CS) system to complement the lip information and make all phonemes of a spoken language clearly visible. As many sounds look identical on lips (e.g., /y/, /u/ and /o/), those sounds can be distinguished using hand information and thus make it possible for deaf people to understand a spoken language using visual information alone.

For French CS which is named *Langue française Parlée Complétée (LPC)* [3], hand positions are used to encode five group of vowels (see Fig. 1), and place near the face. Combining eight hand shapes which code consonants groups, the sounds (phonemes) of traditional spoken languages become visible. Note that other widely used systems such as gestural sign language [4–6] are not efficient to improve reading performance. Cued Speech is a visual representation of a spoken language, and it was developed to help raising the literacy level of deaf individuals.

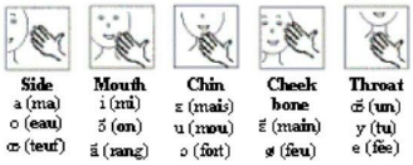

Fig. 1. Manual cues of hand position in *LPC*. The figure is derived from [7].

Hand and lips movements in CS are coherent and complementary to realize an efficient communication. Hand and lips have their own movement rules (not well synchronized, i.e., the hand is more related to the speech syllabic cycle and the lips to phoneme production). The asynchrony problem of lips and hand movement in CS is a challenging issue for the recognition task. The temporal organization of hand movements has been investigated in the literature [7, 8]. Attina et al. found that the hand reaches its target roughly 200ms before the vowel being visible at lips. Indeed, hand movements are very complicated in CS coding process. Although there are some automatic methods to obtain the audio based segmentation [9–11] which can be used for lips, no previous work explored an automatic method to obtain a proper temporal segmentation of hand movements in CS based on the corpus without using artificial marks. In the prior work [12–14], hand positions were extracted by tracking blue colors on the subject's hand. Therefore, the temporal segmentation of hand movement can be obtained based on Gaussian modeling of the hand positions and a minimum of the velocity. However, this method needs both position on the back of the hand and the target finger position. To apply this method to the database without artificial marks is not directly possible.

Long-Short Term Memory (LSTM) [15, 16] have been successful in many fields including audio speech recognition [17] and visual speech recognition [17, 18]. It can exploit a self-learnt amount of long-range temporal information. This ability is helpful to improve noise robustness, e.g. in the case where hand does not reach to its right position.

In this work, as shown in Fig. 2, the main contribution is that we propose a hand preceding model to predict the tempo-

ral segmentation of hand movements for CS by investigating the relationship between hand preceding time and the vowel position in sentences in a subset of the database. Note that the hand movements mean the change of hand positions (no hand shape movements in this work). To evaluate the proposed method, we first use the simple multi-Gaussian classifier (without exploring any temporal information) for hand positions recognition based on the sub-database which is used to build the hand preceding model. Then, applying the hand preceding model to the whole database, we use LSTM to continuous hand positions recognition. The predicted temporal segmentation is compared with the audio based segmentation. The hand positions recognition result confirms the promising performance of the proposed method. In the literature of CS, there was no other published work related to hand positions recognition based on the corpus without using any artificial marks.

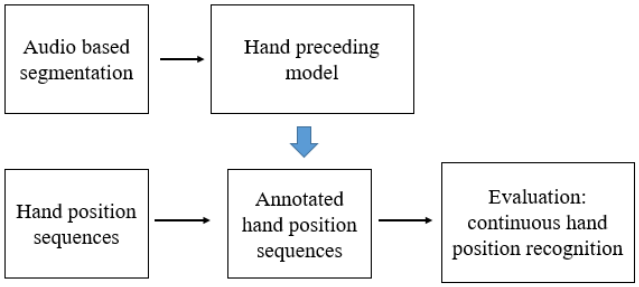

Fig. 2. An overview of this work. The application of the proposed hand preceding model to continuous hand position recognition in CS.

2. DATA ACQUISITION

2.1. Cued Speech material

The database is collected from two female normal-hearing CS speakers. The database of the first subject was recorded without using any artificial marks in 2016, and the database of the second subject was a previous corpus with artificial marks [14] (but the marks are not used in this work). The video images of the speaker's upper body (720x576 RGB images, 50 fps) are recorded in a sound-proof booth in Gipsa-lab, France. Speakers are certified in transliteration speech into Cued Speech in the French language. The first subject pronounces and codes a set of 238 French sentences in CS derived from a corpus described in [12, 19]. Each sentence is repeated twice by the speaker resulting in a set of 476 sentences. The second corpus is made of 44 short sentences which come from a large database [14]. We take a subset of the whole database to build the hand preceding model, which contains 182 sentences including 88 short sentences and 50 long sentences for the first subject (totally 1068 vowels) and 44 short sentences (196 vowels) for the second subject.

The phonetic transcription of each recorded sentence is extracted automatically using Liaphon [20]. The audio based temporal segmentation of each vowel is extracted from a conventional ASR system in HTK 3.4 [11]. Using forced alignment, the acoustic signal synchronized with the video was automatically labeled. Both of them are manually post-checked.

2.2. Hand position tracking

To evaluate the proposed model, the hand position feature is needed to realized the hand position recognition on the whole database. The automatic hand position tracking is a highly difficult task since the hand shape keeps changing and rotating when the hand moves. In the prior CS study, hand position is captured by tracking the artificial marks on the subject's hand. This work is the first time to propose an automatic hand position tracking method in CS without using any artificial marks. It is based on the Mixture GMM foreground extraction approach [21–23]. The mixture Gaussian model uses five Gaussian models to characterize the individual pixel in the image. For a new image, the mixture Gaussian model will be updated, and each pixel in the current image is matched with the mixture Gaussian model. If it is matched, this point will be regarded as the background point. Otherwise, it will be classified as the foreground point. The center gravity of the extracted foreground is taken as the hand position.

However, the above automatic hand tracking method may have some errors. In the following modeling section, we track the hand position manually on the subset of the database to ensure the accuracy of the model.

3. HAND PRECEDING MODEL

The objective of this experiment is to study the relationship between the hand preceding time and the vowel position in sentences. A hand preceding model is built to predict the temporal segmentation of hand movements.

3.1. Hand preceding time

In this paper, we define the hand preceding time as the time difference between the hand target instant and the acoustic target instant. Denote the middle instant of hand target by H_t and the middle instant of the audio based segmentation by A_t (Fig. 3). The hand preceding time Δ_t (in ms) is expressed as $\Delta_t = A_t - H_t$ for each vowel.

In order to calculate the hand preceding time, an accurate hand target instant (H_t) is needed. Since no artificial marks can be used in our database, we detect H_t manually on the sub-database in order to study the hand preceding time accurately. To make the manual determination easier, we use the integrated speed rate of hand movements to help us to localize the hand target instant. The integrated speed rate is calculated as $v = \sqrt{v_x^2 + v_y^2}$, where v_x and v_y is the speed rates in x

and y direction. The minimum value of the curve means that hand is reaching its target instant. With the help of the above integrated speed rate, a semi-manual temporal segmentation of vowels for each sentence is accomplished by using the movie editor Magix [24].

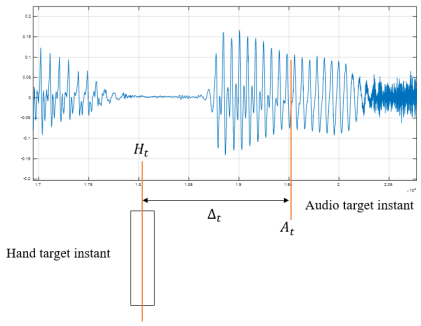

Fig. 3. The illustration of hand preceding time Δ_t, hand position instant H_t and audio signal instant A_t.

3.2. Hand preceding model

The hand preceding time as function of vowel position of 138 sentences for the first subject is plotted in Fig. 4(a). It shows the relationship between hand preceding time (Δ_t) and the instant of the vowel in each sentence. We align all sentences from their end, which is considered as the instant 0. We can see that Δ_t remains stable from the beginning of the sentences and then decreases until the ending of the sentence. The hand preceding model which is a uniform distribution from the beginning to the instant of a turning-point, and follows a linear relationship from this turning-point to the end of the sentence. The uniform distribution (a horizontal line) takes the mean value (0.139) of the data before the instant of the turning-point, and after this instant, a linear regression line (with slope -0.213, and correlation coefficient -0.68) is built. This turning-point is the intersection of two lines. We can see that this model fits not only for the short sentences but also for the long sentences.

One the other hand, as shown in Fig. 4(b), we plot the hand preceding time as function of vowel position of the 88 and 44 short sentences of two subjects, respectively. The same model as Fig. 4(a) is plotted by the black curve. In fact, the linear regression coefficient (with slope -0.228) for the second subject is similar to that for the first subject. Fig. 5 illustrates the application of the hand preceding model for predicting the hand movement temporal segmentation from audio based segmentation.

4. EVALUATION OF THE HAND PRECEDING MODEL

To evaluate the performance of the hand preceding model, hand positions recognition of CS is first carried out based on the sub-database and then on the whole database. The main difference between these two databases is that we have the ground truth hand position and hand segmentation for the sub-database, but not for the whole database.

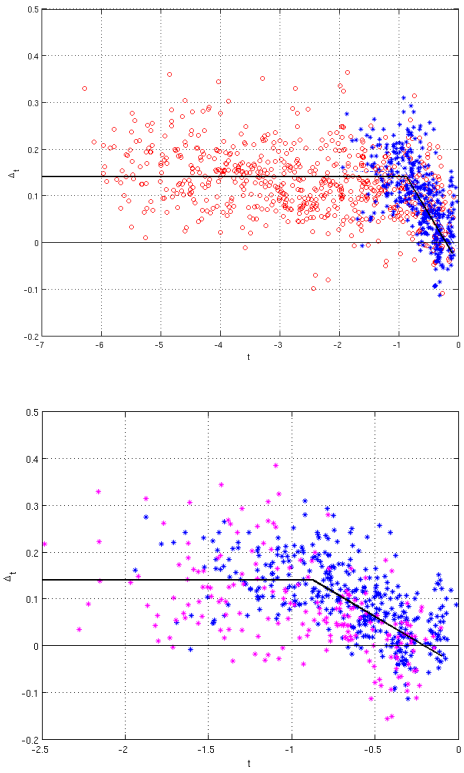

Fig. 4. The top one is (a), and the bottom is (b). The X-axis is the vowel instant in a sentence. Y-axis: the preceding time Δ_t. In (a), the red circles shows the distribution of the long sentences, and the blue start the short sentences. The black curve shows the hand preceding model. In (b), the blue stars show the distribution of short sentences for the first subject and the magenta stars the second subject.

In this work, 80% and 20% of the database (randomly chosen) are used for training and testing (no overlap between them), respectively. Ten repetitions with different training and testing sets are used to refine the recognition results. The standard deviations are small ($\pm$ 0.3) for all results, which do not take into account the CS coding errors. Note that the duration of the ground truth segments is the stable interval around the target instant, and duration of the predicted segments is the same as that of the audio based segmentation.

4.1. Hand positions recognition in CS based on the sub-database

A simple multi-Gaussian model is first used as a recognizer on this sub-database since the aim is just to evaluate the perfor-

Table 1. Five hand positions recognition using multi-Gaussian classifier based on audio based segmentation, the predicted segmentation and the ground truth segmentation.

	auto hand pos	manual hand pos
audio based segmentation	45.41%	59.63%
the predicted segmentation	54.40%	71.33%
the ground truth segmentation	62.26%	86.57%

mance of the predicted segmentation.

In Table 1, we compare the hand positions recognition using the predicted segmentation with that using audio based segmentation and the ground truth segmentation, respectively. Based on the ground truth hand position, upper limit 86.57% using the ground truth segmentation is obtained. A significant improvement is achieved using the predicted segmentation (71.33%) compared with that using the audio based segmentation (59.63%). It shows the efficiency of hand temporal segmentation using the hand preceding model.

On the other hand, we can see that the recognition score using the automatic tracked hand position is less than using the ground truth hand position. When using the predicted segmentation, 54.40% is obtained using the automatic tracked hand position, which is much lower than 71.33% that using the ground truth hand position. It is due to the error of the automatic hand tracker. In the case of using the the automatic tracked hand position, the best performance 62.26% with the ground truth temporal segmentation can be regarded as a reference for hand positions recognition.

4.2. Hand positions recognition in CS based on LSTM for the whole database

To further evaluate the performance of the proposed hand preceding model, we apply the hand preceding model to the whole database (476 sentences, about 6000 vowels) of the first subject (Fig. 6). LSTM is used for the continuous hand positions recognition based on the automatic tracked hand position.

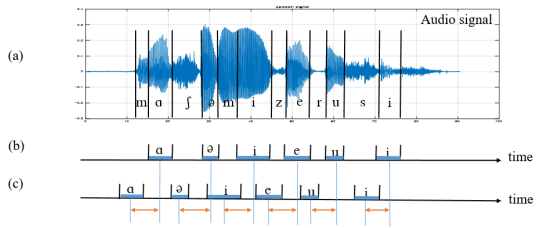

Fig. 5. The predicted Δ_t is shown for the sentence "Ma chemise est roussie". (a) is the audio signal. (b) is the audio based segmentation, while (c) is the segmentation predicted by hand preceding model, the orange interval corresponds to the Δ_t.

In LSTM, two hidden layers of 500 cells, 200 epoch are used. It is trained by backpropagation through time (BPTT) with the cross-entropy cost function. Softmax layer is used to compute the class probability. The final accuracy of LSTM is calculated using max-voting (after softmax layer) which counts the most frequent label as the final label in the corresponding segment. LSTM is implemented using the Keras toolkit [25] based on the GPU-accelerated library.

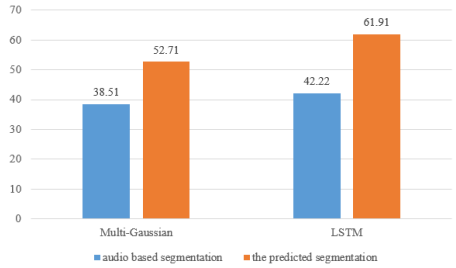

Fig. 6. Hand positions recognition using the multi-Gaussian and LSTM based on the audio based segmentation and the predicted segmentation.

In Fig. 6, the hand positions recognition results confirm the advantages of the the predicted segmentation for both multi-Gaussian and LSTM. Moreover, LSTM obtains higher accuracy than multi-Gaussian since it captures the temporal information of hand movements. More importantly, using LSTM and the predicted temporal segmentation, the accuracy 61.91% almost reached the upper limit 62.26% (mentioned in section 5.1), which uses the ground truth temporal segmentation. The continuous hand positions recognition score can be further improved when more data is applied in LSTM with an accurate hand position.

5. CONCLUSIONS

In this work, we propose a novel hand preceding model to predict the temporal segmentation of hand movements only from the audio based segmentation in CS for the first time. The relationship between hand preceding time and the vowel position in sentences is explored. The evaluation confirms the superior performance of the proposed method. In fact, applying the predicted segmentation to hand positions recognition on the sub-database and the whole database, it significantly outperforms that using the audio based segmentation. Our future work will apply the hand preceding model to hand shape recognition and multi-modal CS recognition.

6. ACKNOWLEDGEMENTS

The authors would like to thank the volunteer speakers for their time spent on Cued Speech data recording, and Duc Canh NGUYEN for his help in LSTM. This conference attending fee is supported by a French CNRS PEPS grant named "LGV".

7. REFERENCES

[1] Gaye H Nicholls and Daniel Ling Mcgill, "Cued speech and the reception of spoken language," *Journal of Speech, Language, and Hearing Research*, vol. 25, no. 2, pp. 262–269, 1982.

[2] Richard Orin Cornett, "Cued speech," *American annals of the deaf*, vol. 112, no. 1, pp. 3–13, 1967.

[3] Carol J LaSasso, Kelly Lamar Crain, and Jacqueline Leybaert, *Cued Speech and Cued Language Development for Deaf and Hard of Hearing Children*, Plural Publishing, 2010.

[4] Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney, "Speech recognition techniques for a sign language recognition system," *hand*, vol. 60, pp. 80, 2007.

[5] Sylvie CW Ong and Surendra Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 6, pp. 873–891, 2005.

[6] Oscar Koller, Hermann Ney, and Richard Bowden, "Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled," in *Proc. CVPR*, 2016, pp. 3793–3802.

[7] Virginie Attina, Denis Beautemps, Marie-Agnès Cathiard, and Matthias Odisio, "A pilot study of temporal organization in cued speech production of french syllables: rules for a cued speech synthesizer," *Speech Communication*, vol. 44, no. 1, pp. 197–214, 2004.

[8] Virginie Attina, Marie-Agnès Cathiard, and Denis Beautemps, "Temporal measures of hand and speech coordination during french cued speech production," in *International Gesture Workshop*. Springer, 2005, pp. 13–24.

[9] David Rybach, Christian Gollan, Ralf Schluter, and Hermann Ney, "Audio segmentation for speech recognition using segment features," in *Acoustics, Speech and Signal Processing, 2009. IEEE International Conference on*. IEEE, 2009, pp. 4197–4200.

[10] SE Tranter, Kai Yu, G Everinann, and Philip C Woodland, "Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech," in *Proc. IEEE-ICASSP*, 2004, vol. 1, pp. I–753.

[11] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al., "The htk book," *Cambridge university engineering department*, vol. 3, pp. 175, 2002.

[12] Noureddine Aboutabit, *Reconnaissance de la Langue Française Parlée Complété (LPC): décodage phonétique des gestes main-lèvres.*, Ph.D. thesis, Institut National Polytechnique de Grenoble-INPG, 2007.

[13] Noureddine Aboutabit, Denis Beautemps, and Laurent Besacier, "Automatic identification of vowels in the cued speech context," in *Proc. AVSP*, 2007, p. 8.

[14] Noureddine Aboutabit, Denis Beautemps, and Laurent Besacier, "Hand and lip desynchronization analysis in french cued speech: Automatic temporal segmentation of hand flow," in *Proc. IEEE-ICASSP*, 2006, vol. 1, pp. I–I.

[15] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins, "Learning to forget: Continual prediction with lstm," *Neural Computation*, vol. 12, pp. 2451–2471, 1999.

[17] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE-ICASSP*, 2013, pp. 6645–6649.

[18] Stavros Petridis, Zuwei Li, and Maja Pantic, "End-to-end visual speech recognition with lstms," in *Proc. IEEE-ICASSP*, 2017, pp. 2592–2596.

[19] Guillaume Gibert, Gérard Bailly, Denis Beautemps, Frédéric Elisei, and Rémi Brun, "Analysis and synthesis of the three-dimensional movements of the head, face, and hand of a speaker using cued speech," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1144–1153, 2005.

[20] Frédéric Béchet, "Lia phon: un systeme complet de phonétisation de textes," *Traitement automatique des langues*, vol. 42, no. 1, pp. 47–67, 2001.

[21] Chris Stauffer and W Eric L Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE-CVPR*, 1999, vol. 2, pp. 246–252.

[22] Kyungnam Kim, Thanarat H Chalidabhongse, David Harwood, and Larry Davis, "Real-time foreground–background segmentation using codebook model," *Real-time imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[23] Derek R Magee, "Tracking multiple vehicles using foreground, background and motion models," *Image and vision Computing*, vol. 22, no. 2, pp. 143–155, 2004.

[24] Daniel V Abreu, Thomas K Tamura, Donald G Keamy Jr, Roland D Eavey, et al., "Podcasting: contemporary patient education," *Ear, Nose & Throat Journal*, vol. 87, no. 4, pp. 208, 2008.

[25] François Chollet et al., "Keras," 2015.