Livrable WP3 - L1 : Système prédictif des moments de transition de positions

Brigitte Bigi

Automatic Cued Speech / LfPC automatisée

La réalité augmentée au service des personnes sourdes

Livrable WP3 - L1 : Système prédictif des moments de transition de positions

Objectif

Dans le cadre de la création d'un système de génération automatique du codage, la deuxième étape requise consiste à déterminer les moments de transitions à partir de la séquence des clés qui ont été définies à partie des phonèmes prononcés. Ce document s'attache à la prédiction des transitions des positions de la main, les transitions de formes ne faisant pas l'objet de ce projet.

Afin de répondre à cet objectif, il est nécessaire de disposer en amont d'un système qui permet d'obtenir la séquence des phonèmes synchronisés avec l'audio, ainsi que les clés qui doivent être codées. Ce système a été développé dans le cadre de ce projet ; il est décrit dans le document WP2-L1.

Le système de prédiction des moments de transition des positions de la main est indépendant de la langue.

Description du système

1. Définitions et objectifs

Dans la suite de ce document, nous utiliserons la terminologie introduite dans les travaux de thèse (Attina, 2005) :

- A1 représente le début de la consonne de la clé
- A3 représente la fin de la voyelle de la clé
- A3A1 désigne la durée de l'intervalle entre les moments A1 et A3, donc la durée des phonèmes prononcés pour la clé

- M1 correspond au début de la transition manuelle vers la position cible
- M2 correspond à l'atteinte de cette cible, donc à la fin de la transition manuelle

Le système décrit dans ce document a pour objectif de prédire automatiquement les valeurs M1 et M2, qui représentent la transition de la main de la position actuelle vers la position cible.

2. Les solutions existantes

Modèle 1 : (Duchnowski et al., 1998)

Pour répliquer le système proposé dans l'article (Duchnowski et al., 1998-1), nous nous sommes appuyés sur le paragraphe suivant extrait de la section 2.2 de cet article: *"We found that cues are often formed before the corresponding sound is produced. To approximate this effect we adjusted the start times of cues to begin 100 ms before the boundary determined from acoustic data by the cue recognizer."*

On retrouve plus ou moins le même énoncé dans l'article (Duchnowski et al., 1998-2) en section 3.2 : *"We observed that human cuers often begin to form a cue before producing the corresponding audible sound. To approximate this effect we adjusted the start times of the cues to begin 100 ms before the boundary determined by the cue recognizer."*

Ainsi, dans ce système, il n'existe pas de temps de transition pour déplacer la main d'une position à l'autre. En revanche, il établit une règle fixe : la main doit se trouver à la position cible 100ms avant le début de l'énonciation des phonèmes. Les intervalles sont donc estimés tels que (valeurs en secondes) : `M1 = [A1-0.1, A1-0.1]` et `M2 = [A1-0.1, A1-0.1]`.

Ci-après se trouve un aperçu du code que nous avons implémenté comme système 1 :

```
m1 = a1 - 0.100
m2 = a1 - 0.100
```

Modèle 2 : (Duchnowski et al., 2000)

Nous avons également répliqué le système proposé dans l'article (Duchnowski et al., 2000). Nous nous sommes appuyés sur la citation suivante, extraite de la section III.C (page 491) : *"The 'dynamic' display used heuristic rules to apportion cue display time between time paused at target positions and time spent in transition between these positions. Typically, 150 ms was allocated to the transition provided the hand could pause at the target position for at least 100 ms. The movement between target positions was, thus, smooth unless the cue was short, in which case it would tend to resemble the original 'static' display."*

Le modèle 'statique' mentionné dans cette citation fait référence au modèle 1 ci-dessus. Ainsi, ce système conserve le principe d'arriver à la position cible 100ms avant de prononcer les phonèmes de la clé. En revanche, il alloue un temps de transition de 150ms pour atteindre cette cible.

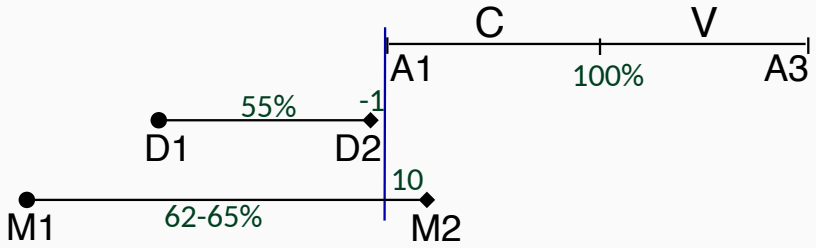
Ci-après se trouve un aperçu du code que nous avons implémenté comme système 2 :

```
m1 = a1 - 0.250
m2 = a1 - 0.100
```

Modèle 3 : (Attina, 2005)

Dans sa thèse, V. Attina (2005) n'a pas, à proprement parlé, proposé de système automatique, mais elle a conduit différentes expériences afin d'analyser la synchronisation main-son. Un schéma représentant cette synchronisation est proposé page 136, pour lequel la durée des phonèmes de la clé est de 399.5ms.

Ces expériences ne permettent pas de couvrir l'ensemble des situations qui peuvent se présenter lors du codage puisque seules les syllabes en contexte ont été étudiées. Pour implémenter un système à partir du modèle qui a été proposé, nous avons dû le généraliser sans pouvoir vérifier la validité de cette généralisation.



Reproduction partielle de la figure 32 pp. 136 (Attina, 2005).

Ci-après se trouve un aperçu du code que nous avons implémenté comme système 3 :

```
a3a1 = 0.399

if target_key in ('CV', 'C'):
      # M1 is 62% of A1A3
      m1a1 = a3a1 * 0.62
      # M2 is 10% later than A1
      a1m2 = a3a1 * 0.10
else:
      # M1 is 46% of A1A3.
      m1a1 = a3a1 * 0.46
      # M2 is 21% later than A1
      a1m2 = a3a1 * 0.21

m1 = a1 - m1a1
m2 = a1 + a1m2
```

3. Le système proposé

Comme pour les systèmes précédents, notre système repose sur un ensemble de règles qui ont été établies grâce à cet état de l'art, mais surtout grâce à notre expertise du codage.

Le système que nous avons implémenté s'appuie donc sur les travaux précédents ainsi que sur ceux présentés dans (Bratakos et al., 1998). En particulier, nous avons noté le paragraphe suivant : *"A critical delay to display the hand cue is +165ms. The max delay is +100ms. A non-significant delay is +33ms."* Ainsi, il est important que la main soit *relativement en avance sur le son*, les retards étant critiques pour la compréhension. De ce fait, nous avons défini M2 en choisissant un compromis entre le modèle 2 (100ms avant A1) et le modèle 3 (10% après A1) : M2 est placé 5% avant A1.

Nous avons également établi les principes suivants pour estimer [M1,M2] :

- Identiquement au système de (Duchnowski et al. 1998), nous ne tenons pas compte de la structure de la clé.
- La transition entre la position neutre et celle de la première clé est très anticipée par rapport au son.
- La transition entre la première clé et la deuxième est relativement plus anticipée par rapport au son que les transitions suivantes.
- La transition entre la dernière clé d'une séquence prononcée et la position neutre est différée.

Contrairement aux systèmes précédents qui utilisent des valeurs fixes de durées, nous faisons l'hypothèse que le temps de transition de la position doit être *modérément ajusté* en fonction du débit de parole. Effectivement, nous supposons que lorsque le débit augmente, c'est le temps d'exposition de la clé qui sera le plus impacté (nettement réduit) tandis que le mouvement d'une position à l'autre le sera dans une moindre mesure (vitesse plus élevée). Nous avons donc utilisé une valeur fixe de durée A3A1 uniquement pour la première clé, et nous avons utilisé la moyenne des durées observées pour les clés suivantes. Cette hypothèse devra toutefois être vérifiée par l'observation d'annotations des mouvements de la main dans un corpus.

Ci-après se trouve un aperçu du code que nous avons implémenté comme système 4 :

```
a3a1 = mean(all(a3a1))

if target_key == 'N':
      m1 = a1 + (a3a1 * 0.2)
      m2 = m1 + (a3a1 * 0.8)
else:
      m2a1 = a3a1 * 0.05
      if rank == 1:
            m1a1 = max(0.500, a3a1 * 1.25)
      elif rank == 2:
```

```
            m1a1 = max(0.250, a3a1)
      else:
            m1a1 = a3a1 * 0.9

      m1 = max(0., a1 - m1a1)
      m2 = max(m1, a1 - m2a1)
```

Accès et utilisation du système

La version stable de ce système est distribuée sous les termes de la licence GNU GPL v3. Elle fait partie du logiciel SPPAS, depuis la version 4.17, et peut être téléchargée à l'adresse : https://sppas.org/ .

Pour obtenir le codage de ce système sur la démo proposée dans SPPAS, il est possible d'utiliser l'interface graphique de SPPAS, ou la commande en ligne suivante:

```
> python sppas/bin/cuedspeech.py -I demo/demo.mp4 -l fra --createvideo=true --handtrans=4
```

L'option "handtrans" permet de choisir le numéro du modèle comme décrit dans ce document. Cette commande permet de créer trois fichiers : deux fichiers de description au format XML qui contiennent l'ensemble des informations prédites par le système, ainsi que la vidéo codée automatiquement.

Références bibliographiques

Maroula S. Bratakos, Paul Duchnowski and Louis D. Braida (1998). *Toward the Automatic Generation of Cued Speech*. Cued Speech Journal VI 1998 p1-37.

PDF

Paul Duchnowski, Louis D. Braida, David S. Lum, Matthew G. Sexton, Jean C. Krause, Smriti Banthia (1998). *AUTOMATIC GENERATION OF CUED SPEECH FOR THE DEAF: STATUS AND OUTLOOK*

PDF

Paul Duchnowski, Louis D. Braida, Maroula Bratakos, David S. Lum, Matthew G. Sexton, Jean C. Krause. (1998) *A SPEECHREADING AID BASED ON PHONETIC ASR*

PDF

Paul Duchnowski, David S. Lum, Jean C. Krause, Matthew G. Sexton, Maroula S. Bratakos, and Louis D. Braida (2000). *Development of Speechreading Supplements Based in Automatic Speech Recognition* . IEEE Transactions on Biomedical Engineering, vol. 47, no. 4, pp. 487-496. doi: 10.1109/10.828148.

Virginie Attina Dubesset (2005). *La langue française parlée complétée (LPC) : production et perception.* PhD Thesis of INPG Grenoble, France.

PDF

Contributeurs

Développement logiciel : Brigitte Bigi (LPL)

Expertise du codage : Datha

Dernière mise à jour : 29 janvier 2024

Nos résultats À propos

- Logiciel SPPAS
- Capsules vidéos
- Publications scientifiques
- Plan du site
- Mentions légales
- Nous contacter
- Accessibilité

Projet financé par la FIRAH (2023-2026)