INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Automatic detection of the temporal segmentation of hand movements in British English Cued Speech

Li Liu[†]*, Jianze Li*[‡]*, Gang Feng*[§] *and Xiao-Ping Zhang*[†]

† Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada
‡ Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China
§ Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-lab, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes

liuli1@ryerson.ca, lijianze@gmail.com, gang.feng@gipsa-lab.grenoble-inp.fr, xzhang@ee.ryerson.ca

Abstract

Cued Speech (CS) is a multi-modal system, which complements the lip reading with manual hand cues in the phonetic level to make the spoken language visible. It has been found that lip and hand movements are asynchronous in CS, and thus the study of hand temporal organization is very important for the multi-modal CS feature fusion. In this work, we propose a novel diphthong-hand preceding model (D-HPM) by investigating the relationship between hand preceding time (HPT) and diphthong time instants in sentences for British English CS. Besides, we demonstrate that HPT of the first and second parts of diphthongs has a very strong correlation. Combining the monophthong-HPM (M-HPM) and D-HPM, we present a hybrid temporal segmentation detection algorithm (HTSDA) for the hand movement in CS. The evaluation of the proposed algorithm is carried out by a hand position recognition experiment using the multi-Gaussian classifier as well as the long-short term memory (LSTM). The results show that the HTSDA significantly improves the recognition performance compared with the baseline (i.e., audio-based segmentation) and the state-of-the-art M-HPM. To the best of our knowledge, this is the first work to study the temporal organization of hand movements in British English CS.

Index Terms: British English Cued Speech, Asynchronous multi-modality, Hand temporal organization, Hand preceding model, LSTM.

1. Introduction

It was reported by the *World Health Organization* (WHO)[1] and *National Deaf Children's Society*[2] that over 5% of the world's population (466 million people) has disabling hearing loss (432 million adults and 34 million children) in the world, while over 45,000 deaf children live in the UK. Different systems have been developed to help the deaf people to understand each other better, such as the *lip reading* [1], *sign language* [2, 3], *cued speech* [4], *finger spelling* [5], etc.

In 1967, Cornett [4] invented the Cued Speech (CS) system, which complements the lip reading with manual cues (hand shapes and hand positions around the mouth). In this system, three streams of lips, hand positions and hand shapes are coherent and complementary to realize an efficient communication [4, 6]. Lip reading has a problem that some phonemes (like [u] and [y]) may look identical on the lips [7], while CS can overcome this problem by using the manual hand cues to distinguish them.

Up to now, CS has been adapted to more than 60 languages including American English, British English, French and so on. In the British English CS system, four hand positions are used to encode the 12 monophthongs and four hand slips are used to encode the 8 diphthongs (see Fig. 1), while eight hand shapes are used to encode the 24 consonants.

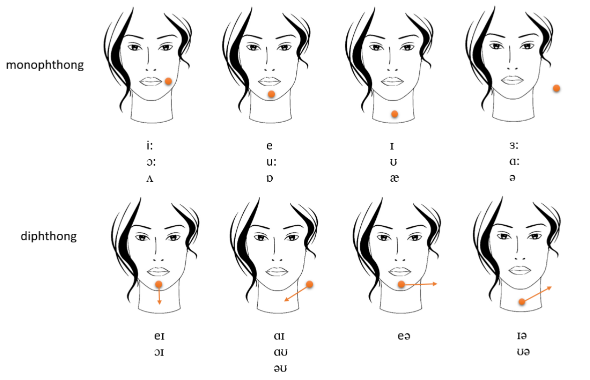

Figure 1: *Coding of vowels in British English CS. Four hand positions for monophthongs and four slips for diphthongs.*

It was found in [4, 6] that lips and hand movements are asynchronous in CS. Lips movement is more related to the phoneme production and hand movement is more related to the speech syllabic cycle. This asynchrony problem was studied in French CS system [8, 9] and it was shown that the hand reaches its target roughly 0.2s before the vowel being visible at lips based on a corpus made of "tatuta" logatome. More recently, in [10], based on a continuous French sentence corpus [11, 12, 13], the time interval that hand precedes lips movement is found to be 0.14s on average. In the automatic CS hand position recognition [14, 15, 16], the temporal segmentation is very important to train a good recognizer. In our previous work of French CS[3] [10], a hand preceding model (HPM) was proposed to predict the temporal segmentation of hand movements from the audio based segmentation. The French CS hand position recognition results show that the predicted temporal segmentation by HPM is superior to the baseline segmentation (i.e., audio based segmentation). Given the fact that there is only monophthong (no diphthong) in the French CS, the HPM in

[1] http://www.who.int/
[2] http://www.ndcs.org.uk/

[3] In French CS, five hand positions are used to code vowels.

http://dx.doi.org/10.21437/Interspeech.2019-2353

[10] is called monophthong-HPM (M-HPM) in this work. M-HPM is only suitable for the monophthong and thus it is not suitable for the diphthong in British English CS. Therefore, the motivation of this work is to explore a method to automatically detect the temporal segmentation of hand movement in British English CS.

This work mainly has the following contributions: 1) We investigate and confirm the phenomenon of hand preceding lips in British English CS for the first time, based on a new dataset recorded specifically for this work; 2) By analyzing the diphthong in the dataset, we find that the hand preceding time (HPT) of the first and second parts of the diphthong has a very strong correlation, and propose a novel diphthong-HPM (D-HPM) for the diphthong; 3) We present a hybrid temporal segmentation detection algorithm (HTSDA) by combining the M-HPM [10] and D-HPM that takes into account the case of the diphthong for British English CS. Based on the HTSDA, we derive the temporal segmentation of hand movement, and evaluate it by hand position recognition experiments using the multi-Gaussian and Long-Short Term Memory (LSTM) [17, 18]. It shows that the hand position recognition performance based on the HTSDA outperforms that of the baseline audio based segmentation (around 13%) and the M-HPM in [10].

2. Experiment setup

2.1. Cued Speech material

The first British English CS dataset[4] is recorded for this work in *Cued Speech UK*[5] association without using any artificial mark, and it is also the first one specially for the continuous recognition in British English CS. A professional CS interpreter (with no hearing impairment) is asked to simultaneously utter and encode a set of 97 British English sentences (e.g., *I feel it is a time to move to a new chapter in my career*). There are totally 907 monophthongs and 138 diphthongs in the dataset. Color video images of the interpreter's upper body are recorded at 25 fps, with a spatial resolution of 720x1280.

We assume that lips movement is synchronous with the audio signal. Then the temporal segmentation of audio speech can be used for lips movement. The audio based temporal segmentation is obtained using the viterbi algorithm based force alignment [19]. Since we find that the automatic alignment has some errors, we do a post-check based on the results of the force alignment.

We manually label the target temporal segmentation of hand movements for all vowels. In the lips stream, we use *Praat* [20] to get the temporal segmentation of audio signal. In the hand stream, we use the *MAGIX* software [21] to obtain the temporal segmentation of hand movements. For each vowel in the corpus, we denote by A_t and H_t the middle instant of the manually determined temporal segmentation of this vowel in the lips stream (audio signal) and hand stream, respectively. Then we define Δ_t (in s) to be the time difference:

$$\Delta_t = A_t - H_t. \tag{1}$$

For the audio speech signal, we notice that there is no time gap between two parts of the diphthong. The phonetic transcript is obtained automatically by the Lliaphon [22]. We automatically check all the audio based temporal segmentation of all the vowels and the corresponding phonetic transcription to determine the diphthong. For simplicity, in this work, we call the first and second parts of the diphthong D_1 and D_2, respectively (see Fig. 3).

2.2. Hand position feature

In the hand position recognition experiments, the hand position feature is extracted by the adaptive background mixture model (ABMM) [23], which was first proposed for the real-time segmentation of moving regions (car) in a video. In fact, in CS, the moving hand can be seen as the foreground, while other regions can be seen as the background. In this method, the mixture Gaussians are used to model the background. Any new pixel that does not match the background mixture Gaussian model is determined as the foreground hand. The center gravity of the detected hand pixels is defined as the hand position feature.

3. Hybrid temporal segmentation detection algorithm

The proposed hybrid temporal segmentation detection algorithm (HTSDA) takes into account the monophthong and diphthong in British English vowel by combining the M-HPM and D-HPM. The M-HPM was first proposed in [10] to obtain the temporal segmentation of hand movement for the monophthong in French CS. In this work, we propose a novel D-HPM that can obtain the temporal segmentation of hand movement for the diphthong.

3.1. Temporal organization of monophthong

The HPT (Δ_t) of all the 1045 vowels in the dataset is obtained by (1). We plot them in Fig. 2, and align all the vowels by the end of the sentences, which is considered as the instant 0. Generally, we can see that Δ_t remains stable from the beginning of the sentence to a turning point and then decreases from this point to the end. This phenomenon is very similar to the distribution between HPT and vowel instant in sentences for French CS [10].

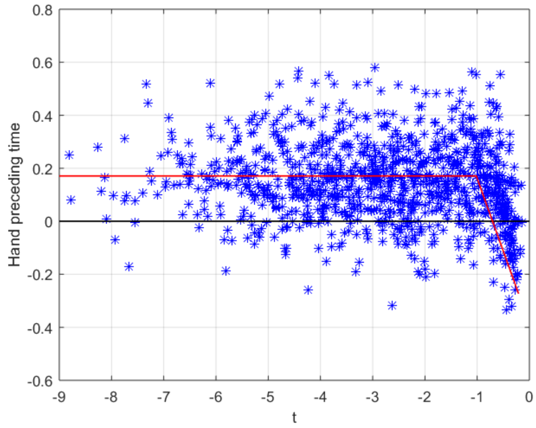

Figure 2: *Distribution of Δ_t in British English CS dataset. Abscissa is the vowel instant in the sentence, and Y axis is the corresponding HPT.*

The main difference is that the distribution for British English CS (see Fig. 2) has a larger variance than French CS case (0.150s vs. 0.077s). Besides, there are more negative Δ_t before the turning point in Fig. 2. Even these two studies are based on different speakers using different languages, we find that

[4] This dataset will be made publicly available on Zenodo.
[5] http://www.cuedspeech.co.uk/

the existence of diphthongs in British English (no diphthong in French) may be the main reason.

Since the temporal organization of D_1 is similar to the monophthong, based on the analysis of the temporal organization between lips and hand movements, for the monophthong and D_1, we develop a M-HPM (see the red polyline in Fig. 2) in the same way as [10]. It maintains a constant from the beginning to the turning point, and then follows a linear relationship from this turning point to the end of the sentence. More precisely, this model can be shown as follows:

$$\Delta_1(t) = \begin{cases} \overline{\Delta_t} & 0 \leq t < t_0^1, \\ at + b & t_0^1 \leq t \leq L, \end{cases} \tag{2}$$

where $\overline{\Delta_t}$ is the mean value (0.172s) of all the Δ_t between 0 and t_0^1. We experimentally set the turning point as 1s before the end, that is, $t_0^1 = L - 1$, where L is the length of the sentence. From t_0^1 to the end, a linear line with slope $a = -0.494$ is built from the dataset.

3.2. Temporal organization of the diphthong

Observing the British CS data, we find that diphthongs are considered as two consecutive monophthongs in hand movement. However, from the audio speech point of view, the audio speech cannot be regarded as the combination of two separated monophthongs [24]. This causes differences between the temporal organizations of D_1 and D_2. Taking the diphthong [eɪ] in the word *name* as an example, in Fig. 3, the audio signal and the hand position frames show the diphthong [eɪ]. We denote by A_1 and A_2 the target instant of [e] and [ɪ] in the audio signal, respectively, and denote by H_1 and H_2 the target instant of them in the hand movement, respectively. Then $\Delta_1 = A_1 - H_1 = 0.158s$, and $\Delta_2 = A_2 - H_2 = 0.063s$. We can see that the HPT of the first part of the diphthong [e] is much lager than that of the second part [ɪ], showing the temporal inconsistency between the two parts of diphthongs.

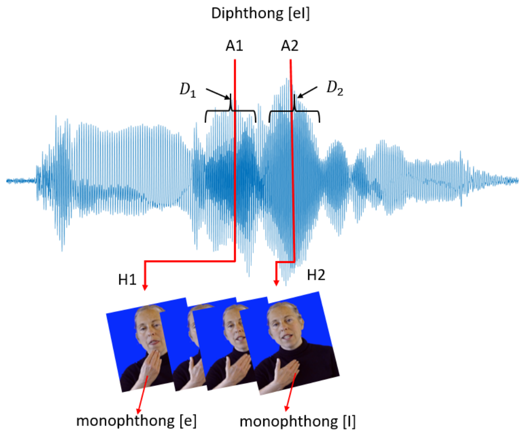

Figure 3: *Illustration of asynchrony for diphthong [eɪ] in British English CS.*

Since the temporal organization of D_1 can be modeled by M-HPM (see Section 3.1), it is important to know the relation of the HPT between D_1 and D_2 in order to explore the temporal organization of D_2. For this purpose, we calculate $\Delta\Delta$ (i.e., the distance of HPT between D_1 and D_2) by

$$\Delta\Delta = (t^p_{D_2} - t^a_{D_2}) - (t^p_{D_1} - t^a_{D_1}) \tag{3}$$

for all 138 diphthongs in the dataset. We denote the target instant of D_1 as $t^a_{D_1}$ and the target instant of D_2 as $t^a_{D_2}$ for all the diphthongs in the dataset. Concerning the hand temporal segmentation, the target instant of D_1 is denoted by $t^p_{D_1}$, and the target instant of D_2 is denoted by $t^p_{D_2}$, for all the diphthongs in the dataset.

Algorithm 1 Hybrid temporal segmentation detection algorithm (HTSDA)

```
Input: Audio based temporal segmentation $[A_{t_1}, A_{t_2}]$ for all vowels;
Output: Temporal segmentation for hand movement $[H_{t_1}, H_{t_2}]$ for all vowels.
  for each vowel $v$ do
    if $v$ is monophthong then
      transform $[A_{t_1}, A_{t_2}]$ by M-HPM in Eq. (2).
    else
      for each $D_1$ of the diphthong do
        transform $[A_{t_1}, A_{t_2}]$ by M-HPM in Eq. (2).
      end for
      for each $D_2$ of the diphthong do
        transform $[A_{t_1}, A_{t_2}]$ by D-HPM in Eq. (4).
      end for
    end if
  end for
```

$\Delta\Delta$ for all the 138 diphthongs is plotted in Fig. 4, where abscissa is the target time instant of D_2 (i.e., $t^a_{D_2}$) in the sentences and Y-axis shows the $\Delta\Delta$ for each diphthong. In order to show the distribution clearly, we fix $t = 0$ as the end of the sentence in the figure. The standard deviation (std) of all $\Delta\Delta$ is about 0.077s, which is small. A statistic significant test shows that the HPT of D_1 and D_2 has a significant correlation with a P-value near 0, and their distance stays at about 0.1s on average. Besides, we find that after 0.8s until the end of the sentence, the $\Delta\Delta$ of the diphthong increases linearly.

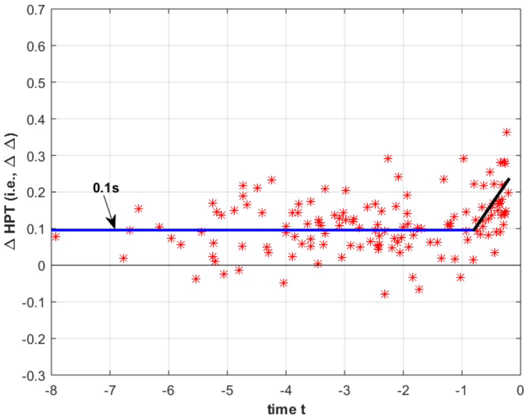

Figure 4: *Distribution of HPT's distance between D_1 and D_2 (i.e., $\Delta\Delta$) for 138 diphthongs in the dataset.*

Concerning the above analysis, we build the following D-HPM for the D_2 of the diphthong

$$\Delta_2(t) = \begin{cases} \overline{\Delta_t} - \overline{\Delta\Delta_t} & 0 \leq t < t_0^2, \\ at + b - (ct + d) & t_0^2 \leq t \leq L, \end{cases} \tag{4}$$

where $\overline{\Delta_t}$ is the same as Eq. (2), which is the average HPT of the monophthong. $\overline{\Delta\Delta_t} = 0.1s$ is the average Δ_t between D_1 and D_2 (see the blue line in Fig. 4). a, b are the same as Eq. (2), and $c = 0.225$, $d = 0.28$ are the slope and intercept of the black linear line in Fig. 4. t_0^2 is the turning point in Fig. 4 which is about $L - 0.8s$, where L is the length of the sentence.

3.3. Hybrid temporal segmentation detection algorithm for British English CS

Based on the above M-HPM and D-HPM, we propose a HTSDA to automatically detect the temporal segmentation of hand movement in CS (see Algorithm 1). The temporal segmentation is obtained by shifting the audio based segmentation with the corresponding $\Delta_1(t)$ in Eq. (2) for each monophthong and D_1 of the diphthong, and the corresponding $\Delta_2(t)$ in Eq. (4) for each D_2 of the diphthong.

4. Evaluation and Discussion

In order to evaluate the HTSDA composed of the M-HPM and D-HPM, we carry out experiments with the multi-Gaussian classifier and LSTM to automatically recognize the hand positions. We take 80% of the dataset as the training set, while the rest is the test set. The hand feature is a two-dimensional vector obtained by ABMMs (introduced in Section 2.2). All the final results are the average of 100 experiments with different training and test sets. The recognition results are shown in Fig. 5.

In the multi-Gaussian model, four Gaussian models are trained for the four hand positions. Given any test data, the Gaussian model with the maximum probability will be the right class. It can be seen in Fig. 5 that the audio based segmentation only obtains 47.98% recognition score, while a significant improvement is obtained (58.74%) using the temporal segmentation based on the temporal segmentation predicted by the M-HPM. Furthermore, by using the proposed D-HPM, the hand position recognition accuracy increases about 2% retaining 60.68% (std is about 0.2%). However, it is still worse than the one using the ground truth temporal segmentation (77.49%).

Note that LSTM achieves good performance in audio and visual speech recognition [25, 26], since it can exploit a self-learnt amount of long-range temporal information. This ability may be able to deal with some noise in CS coding (e.g., the speaker's hand does not reach the the right position), and affect the recognition performance.

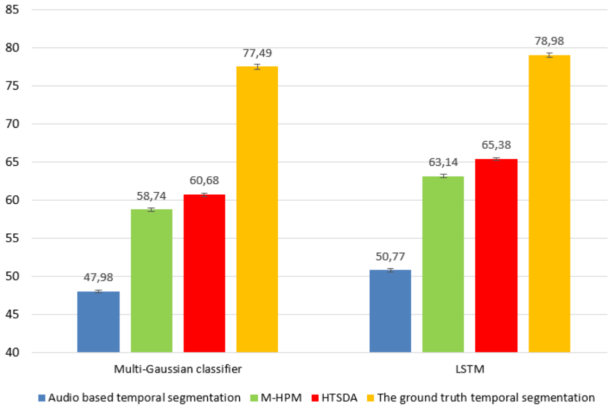

Figure 5: *Hand position recognition accuracy using the multi-Gaussian and LSTM based on the audio based segmentation, the predicted segmentation and the ground truth segmentation.*

In LSTM, two hidden layers of 100 cells, 80 epochs and the dropout rate of 0.5 are used. It is trained by the classical Back-Propagation Through Time (BPTT) with cross-entropy as the loss function. The posterior probability is calculated by the softmax function and the final label is obtained using the max-voting (i.e., the most frequent label in one temporal segmentation is the final decision). Keras toolkit [27] is used for the LSTM implementation with the GPU-accelerated library.

In Fig. 5, firstly, we can see that LSTM obtains a higher accuracy than multi-Gaussian since it captures some temporal information of hand movements to some extent. Then, we compare the hand position recognition result using the temporal segmentation predicted by the HTSDA with that using the audio based and ground truth segmentation, respectively. The results confirm the advantages of the proposed method. More precisely, the temporal segmentation based on M-HPM obtains 63.14% accuracy, which significantly outperforms the audio based segmentation (50.77%). More importantly, the proposed HTSDA obtains 65.38% accuracy, outperforming the M-HPM by 2.24% (std is about 0.2%).

Compared with the results obtained for French CS [10] using the ground truth temporal segmentation and hand position obtained by the ABMMs, a huge difference on the hand position recognition accuarcy (77.49% vs. 62.26%) is observed using the multi-Gaussian classifier. This is probably caused by the fact that there are only four hand positions to be recognized in British English CS, making the recognition easier than the French CS with five hand positions. However, this difference is not obvious when using the non-ground truth temporal segmentation (e.g., 47.98% vs. 45.41% using the audio based temporal segmentation). It seems that hand position recognition result for British English CS is more sensitive to the temporal segmentation than the French CS. Indeed, British CS has a much larger variance (std = 0.150 s) in the distribution of HPT Δ_t (see Fig. 2) than French CS (std = 0.077s). This sensitivity could also explain the huge difference between 77.49% and 60.68% when using the ground truth temporal segmentation and the non-ground truth segmentation in British English CS.

5. Conclusion

In this work, we propose a novel D-HPM concerning the diphthong temporal organization and develop a HTSDA to automatically detect the temporal segmentation of the hand movement in British English CS. Compared with French CS case, the main difference in the distribution of HPT caused by the diphthongs in British English is explored. The proposed HTSDA forms a hybrid HPM which takes into account the particularity of the diphthong by D-HPM and the monophthong by M-HPM. The evaluation on the predicted temporal segmentation by the HTSDA is carried out by hand position recognition experiments using multi-Gaussian classifier and LSTM. They both confirm the efficiency of the proposed HTSDA. In the future, we will adjust the proposed algorithm to a multi-speaker dataset with CS of other languages.

6. Acknowledgement

The authors would like to thank the professional CS teacher Cate Calder from *Cued Speech UK* for the British English CS dataset recording. The work was supported in part by the Natural Sciences Engineering Research Council of Canada (NSERC, Grant No. RGPIN239031) and Shenzhen Peacock Plan under Grant No. KQTD2015033114415450.

7. References

[1] B. Dodd, "Lip reading in infants: Attention to speech presented in-and out-of-synchrony," *Cognitive psychology*, vol. 11, no. 4, pp. 478–484, 1979.

[2] W. C. Stokoe, D. C. Casterline, and C. G. Croneberg, *A dictionary of American Sign Language on linguistic principles*. Linstok Press, 1976.

[3] W. C. Stokoe Jr, "Sign language structure: An outline of the visual communication systems of the american deaf," *The Journal of Deaf Studies and Deaf Education*, vol. 10, no. 1, pp. 3–37, 2005.

[4] R. O. Cornett, "Cued speech," *American annals of the deaf*, vol. 112, no. 1, pp. 3–13, 1967.

[5] C. Padden and C. Ramsey, "American sign language and reading ability in deaf children," *Language acquisition by eye*, vol. 1, pp. 65–89, 2000.

[6] C. J. LaSasso, K. L. Crain, and J. Leybaert, *Cued Speech and Cued Language Development for Deaf and Hard of Hearing Children*. Plural Publishing, 2010.

[7] G. H. Nicholls and D. L. Mcgill, "Cued speech and the reception of spoken language," *Journal of Speech, Language, and Hearing Research*, vol. 25, no. 2, pp. 262–269, 1982.

[8] V. Attina, D. Beautemps, M.-A. Cathiard, and M. 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.

[9] V. Attina, M.-A. Cathiard, and D. Beautemps, "Temporal measures of hand and speech coordination during french cued speech production," in *International Gesture Workshop*. Springer, 2005, pp. 13–24.

[10] L. Liu, G. Feng, and D. 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.

[11] ——, "Extraction automatique de contour de lèvre à partir du modèle clnf," in *Actes des 31èmes Journées d'Etude de la Parole*, 2016.

[12] ——, "Automatic tracking of inner lips based on clnf," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference*, 2017, pp. 5130–5134.

[13] ——, "Inner lips parameter estimation based on adaptive ellipse model," in *14th International Conference on Auditory-Visual Speech Processing (AVSP 2017)*, 2017.

[14] ——, "Inner lips feature extraction based on clnf with hybrid dynamic template for cued speech," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 88, 2017.

[15] L. Liu, T. Hueber, G. Feng, and D. Beautemps, "Visual recognition of continuous cued speech using a tandem cnn-hmm approach," in *Interspeech, 2018*, 2018, pp. 2643–2647.

[16] L. Liu, "Modeling for continuous cued speech recognition in french using advanced machine learning methods," Ph.D. dissertation, Université Grenoble Alpes, 2018.

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

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

[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[20] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[21] D. V. Abreu, T. K. Tamura, D. G. Keamy Jr, R. D. Eavey *et al.*, "Podcasting: contemporary patient education," *Ear, Nose & Throat Journal*, vol. 87, no. 4, p. 208, 2008.

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

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

[24] E. Ferragne and F. Pellegrino, "Formant frequencies of vowels in 13 accents of the british isles," *Journal of the International Phonetic Association*, vol. 40, no. 1, pp. 1–34, 2010.

[25] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE-ICASSP*, 2013, pp. 6645–6649.

[26] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with lstms," in *Proc. IEEE-ICASSP*, 2017, pp. 2592–2596.

[27] F. Chollet *et al.*, "Keras," 2015.