A novel resynchronization procedure for hand-lips fusion applied to continuous French Cued Speech recognition

Li Liu[1], Gang Feng[2], Denis Beautemps[2], Xiao-Ping Zhang[1]
1 Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada.
2 Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France.

***Abstract*—Cued Speech (CS) is an augmented lip reading with the help of hand coding. Due to lips and hand movements are asynchronous and a direct fusion of these asynchronous features may reduce the efficiency of the recognition, the fusion of them in automatic CS recognition is a challenging problem. In our previous work, we built a hand preceding model for hand positions (vowels) by investigating the temporal organization of hand movements in French CS. In this work, we investigate a suitable value of the hand preceding time for consonants by analyzing the temporal movements of hand shapes in French CS. Then, based on these two results, we propose an efficient resynchronization procedure for the fusion of multi-stream features in CS. This procedure is applied to the continuous CS phoneme recognition based on the multi-stream CNN-HMMs architecture. The result shows that using this procedure brings an improvement of about 4.6% in the phoneme recognition correctness, compared with the state-of-the-art, which does not take into account the asynchrony of multi-modalities.**

***Index Terms*—Cued Speech, multi-modal fusion, hand preceding time, resynchronization procedure, CNN-HMMs**

I. Introduction

To overcome the problems of lip reading [1] and improve the reading ability of deaf children, in 1967, Cornett [2] invented the Cued Speech (CS) system, which complements the lip reading and makes all the phonemes of a spoken language clearly visible. In the French CS named *Langue franaise Parle Complte (LPC)* [3], five hand positions are used to encode the vowel groups, and eight hand shapes are used to encode the consonant groups [4]. In this system, these sounds, which may look similar on lips (e.g., /y/, /u/ and /o/), can be distinguished using the hand information (three different hand positions for /y/, /u/ and /o/), and thus it is possible for the deaf people to understand a spoken language using visual information alone.

The automatic continuous CS recognition is a multi-modal task, as it includes the lips, hand position and hand shape information. To realize this task, one challenging problem is the fusion of these multi-stream features given the fact that lips and hand movements are asynchronous. In fact, it was investigated that the hand reaches its target on average $239ms$ [5] (based on non sense syllables logatome, like 'tatuta'), and $144.19ms$ [6] (based on syllables extracted from French sentences) before the vowel being visible at the lips in case of CV syllables, respectively. This hand preceding phenomenon is illustrated in Fig. 1 by an example where the CS speaker utters petit ([p ø t i]). In Fig. 1(a), the speaker points to her cheek position to indicate the vowel [ø], while the corresponding instant (red line) in the acoustic signal is not yet the vowel [ø]. Fig. 1(b) shows the transition of the syllable [p ø], and the speaker is preparing to utter [p]. In Fig. 1(c), the speaker pronounces the vowel [ø], while the hand position has already indicated the next vowel [i].

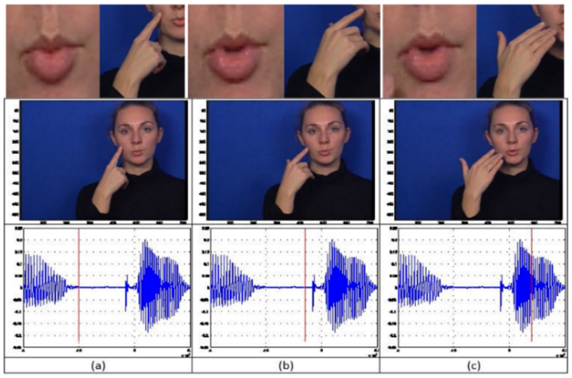

Fig. 1. Illustration of the asynchrony phenomenon in CS production. Top: lips and hand zoomed from the middle image. Bottom: the audio speech signal. Red vertical lines: the instant where the middle image is taken.

In [4], [7], a direct feature fusion was applied to the isolated[1] French CS recognition without taking into account the asynchrony problem. In our recent work [8], the tandem architecture that combines convolutional neural networks (CNN) [9], [10] with multi-stream hidden markov model (MSHMM) [11] was used for the continuous CS recognition. In this architecture, MSHMM merges different features by adding different weights, but it does not take into account the asynchrony between different feature modalities. Therefore, there is still a room for us to improve the CS recognition performance by exploring a reasonable approach to tackle the fusion of asynchronous multi-modalities.

We remark that the deep leaning method *encoder-decoder* [12] with the recurrent neural network (RNN) [9], [10] and attention mechanism could learn the contexts and variabilities

[1]the temporal boundaries of each phoneme to be recognized in the video are known at test stage.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

of the multi-stream features if sufficient data is available. In this work, instead of exploiting the the deep learning based methods which need large data set, we explore a study that is able to give a more clear explanation for us to deeply understand the principle of the CS multi-modal fusion.

In this work, based on the hand preceding time (i.e., the time difference that hand precedes lip movement) for vowels that has been studied in [13], we deal with the optimal hand preceding time for consonants, and then propose a resynchronization procedure to align the hand position and shape features with lips features. One important point is that we use two different hand preceding time for all vowels and consonants, respectively, instead of resynchronizing them by their own corresponding hand preceding time. For the evaluation, we build a new automatic CS recognition architecture S_{re} (see Fig. 2), where the resynchronization procedure is added to process the CNNs based features before the MSHMM-GMM [14] decoder. It is shown that this method significantly improves the CS recognition performance compared with the state-of-the-art of the continuous/isolated CS recognition [8] and [4], respectively. To the best of our knowledge, this is the first work that proposes the resynchronization procedure for multi-modal features fusion in automatic continuous French CS recognition system.

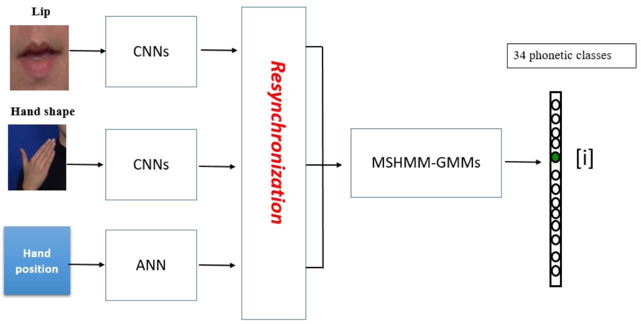

Fig. 2. Proposed architecture S_{re} in this work. The main difference with [8] (architecture $S3$ in [8]) is adding a new resynchronization procedure.

II. Related works

Regarding the modeling of hand preceding time, in our previous work [13], the relationship between the hand preceding time for vowels and their target time instant was analyzed. We found that hand preceding time follows a Gaussian distribution that remains almost the same for all the instants of vowels, except a small time interval just before the end of each sentence. Based on the hand preceding time for vowels that was studied in [13], in the present work, we explore the optimal hand preceding time for consonants and propose a novel resynchronization procedure to align the hand position and shape features with lips features for the automatic continuous CS recognition.

Concerning the feature extraction, in previous works on CS recognition, the video images were recorded with artifices applied to the CS speaker (blue sticks on the lips, blue marks on the hand and forehead) to mark the pertinent information and make their further feature extraction easier. In [15], [16], a modified *constrained local neural fields* (CLNF) was proposed to extract the inner lips height and width, and an adaptive ellipse model was proposed for inner lips parameters extraction in CS [17]. In this work, we adopt the deep CNN for the feature extraction of lips and hand shape, and use the *artificial neural network* (ANN) [9] to process the hand position feature.

As for the automatic continuous CS recognition, a tandem CNN-HMM architecture which extracts the CS feature from raw image was proposed in [8]. However, it did not take into account the asynchrony of the multi-modalities when merging multi-stream features in the automatic CS recognition. In the present work, in order to tackle the multi-modal feature fusion in the automatic CS recognition, we propose a new automatic CS recognition architecture S_{re} (see Fig. 2) by adding a novel resynchronization procedure to process features extracted by CNNs and ANN before feeding them to the MSHMM-GMM decoder. The result shows that this resynchronization procedure significantly improves the CS recognition performance compared with the state-of-the-art of the isolated/continuous CS recognition [4] and [8], respectively.

III. Problem formulation

In the automatic continuous CS phoneme recognition task, the features of lips $O^{(L)}$, hand position $O^{(P)}$ and hand shape $O^{(S)}$ are merged and fed to the phonetic decoder. Let phoneme Υ be extracted from a continuous French sentence with a certain time step t. It is determined by

$$\Upsilon = \arg\max_{\Upsilon} P(O^{(LPS)}|\Theta_{\Upsilon}), \tag{1}$$

where $O^{(LPS)} = [{O^{(L)}}^T, {O^{(P)}}^T, {O^{(S)}}^T]$ is the merged feature and Θ_{Υ} is the model parameter for Υ.

As introduced in Section I, the lips features, hand shapes and positions are asynchronous in CS, which results in the fact that features corresponding to different phoneme classes may be merged to represent one common phoneme. Therefore, at time t, the direct concatenated feature will be interfered and thus not suitable to train one particular phoneme class Υ.

The aim of this work is to propose a way to align the hand position $O^{(P)}$ and shape features $O^{(S)}$ with lips feature $O^{(L)}$, i.e., to build two transformations τ_1 and τ_2 such that:

$$O^{(P)}_{\text{resy}} = \tau_1(O^{(P)}), \tag{2}$$

$$O^{(S)}_{\text{resy}} = \tau_2(O^{(S)}), \tag{3}$$

are synchronized with lips feature $O^{(L)}$, respectively. Then the merged feature of resynchronized features $O^{(LPS)}_{\text{resy}} = [{O^{(L)}}^T, {O^{(P)}_{\text{resy}}}^T, {O^{(S)}_{\text{resy}}}^T]$ for phoneme Υ can be used to train the model of the phoneme without interference.

IV. Methodologies

In this section, we will first introduce the hand preceding time for vowels and consonants. Then, based on these two results, the resynchronization procedure will be proposed.

A. *Hand preceding time for vowels*

In our previous work [13], the relationship between the hand preceding time for vowels (Δ_v) and their target time instant was analyzed. We found that Δ_v follows a Gaussian distribution that remains almost the same for all the instants of vowels, except a small time interval just before the end of each sentence (about one second). In this work, instead of following the piece-wise linear relationship, which gives different Δ_v for each vowel, we assume that the mean value $\overline{\Delta_v}$ (about $140ms$) of the Gaussian distribution is suitable for all vowels. Indeed, we have tried the complex way by using their corresponding Δ_v for each vowel. However, only minor gains were obtained.

B. *Hand preceding time for consonants*

Without loss of generality, we consider the hand preceding time for consonants in the CV (i.e., consonant vowel) syllable context. We carry out a statistical study on the average distance of consonant and vowel based on our database, and it shows that the average distance is about $110ms$. It is observed from our data that the stable time interval for vowels and consonants is about $60ms$ (three images). Therefore, we can deduce that the hand preceding time for hand shape movement Δ_c is about $60ms$ (see Fig. 3).

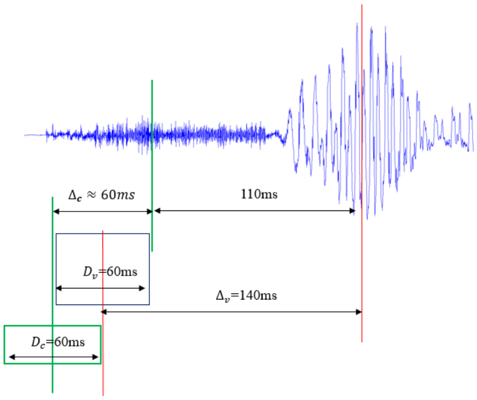

Fig. 3. Relationship between different parameters. Δ_c and Δ_v are the hand preceding time for consonants and vowels, respectively. D_v and D_c are the time duration for target hand position and shape, respectively.

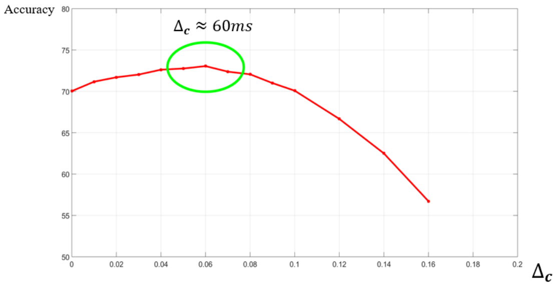

Fig. 4. Eight hand shapes recognition using Gaussian classifier with the feature extracted by CNNs. The hand position stream is shifted by increasing Δ_c values.

To further confirm this value, a Gaussian classifier is applied to recognize eight classes of hand shapes, based on the CNN hand shape features (i.e., the features before the softmax layer of CNN). The temporal segmentation is obtained by shifting a Δ_c value based on the audio-based temporal segmentation. By modifying this value, different recognition scores are obtained in Fig. 4, which confirm that the maximum score is obtained with $\Delta_c = 60ms$.

C. *Resynchronization procedure*

Based on the hand preceding time for vowel and consonant, the proposed resynchronization procedure contains two steps:

1) By applying τ_1 to the hand position feature stream $O^{(P)}$, which positively shift the $O^{(P)}$ by Δ_v^* temporally. More precisely, the pre-aligned hand position feature $O_{\text{resy}}^{(P)}$ is obtained by

$$\tau_1(O^{(P)}(t)) = O^{(P)}(t - \Delta_v^*), \tag{4}$$

where $\Delta_v^* = 140ms$, and t is the time step.

2) By applying τ_2 the hand shape feature stream $O^{(S)}$, which positively shift $O^{(S)}$ by Δ_c^* temporally. More precisely, the pre-aligned hand shape feature $O_{\text{resy}}^{(S)}$ is obtained by

$$\tau_2(O^{(S)}(t)) = O^{(S)}(t - \Delta_c^*), \tag{5}$$

where $\Delta_c^* = 60ms$, and t is the time step.

We take the vowel case (see Fig. 5) as an example to illustrate this procedure. The audio signal is shown in Fig. 5(a) with its phonetic annotation for the French sentence *Ma chemise est roussie*. Note that the lips feature stream is assumed to be synchronous with the audio signal [18]. In Fig. 5(b), the hand position is presented by the x coordinate of the hand back point. We can clearly observe that the hand position stream is not synchronous with the audio signal, and thus a direct fusion of these two streams will not be optimal for the fusion. In Fig. 5(c), the aligned hand position stream is obtained by positively shifting the original one (see Fig. 5(b)) with $\Delta_v = 140ms$ [13]. With this alignment, the hand position stream is resynchronized with the audio signal on average. For consonants, the alignment of the hand shape feature is similar, with $\Delta_c = 60ms$ as introduced in Section IV-B.

In fact, we observe that the hand position feature is more sensitive to the asynchrony problem than the hand shape. This may be due to the intrinsic fact that the hand often stays in its target position for a very short time, while the full realized hand shape keeps longer time in the CS coding.

V. Experiment and Results

In order to evaluate the proposed resynchronization procedure, we carry out the continuous CS phoneme recognition experiments with both $S3$ architectures and S_{re}.

A. *Cued Speech material*

A professional CS interpreter was asked to utter and encode simultaneously a set of 476 French sentences [19] (about 11770 phonemes totally). Color video images of the interpreter's upper body were recorded at 50 fps, with a spatial

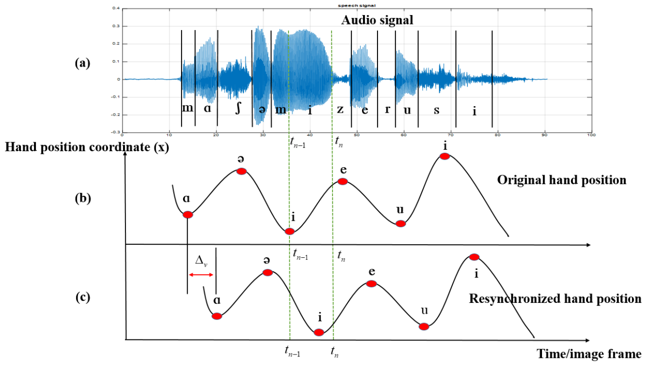

Fig. 5. Proposed resynchronization procedure. (a) The audio speech with its phonetic annotation. (b) The original hand position stream. (c) The aligned hand position stream shifted by Δ_v. Two green lines correspond to the temporal boundaries of vowel [i].

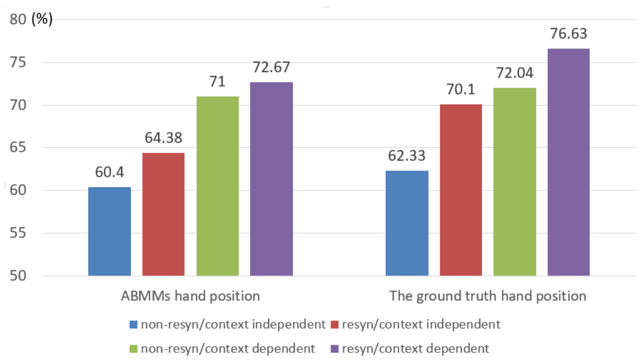

Fig. 6. The result of the continuous CS phoneme recognition with and without using the proposed resynchronization procedure and context-dependent modeling. non-resyn means the case that does not use the proposed resynchronization procedure, and resyn means the case using the proposed resynchronization procedure.

resolution of 720x576. This dataset was made publicly available on Zenodo (https://doi.org/10.5281/zenodo.1206001). The phonetic transcription was extracted automatically using Liaphon [20] and post-checked manually. We remark that the French language is normally described with a set of 34 phonetic classes (14 vowels and 20 consonants). The audio based temporal segmentation for vowels and consonants are obtained based on the force-alignment using HTK [21]. The ground truth hand position in this work is manually determined for all the images of the corpus. We choose this position in the following way: the 2D position of the index finger extremity is assumed if no middle finger appears.

B. CNN-HMMs based CS recognition

The tandem CNN-HMMs structure (see Fig. 2) is used in this work. CNNs are used as the feature extractor and a triphone HMM-GMM is used as the CS phonetic decoder.

As $S3$ in [8], in this work, for the proposed CS recognition architecture S_{re}, each phoneme is modeled by a context-dependent triphone MSHMM (i.e., takes into account the contextual information about the left and right phoneme) [22], and three emitting states are used with GMM to model the features of lips, hand position and hand shape together with their first derivatives. The main difference between $S3$ amd S_{re} is that MSHMM-GMMs are used to model the resynchronized multi-modal features (i.e., $O_{\text{resy}}^{(LPS)}$) in S_{re}, while in $S3$, MSHMM-GMMs are used to model the asynchronous multi-modal features.

In the CNN-HMM architecture, lips and hand shape features are extracted by CNN, and hand position coordinates are processed by ANN. These features with their first derivatives are modeled together in MSHMM-GMM for phonetic decoding. For $S3$, lips and hand information are combined at the state level using the three-stream MSHMM-GMMs. The stream weights are optimized empirically using the cross-validation, resulting in the optimal weights 0.4 for lips, 0.4 for hand shapes and 0.2 for hand positions. It should be noted that neither the pronunciation dictionary nor language model is used in this architecture.

C. Evaluation: Cued Speech recognition system based on the novel resynchronization procedure

In this experiment, 80% of the data is used as the training set, while the rest is the test set. The measure is the correctness $T_c = \frac{N-D-S}{N}$, where D is the number of deletion errors, S is the number of substitutions and N is the data size. For all the results, we take the average of ten experiments with different training and test sets. The results are shown in Fig. 6.

We observe that, based on the hand position features given by the Adaptive Background Mixture Models (ABMMs) [23], [24], using the architecture $S3$ in the state-of-the-art [8], the phoneme recognition obtained a recognition correctness 71.0%, without using any resynchronization procedure. When the proposed resynchronization is incorporated (i.e., using S_{re}), it increases to 72.67% (see the 3rd and 4th columns in Fig. 6). As we know, in the current recognition system, the triphone context-dependent modeling is helpful to correct the recognition errors due to the co-articulation or the asynchrony of multi-modalities [25]. Thus, the context-dependent modeling may hide the effect of the proposed resynchronization procedure. In order to get rid of this effect, we examine the recognition scores without using the context-dependent modeling. In this case, a correctness of only 60.4% is obtained without any resynchronization, while it increases to 64.38% when using the proposed resynchronization procedure (see the 1st and 2nd columns in Fig. 6). This improvement (about 4%) is more evident than the case using the context-dependent modeling (about 1.6%).

In fact, there are two possible reasons for the above weak improvements: (1) only a small weight of 0.2 is applied to the hand position stream, and this weight reduces the effect of the resynchronization procedure given the fact that hand position is more sensitive to the asynchrony problem than hand shape (introduced in Section IV-C); (2) the hand position stream extracted by the ABMMs may have some errors, which directly reduce the efficiency of the resynchronization procedure, since the hand position target can be identified only when the correct hand position[2] is selected with a good temporal boundary for a given vowel.

To reduce the effect of the above second reason, instead of using the hand positions given by the ABMMs, we use the ground truth hand positions, which are manually determined for all the images. The results are shown in the 5th to 8th columns of Fig. 6. We see that, without the context-dependent modeling and resynchronization procedure, a score of 62.33% is obtained (5th column), which is close to the result 60.4% (1st column). This can be explained by the above first reason. When the resynchronization procedure is used, a correctness of 70.1% is achieved (6th column), which shows a significant improvement (7.3%). In this case, the real benefit of the proposed resynchronization procedure is shown. Finally, we consider the case using the context-dependent modeling (see 7th to 8th columns in Fig. 6). Without the resynchronization procedure, the recognition correctness is 72.04%. However, when combined these two in the recognition system, an evidently higher score of 76.63% is obtained (with an improvement of 4.6%), outperforming the state-of-the-art [8], as well as the work of Heracleous et al., [4] with 74.4% correctness (in case of the isolated CS phoneme recognition).

VI. Conclusion

In this work, we propose a novel resynchronization procedure for the CS feature fusion in a CNN-HMMs continuous French CS recognition system. By exploring the optimal hand preceding time for all vowels ($140ms$) and for all consonants ($60ms$) in the sentences, and delaying the hand position and shape feature streams by these two different optimal hand preceding time, respectively, the lips and hand features can be resynchronized on average. The evaluation on the continuous phoneme CS recognition shows a significantly improvement (about 4.6%) after using this resynchronization procedure. In the future, we will 1) improve the accuracy of the automatic hand position tracking; 2) record more CS data, and explore the deep learning fusion methods, which might be able to exploit the asynchrony delay in the neural network training.

VII. Acknowledgement

The authors would like to thank the CS speaker for her time spent on the French CS data recording, and Thomas Hueber for his help in CNN-HMM. This work has been realized in Li Liu's PhD thesis at GIPSA-lab with a grant of the Universit Grenoble Alpes, France.

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] Panikos Heracleous, Denis Beautemps, and Noureddine Aboutabit, "Cued speech automatic recognition in normal-hearing and deaf subjects," *Speech Communication*, vol. 52, no. 6, pp. 504–512, 2010.

[5] 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.

[6] 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.

[7] Panikos Heracleous, Denis Beautemps, and Norihiro Hagita, "Continuous phoneme recognition in cued speech for french," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 2090–2093.

[8] Li Liu, Thomas Hueber, Gang Feng, and Denis Beautemps, "Visual recognition of continuous cued speech using a tandem cnn-hmm approach," in *Interspeech, 2018*, 2018, pp. 2643–2647.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[11] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[12] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[13] Li Liu, Gang Feng, and Denis Beautemps, "Automatic temporal segmentation of hand movement for hand position recognition in french cued speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference*, 2018, pp. 3061–3065.

[14] Lawrence R Rabiner and Biing-Hwang Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[15] Li Liu, Gang Feng, and Denis Beautemps, "Extraction automatique de contour de lèvre à partir du modèle clnf," in *Actes des 31èmes Journées d'Etude de la Parole*, 2016.

[16] Li Liu, Gang Feng, and Denis Beautemps, "Automatic tracking of inner lips based on clnf," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference*, 2017, pp. 5130–5134.

[17] Li Liu, Gang Feng, and Denis Beautemps, "Inner lips parameter estimation based on adaptive ellipse model," in *14th International Conference on Auditory-Visual Speech Processing (AVSP 2017)*, 2017.

[18] Jean-Luc Schwartz and Christophe Savariaux, "Data and simulations about audiovisual asynchrony and predictability in speech perception," in *12th International Conference on Auditory-Visual Speech Processing (AVSP 2013)*, 2013, pp. 147–152.

[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] Steve J Young and Sj Young, *The HTK hidden Markov model toolkit: Design and philosophy*, University of Cambridge, Department of Engineering, 1993.

[22] Steve J Young, Julian J Odell, and Philip C Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 307–312.

[23] 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.

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

[25] Jean-Luc Schwartz, Pierre Escudier, and Pascal Teissier, "Multimodal speech: Two or three senses are better than one," *Language and Speech Processing*, pp. 377–415, 2009.

[2] It has been reported in [8] that the lips and hand shape features extracted by CNNs are good.