On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French

Xiaoou Wang*, Thomas Hueber, Pierre Badin

GIPSA-Lab / DPC, UMR 5216, CNRS – Grenoble Alpes University, France

rosingle871113@gmail.com, (thomas.hueber, pierre.badin)@gipsa-lab.grenoble-inp.fr

Abstract

This study investigates the usefulness of a French Articulatory Talking Head (ATH) to train Mandarin Chinese speakers to produce French vowels that do not exist in their phonetic repertoire. Two groups of Mandarin Chinese speakers were trained using either auditory or audiovisual stimuli displayed with the ATH, based on a speech shadowing task of immediate repetition of vowels and VCV sequences. We found that the F1 and F2 formants of vowels improved for both groups and that the audiovisual group (AVG) improved more than the auditory group (AG), which tends to validate the interest of the vision of articulators for the pronunciation training.

Keywords: *articulatory talking head (ATH), computer aided pronunciation training (CAPT), second language acquisition, Chinese Mandarin, French*

1. Introduction

Human begins to use articulatory information in speech perception at a very young age. Kuhl & Meltzoff (1982) have found that 18- to 20-week-old infants recognize already the correspondence between auditorily and visually presented sounds. This behavior will persist into adulthood, as shown by the McGurk Effect (McGurk & MacDonald, 1976). The use of such information in second language pronunciation training could thus be beneficial and has already started to spread. Traditional studies tended to show learners their own articulation with different instruments for pronunciation training, including ElectroPalatoGraphy (Schmidt & Beamer (1998)) or ultrasound (Bressmann *et al.*, 2005), while other studies used what we call here an “articulatory talking head” (ATH) capable of displaying the movements of both external and internal articulators (such as the tongue) (Massaro *et al.*, 2008; Engwall, 2012). The present study adopted the second approach.

Although the overall contribution of ATH in pronunciation training has already been investigated (*cf. e.g.* Massaro & Light, 2003), the benefit of articulatory information compared with traditional trainings in which only auditory instructions are available has not so far received such attention. For this purpose we recruited two groups of Chinese students learning French to participate respectively in an auditory training (Auditory Group, AG) and an audiovisual training (AudioVisual Group, AVG). Due to the influence of the first language, Chinese learners typically have difficulty pronouncing correctly /e o ɔ ø œ/. According to the Perceptual Assimilation Model (PAM: Best, 1991), they would probably assimilate the first three vowels to three diphthongs /ei/, /ɤʊ/, /ɑʊ/ and the last two to /ɤ/, a high-mid back unrounded vowel. Our training centered thus on these five vowels.

2. Methods

2.1. Participants

We report the data from 2 groups of 7 female native speakers of Mandarin Chinese recruited in a language institute, including 4 speakers from Peking (in the AG group), 4 from north-eastern China (in the AVG group) and 6 from various districts of northern China. They had all studied French for 6 months and their ages ranged from 22 to 23. All of them had basic knowledge of English and none reported hearing or reading disorders.

2.2. The articulatory talking head

We used the ATH developed at GIPSA-lab (Badin *et al.*, 2008) to generate the stimuli used all along the training. Two advantages of this ATH are worth mentioning. Besides its ability to display the complete set of articulators (*e.g.* lips and tongue), the stimuli produced by this ATH were built from original natural speech sounds and articulatory movements recorded synchronously by an ElectroMagnetic Articulograph (EMA) device on one French speaker. This approach results in very naturally moving animations, in contrast to approaches based on a re-synthesis of these movements from acoustic or phonetic specifications, which may produce less realistic movements. For a detailed presentation, see Badin *et al.* (2008).

2.3. Training Corpus

The training corpus was composed of single vowels and of VCV sequences built with the aforementioned ATH. The single vowels were /ɛ u y i e o ɔ ø œ/ and the consonants in the VCV sequences were /p t k f s/. Since the durations of the original sequences were too short, we used a time stretching algorithm (based on Harmonic plus Noise modeling (Stylianou, 1996)) to artificially increase the vowel length so that the learners had enough time to observe the whole movement. In order to preserve natural coarticulation patterns, time stretching was applied only to stable parts of the vowels.

The training with single vowels was organized into five blocks of three drills. Each block focused on one of the five vowels that the subjects should learn. In each block, the three drills presented successively a reference vowel and the vowel to be learned. At the end of the drill, the images corresponding to the stable state of the vowels were retained on the screen for 2 seconds, allowing the learners to better understand their articulation by comparison. Figure 1 illustrates this method with /ø/ and /œ/. Table 1 lists the drills used for each block.

[*] At present: Laboratoire de Phonétique et Phonologie, Paris, France.

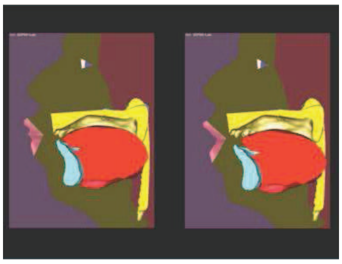

Figure 1: *A drill presenting /ø/ (left) and /œ/ (right)*

Table 1: *List of drills for each block (single vowels)*

Block	Drill
/e/	/ɛ ø i/ vs. /e/
/o/	/u ø ɔ/ vs. /o/
/ɔ/	/u œ o/ vs. /ɔ/
/ø/	/e o œ/ vs. /ø/
/œ/	/ɛ o ø/ vs. /œ/

The training with VCV sequences was organized into three blocks of five drills. In each block we provided an intensive training for two confusing vowels combined with consonants /p t k f s/ to form five contrastive drills. This procedure was also designed with an eye to training coarticulation. Table 2 lists the corresponding drills and blocks.

Table 2: *List of drills for each block (VCV sequences)*

Block	Drill
/i/ ~ /e/	/ipi/ vs. /epe/, and so on
/o/ ~ /ɔ/	/opo/ vs. /ɔpɔ/, and so on
/ø/ ~ /œ/	/øpø/ vs. /œpœ/, and so on

2.4. Pronunciation test corpus

Each learner was required to read a speech corpus before and after the training so that the changes of her pronunciation could be assessed in later analysis. The corpus contained all French oral vowels except /y/. 2 isolated productions and 5 VCV sequences were recorded for /a i u ɛ/, while 9 VCV sequences and 5 words were recorded for the other vowels. The detail is presented in Table 3.

Table 3: S*peech corpus used for pronunciation tests*

Vowel	Syllable	Word
/a i u ɛ/	Two isolated productions + VCV sequences with /p t k f s/ as consonants, no words	
/e/ /o/ /ɔ/ /ø/ /œ/	Two isolated productions + VCV sequences with /p t k f s z m n l/ as consonants	Pépé, mémé, café, casser, télé Peau, mot, faux, morceau, tôt Porc, molle, folle, sort, tort Peu, fameux, feu, paresseux, deux Peur, fumeur, fleur, seul, auteur

2.5. Speech perception test corpus

A perception test (forced choice identification) was also presented before and after the training. The corpus contained one isolated vowel production and three VCV sequences with /t s l/ as consonants for /e ɛ o ɔ ø œ/. Each stimulus was played once. The results of this test could inform us of the learners' discriminative capacities.

2.6. Protocol

The whole experiment was performed using the *Presentation*® software (Version 16.5, www.neurobs.com). During the experiment the operator could monitor the learner's performance via a screen located outside the soundproof room where the learner was seated and could communicate with the latter by an intercom system. Before the experiment started, each learner was required to listen to an isolated production of /i/ to allow her to adjust the earphone volume to a comfortable level.

The experiment was divided into 6 phases. The learners were at first given two minutes to get accustomed with the phonetic symbols of each vowel (phase 1). Since the IPA was abstruse for some learners, we used /E O eu ou u/ to represent /ɛ ɔ ø u y/. They heard successively a word, an isolated vowel production and two VCV sequences using /t k/ as consonants. They were then submitted to the pronunciation test (phase 2). Next, they performed the perception test during which they gave their responses by clicking one of the six buttons representing /e ɛ o ɔ ø œ/ after hearing a stimulus (phase 3). The stimuli in the perception test were randomized using the default algorithm in *Presentation*®.

Phase 4 was intended for training. At first the learners watched a three-minute video. At the beginning, some animations were displayed in order to get the AVG learners familiar with the ATH whereas the AG learners simply listened. Then the video explained the speech shadowing task (*cf.* Shockley *et al.*, 2004) demonstrated by a French speaker. This task required each subject to repeat the stimulus as soon as they perceived it. The underlying idea was that it would help them imitate better the stimuli, as suggested by Shockley *et al.* (2004). The learners were then given one minute to practice the speech shadowing, using /i/~/y/ and /u/~/y/ as drills.

The training per se started with single vowels training. The learners saw and heard each drill twice. The first time they were requested to observe the tongue movement and the second time the tongue and lips movement. Then they saw the animations on the screen and started to imitate each animation using speech shadowing. The animations were also presented two times. The training with VCV sequences followed the same procedure, except that the learners saw and heard each drill only once. These two training periods lasted 15 minutes. After the training, the learners passed again the pronunciation test (phase 5) and the perception test (phase 6).

The experiment for the AG learners followed exactly the same procedure, except that the animations of the ATH were not provided. The whole experiment never exceeded 40 minutes.

2.7. Data collection and acoustic measurements

For each group, the 672 tokens of collected /a i u ɛ/ consisted of 12 tokens each produced by 7 speakers before and after training (4×12×7×2) and the 1750 tokens of collected /e ɛ o ɔ ø œ/ consisted of 25 tokens each produced by 7 speakers before and after training (5×25×7×2).

The audio data were recorded in a soundproof room at a 44,100 Hz sampling rate with 16-bit resolution and later down sampled to 22,050 Hz before acoustic analyses.

Vowels were segmented by means of the *Praat* software (Boersma & Weenink (2005)). We used the second formant as principal cue for the segmentation of the vowel in VCV sequences. Formant frequency analysis was performed using the LPC (autocorrelation) algorithms available in *Praat* with default settings for females. A rapid overview of the data revealed great variations among learners for the pronunciation of /e o ɔ/. Some diphthongized, while the others pronounced differently, probably due to their knowledge of American English. The pronunciation of /ø/ and /œ/ revealed a similar pattern, though not completely homogenous, as we shall see later. This observation led us to the following decision. Formants of the vowels other than /ø/ and /œ/ were extracted automatically, while formants of /ø/ and /œ/ were extracted manually in the stable part of each vowel production. Some accidental pronunciation mistakes (/y/ for /œ/) were also discarded during this process so that 170 tokens of /ø/ and /œ/ were used for acoustic analysis. For the other vowels, all the collected tokens were used.

3. Results

3.1. Overall improvement

Since the pronunciation of /e o ɔ/ varied considerably in each group, we decided to limit the analysis for the moment to /ø/ and /œ/. To visualize the overall change of all the learners, formants measured on vowels before and after training were displayed in the F1/F2 space, as shown in Figure 2 and Figure 3. Means of 9 French female speakers (Calliope, 1989), extracted from two repetitions of /pVʁ/ for /œ ɔ i ɛ/ and /pV/ for the other vowels, were superposed for reference.

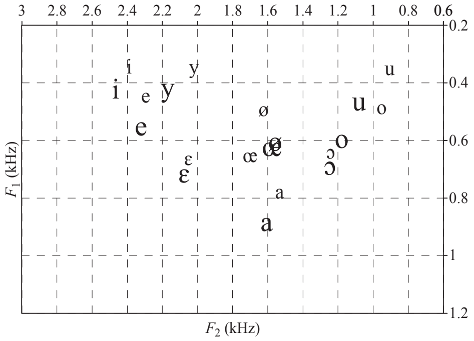

Figure 2: *Centroids of F1-F2 for French vowels produced by Chinese learners* before training *(large symbols) averaged over both groups. The reference values (Calliope, 1989) are marked by small symbols.*

Figure 2 confirms the predictions of the PAM model according to which /ø/ and /œ/ would both be assimilated to /ɤ/. The F1s of these two vowels are close and lie between the natives' values, while the F2s are lower than standard values.

The progress could be thus assessed by analyzing the values of F1 and F2 both in terms of differences between before and after training and in terms of distances from typical values of French. Namely, the Chinese learners should raise their F2s for both vowels, while they should lower their F1s for /ø/ and raise their F1s for /œ/. Figure 3 shows that after training the learners began to differentiate the two vowels.

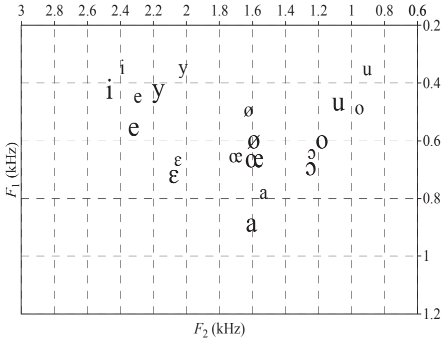

Figure 3: *Centroids of F1-F2 for French vowels produced by Chinese learners* after training (large symbols) *averaged over both groups. The reference values (Calliope, 1989) are marked by small symbols.*

3.2. Progress in each group

The mean frequencies of the F1 and F2 vowel formants before and after training were computed for each group. A paired-samples t-test was conducted to quantify the formant change before and after training, while an independent-samples t-test was conducted to quantify the difference of the formant change between two groups. To better understand the changes in each group, the statistical distributions of the F1 and F2 measured on all learners were displayed as boxplots, as shown in Figure 4 and Figure 5.

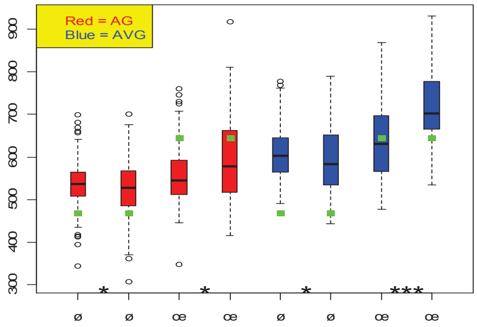

Figure 4: *Distributions of F1 for /ø/ and /œ/ before and after training. For each vowel the first bar represents the state before training and the second one the state after training (* denotes differences between before after training that are significant at $p<0.05$ and *** at $p<0.001$). The green squares represent the values from Calliope (1989).*

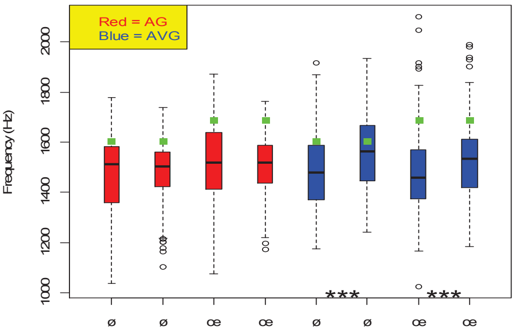

Figure 5: *Distributions of F2 for /ø/ and /œ/ before and after training. Same conventions as in Figure 4.*

It can be seen that the overall differentiation of these two vowels was mainly due to the F1 shift of the two vowels in both groups and the F2 increase for both vowels in AVG.

The changes of the means of F1 and F2 before and after training for each vowel in each group led to the first conclusion: all formants moved significantly towards the typical values given in Calliope (1989). Two exceptions must be noted: (1) the F2 of both vowels did not change for AG, and (2) the F1 of /œ/ shifted excessively for AVG. The second important conclusion is that AVG changed significantly more than AV in all cases, except for the F1 of /ø/ (t(338)=0.7, n.s.). Note that the AV and AVG groups, though randomly selected, were not homogeneous, possibly due to the differences among Mandarin varieties. Indeed, the F1 means of /ø/ and /œ/ for the four north-eastern Chinese (AVG) were respectively 637 Hz and 662 Hz, whereas for the rest of the learners the values were only 548 Hz and 561 Hz. Note also that the four north-eastern Chinese's F1 of /œ/ shifted from 662 Hz to 744 Hz, whereas the other three leaners of AVG shifted from 603 to 664 Hz (to compare with 647 Hz in Calliope (1989)). This could explain the aforementioned excessive shift for the mean of F1 measured over all the learners of AVG.

4. Discussion and conclusions

Some other aspects of these results are worth discussing. The F2 increased for both vowels for AVG but not for AV. This might be attributed to the fact that the vision of the tongue may have prompted the learners to centralize their articulation, as shown by Figure 6. Note also that the F1 difference between /ø/ and /œ/ after training was 62 Hz for AG and 117 Hz for AVG, to compare with 178 Hz in the Calliope (1989) data: an unpaired t-test showed that the difference of these two F1 differences was highly significant. (t(338)=-4.56, p<0.001). This indicates some better learning effect for the AVG group, though more progress remains to be done.

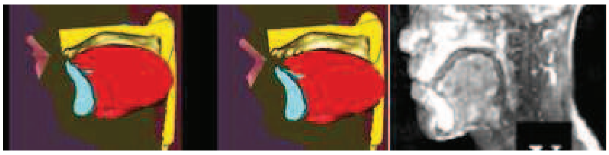

Figure 6: *Comparison of /ø/ (left), /œ/ (middle) and /ɤ/ (right, from Wang et al. (2008)).*

Besides, the perception test showed that the correct response percentage for /œ/ rose from 61% (42-76%) to only 68% (49-82%) for AG and from 50% (32-67%) to 86% (68-94%) for AVG (the 95% confidence interval of correct response percentage was defined as the Wilson score interval (Wilson (1927)). This result contributes to validate the advantage of the audiovisual approach.

Some of the learners reported the wish to have a frontal view of lips instead of the sagittal view. If we consider the decrease of F3 as a sign of lip rounding for front vowels (Harrington (2010)), then no such strong sign was observed in AVG. The F3 of /ø/ and /œ/ changed respectively from 3108 Hz to 3100 Hz and from 3112 Hz to 3199 Hz (to compare with 2581 Hz and 2753 Hz in Calliope (1989)). Since lip rounding induces also the lowering of F1, it may also account for the fact that the F1s of /ø/ remained higher than the reference value after training. In the future, it might also be interesting to extend this type of approach to provide learners with a display of their own articulation (*cf. e.g.* Engwall, 2012; Ben Youssef *et al.*, 2011). In practice, the presence of an instructor could certainly help a lot. Last but not least, the learners in the AVG group being generally more motivated by the game side of the ATH animation, more studies are needed to assess the benefit of the use of articulatory information per se.

5. Acknowledgements

We would like to thank Cédric Gendrot for the *Praat* scripts used in this study (http://gendrot.ilpga.fr/scripts.htm) and Gérard Bailly for the idea of using speech shadowing as a supplementary learning strategy. This work was partially supported by the Centre for Chemistry, Life and Healthcare Sciences and Bioengineering of University Joseph Fourier, Grenoble, France (Vizart3D project).

6. References

Badin, P., Elisei, F., Bailly, G. & Tarabalka, Y. (2008). An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data. In Vth Conference on Articulated Motion and Deformable Objects (AMDO 2008, LNCS 5098) (F.J. Perales & R.B. Fisher, editors), pp. 132–143. Berlin, Heidelberg, Germany: Springer Verlag.

Ben Youssef, A., Hueber, T., Badin, P. & Bailly, G. (2011). Toward a multi-speaker visual articulatory feedback system. In Interspeech 2011 (12th Annual Conference of the International Speech Communication Association), pp. 589-592. Florence, Italy.

Best, C.T. (1991). The emergence of native-language phonological influences in infants: A perceptual assimilation model. Haskins Laboratories Status Report on Speech Research, 107-108, 1-30.

Boersma, P. & Weenink, D. (2005). Praat: doing phonetics by computer (Version 5.3.30) [Computer program]. Retrieved Oct 06, 2012, from http://www.praat.org/.

Bressmann, T., Heng, C.-L. & Irish, J.C. (2005). Applications of 2D and 3D ultrasound imaging in speech-language pathology. Journal of Speech Language Pathology and Audiology, 29(4), 158-168.

Calliope. (1989). La parole et son traitement automatique. Paris, Milan, Barcelone, Mexico: Masson.

Engwall, O. (2012). Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. Computer Assisted Language Learning, 25(1), 37-64.

Harrington, J. (2010). Acoustic phonetics. In The Handbook of Phonetic Sciences (W.J. Hardcastle, J. Laver & F.E. Gibbon, editors): Wiley-Blackwell.

Kuhl, P.K. & Meltzoff, A. (1982). The bimodal perception of speech in infancy. Science, 218(4577), 1138-1141.

Massaro, D.W., Bigler, S., Chen, T., Perlman, M. & Ouni, S. (2008). Pronunciation training: the role of eye and ear. In Proceedings of Interspeech 2008, pp. 2623-2626. Brisbane, Australia.

Massaro, D.W. & Light, J. (2003). Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/. In Eurospeech 2003, pp. 2249-2252. Geneva, Switzerland.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746-4748.

Schmidt, A.-M. & Beamer, J. (1998). Electropalatography treatment for training Thai speakers of English. Clinical Linguistics & Phonetics, 12(5), 389-403.

Shockley, K., Sabadini, L. & Fowler, C.A. (2004). Imitation in shadowing words. Percept Psychophys, 66(3), 422-429.

Stylianou, Y. (1996). Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. Paris.

Wang, G., Lu, X. & Dang, J. (2008). A study of Mandarin Chinese using X-Ray and MRI. Journal of Chinese Phonetics, 2, 51-58.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158), 209-212.