MULTIPHONIA: a MULTImodal database of PHONetics teaching methods in classroom InterActions.

Charlotte Alazard, Corine Astésano, Michel Billières
U.R.I Octogone-Lordat, EA 4156, Université Toulouse II
5 allées Antonio-Machado, F-31058 Toulouse Cedex 1
E-mail: alazard@univ-tlse2.fr, astesano@univ-tlse2.fr, billieres@univ-tlse2.fr

Abstract

The *Multiphonia Corpus* consists of audio-video classroom recordings comparing two methods of phonetic correction (the 'traditional' articulatory method, and the Verbo-Tonal Method) This database was created not only to remedy the crucial lack of information and pedagogical resources on teaching pronunciation but also to test the benefit of VTM on Second Language pronunciation. The VTM method emphasizes the role of prosody cues as vectors of second language acquisition of the phonemic system. This method also provides various and unusual procedures including facilitating gestures in order to work on spotting and assimilating the target language prosodic system (rhythm, accentuation, intonation). In doing so, speech rhythm is apprehended in correlation with body/gestural rhythm. The student is thus encouraged to associate gestures activating the motor memory at play during the repetition of target words or phrases. In turn, pedagogical gestures have an impact on second language lexical items' recollection (Allen, 1995; Tellier, 2008). Ultimately, this large corpus (96 hours of class sessions' recordings) will be made available to the scientific community, with several layers of annotations available for the study of segmental, prosodic and gestural aspects of L2 speech.

Keywords: database, multimodality, phonetic correction, second language acquisition.

1. Introduction

Prosody and multimodality are not only the key to language acquisition but also necessary and irrepressible in everyday communication (Di Cristo, 2004; Kendon, 2004; Mac Neill, 2005). The number of studies in L2 prosody is however rather limited compared to the amount of work carried out on L2 segmental aspects.

The lack of database in didactics, particularly in French as a Second Language (hereafter FSL), could explain the lack of experimental researches in this field. Even if more focus is now put on communication in foreign language teaching methods, some main aspects of oral communication, such as phonetics and prosody, remain remarkably left out

The aim of our research is thus to provide a multimodal database of real classroom interactions in FSL. It is also aimed at confronting theoretical predictions and real class situations, in order to favour phonetics teaching in foreign language courses.

It is expected by the *Common European Framework of Reference for Languages* (hereafter, *CEFRL*) that advanced level students (level B) should have '*a clear and natural intonation'* and read with fluency. No mention is made of pronunciation training and oral skills' mastering. It is as if good fluency and prosody in L2 came naturally to advanced L2 students. Experience in teaching L2 however contradicts this view, insofar as advanced students still transfer the prosodic characteristics of their L1 onto the L2, in both unscripted and read speech. In other words, foreign accent remains persistent at an advanced level.

This apparent contradiction has various roots: first of all, even if recent researches have shown the importance of phonetic training in the improvement of speaking fluency in spontaneous and read speaking skills (Freed, 1995; Freed & al, 2004 ; Alazard & al, 2009, 2010), the idea that pronunciation will improve naturally thanks to mere repeated contacts with the foreign language is persistent. Secondly, in spite of the recognized role of prosody in both first and second language acquisition (Di Cristo, 2004), L2 traditional teaching methods - in the rare cases where pronunciation is taken into account - focus exclusively on the segmental level. It is worth noticing at this point that L2 teachers very rarely perform phonetic correction in their classroom activities, due to their lack of expertise in this discipline. The emphasis is rather on grammatical and lexical aspects of the target language; very rare moments are dedicated in L2 classes to phonetics and segmental pronunciation correction. Finally, and because of this lack of phonetics teaching practice in the classroom, phonetics correction methods have never been experimentally tested or validated.

In order to confront this last point, we propose to question and to experimentally test two different pronunciation teaching methods - the Articulatory Method (hereafter AM) and the Verbo-Tonal Method (hereafter VTM) - and to make these teaching-methods recordings available for researchers through our database.

According to AM, by far the most widespread method, a good production implies the metalinguistic knowledge of how we articulate sounds. Thus, the teacher will provide an articulatory description of the different sounds of the foreign language, then prompt the student to repeat the correct articulatory gestures in order to produce the target sound. For example, to produce [u] the teacher will tell the student to place the tongue at the back of the mouth and to round the lips, in opposition to the fronted [i] for which the lips should be stretched. In this method, the emphasis will first be put on the production and repetition of isolated sounds, then isolated words containing the target sound and finally sentences. There will be no real focus on prosodic parameters such as rhythm and intonation. The AM thus focuses on explicit learning of phonetic articulatory gestures.

The VTM, on the contrary, uses the prosodic structure of the target language as the 'shell' for pronunciation skills' improvement. More specifically, the rhythmic pattern of the target language is used to bring to light the phonetic specificities of the target language. The teacher first helps the learners familiarize themselves with the prosodic structure of the target language through the repetition of prosodic patterns using logatoms (/dadada/) or the use of facilitating gestures (for example rising hand movement for salient syllables). In a second phase, the prosodic structure is used to facilitate phoneme perception and re-production, on the basis that there is a phonological loop between the production and the perception of phonetic features. (Borrell, 1996) For example, if the learner darkens the timbre of a target phoneme, the teacher will pronounce the phoneme in a prosodically brightening context (accented syllable) and have the learner repeat it in the same context. Namely, a facilitating production context will help the learner perceive the proper phonemic features of the target language and thus help them correctly re-produce these features in any other prosodic contexts (Billières, 2005). The VTM thus focuses on non-explicit prosodic learning. Despite extremely positive results both in didactics and speech therapy, teachers are wary of this method as it implies a different teaching approach and an expertise in phonetics and prosody. Furthermore, because this method remains confidential to a small group of international experts, its validity remains to be demonstrated to a larger audience in order to, one day, be included in comprehensive L2 teaching methods.

The originality of our database is thus to record and compare for the first time these two different methods in an ecological classroom situation, and to propose an enrichment of this database for future L2 researches.

2. Collection of the database

The database consists of a longitudinal recording of classroom teaching of phonetics over eight weeks, with twenty participants, all English Speakers (15 female; mean age: 32; age range: 20-60). An oral interview allowed us to evaluate their level in French according to the CEFRL: ten of the participants were judged to have an elementary level in French (level A) and ten were judged to have an advanced level in French (level B).

The participants were equally divided into four groups: two groups per method according to their level. Each group received two pronunciation trainings per week - lasting one hour and a half each - for eight weeks.

Both methods were taught by the same teacher – the first author - and recorded in the same experimental conditions. All class sessions were recorded in the professional audio-visual recording studio of the Direction of Information Technology and Communication for Teaching service (DTICE) of the University of Toulouse II.

Depending on the methodological approach, the classroom stage was reorganized as follows: for the AM classes, the participants were sitting around an U-shaped table while the teacher was explaining articulatory features, using oro-facial and vocal tract gestures or diagrams (Figure1). The participants were then asked to repeat one by one the target sound, presented in isolation or in single words. During the second part of the class, they would listen to a dialogue or an authentic document, before answering a few questions on the audio documents.

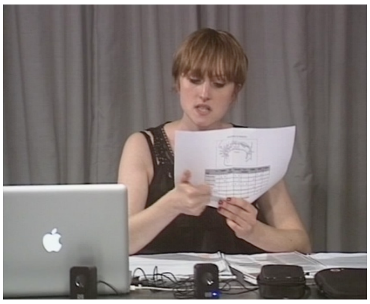

Figure 1: Example of an AM classroom layout. Participants are sitting at a table around the teacher who uses diagrams and meta-discourse to describe the articulation of the different phonemes.

For the VTM classes, the participants were sitting around the teacher, with no table, while the teacher used prosody and multimodality to help participants perceive the target sound and the prosodic features of the target language (Figure2). Participants were free to reproduce or not the hands movement (following the variation of intonation) made by the teacher, while repeating the sentences or the short dialogues. The second part of the course would be identical in both methods.

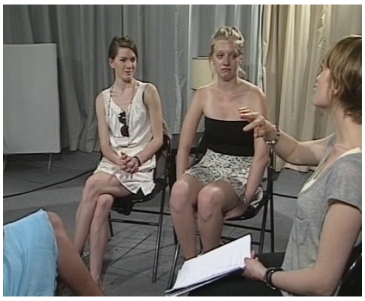

Figure 2: Example of a VTM classroom layout, with no table between the participants and emphasis on body language.

These different set-ups imply a specific technical organization of the classroom stage. The studio combines three video cameras (BVP50 Sony) – an overall shot of the teacher, two overall shots of the students – and six microphones (half-track AKG hanging down from the ceiling) – one microphone for the teacher and the other 5 for the students.

The classes were recorded in an artificial lightening: one spotlight of 800 w for the reserve angle on the teacher, 4 spotlight of 800w for the students and one spotlight for the backlighting.

The stagte controler used a Panasonic video mixer, a Tascam eight tracks audio mixer, a TL 12 Coyllins light controler and dvcam 4/3 DSR 45 Sony of 184mn for each class.

We used dvcams for the production rushes and avid media-composer for the post-production. The data were transferred onto dvd for the trimming. In order to be used on the Internet, the masters were encoded with Adobe cs5.

The database is constituted of ninety-six hours (3h/week*4groups*8 weeks) of classroom recording.

3. Enrichment of the database

This multimodal database constitutes an important resource for Second Language Acquisition's (hereafter SLA) researchers.

Hence, MULTIPHONIA will be enriched at many different levels, to allow for segmental, prosodic, morphosyntaxic, syntaxic, lexical and gestural analyses of L2 speech. The different levels of automatic annotation will be done on short excerpts of the database at first, and confronted to manual annotations by three experts, before extending the procedure to larger parts of the corpus (see section 4- below).

The automatic annotation of such a corpus, consisting of interactional speech in a classroom environment and recorded with several multi-directional microphones on a single sound-track, represents an interesting challenge for the automatic tools' developers to test for the portability of their tools to more challenging speech corpora.

3.1 Transcription (cf. Bertrand et al, 2008).

3.1.1 Segmentation in Interpausal Units

Before any annotation, some significant audio extracts of the recording will be automatically segmented in Interpausal Unit (hereafter IPU). IPU are constituted of blocks of speech separated by 200 ms silent pauses. The IPU segmentation has been commonly used for large corpora as it then facilitates sound and transcription alignment

3.1.2 Enriched OrthographicTranscription (EOT)

The advantage of the EOT is to provide an orthographic transcription as well as specifying natural speech production phenomena such as pauses, false starts or repetitions. According to the EOT transcription convention, the transcribers will be asked to annotate silent pauses, filled pauses, elisions, false starts, word truncation, liaisons (absence of a standard liaison, presence of an unusual liason), assimilation phenomena and specific phenomena, such as, in our case, deviant pronunciations or code switching from L2 to L1 (those phenomena will be labelled as 'interlanguage phenomena') (see Bertrand et al, 2008 for transcription conventions).

Two transcriptions will then be automatically generated from the EOT.

First, a standard orthographic transcription from which the orthographic tokens are extracted for semantics, syntax or discourse analyses (see Blache et al. 2009 for example)

Second, a selective transcription from which the phonetic tokens are extracted for grapheme-phoneme conversion.

3.2 Phonetic annotations

3.2.1 Phonetization

The phonetic annotations will be done with the Speech Phonetization Alignment and Syllabification (SPPAS) tool (Bigi and Hirst, 2012) on the significant extracts. The aim of this tool is to provide automatic utterance, words syllables and phonemes segmentations annotations from a speech recording and its transcription.

SPPAS produces a phonetic transcription based on a phonetic dictionary. The program offers the possibility to select (automatically or manually) among all the phonemics variants that are proposed.

3.2.2 Alignment

The phonetic alignment will consist of an automatic temporal matching between a speech utterance and its phonetic representation. The alignment will be done in the frame of each IUP, to maximize the alignment.

In order to evaluate the errors rate of the aligner the automatic alignment will be compare with the manual alignment of two experts.

3.2.2 Syllabification

The syllabication will be done according to two main principles: (1) a syllable contains only one vowel and (2) a pause signals a syllable boundary, as previously described in Bigi et al (2010).

3.2.3 Disfluencies

Disfluencies can be prosodic (lengthening, silent and filled pauses, mean length of speech runs, etc.) or lexicalized (word or phrase truncation, repetitions, *etc.*) . The prosodic ruptures of the speech flow will be annotated according to the quantitative measures detailled in Kormos (2006) while for the lexicalized disfluencies we will annotate three distinctive parts of the disfluencie: the Reparandum (what precedes the interruption point), the Break Interval (optional event between the Reparandum and the Reparans) and the Reparans (the part following the break) (Blanche-Benveniste, 1987).

3.3 Prosodic annotations

The annotation of prosody is very complex and cannot yet be done automatically. Bertrand *et al.* (2007) & Blache *et al.* (2009) propose to distinguish different levels of relevant prosodic information to annotate in order to account for discourse prosodic structuring:

- Prosodic phrasing, with two hierarchical levels consisting of the Intonational Phrase (IP) and the Accentual Phrase (AP). These units are marked by a Final Accent and a (potential) Initial Accent.

Intonation contours, which can be characterized as follows: minor contour (m), major continuative rising contour (RMC), major list rising contour (RL), major falling contour (F), major terminal rising contour (RT), major questioning rising contour (RQ), rising-falling contour (RF1), falling from the penultième contour (RF2), no melodic variation (f1)

3.4 Gestural annotations

The annotations of gestures can be done manually with the ELAN or ANVIL softwares, according to Mac Neill's typology (MacNeill, 2005): metaphoric (gestures representing an abstract idea), iconic (gestures representing an action or concrete object),, deictic (pointing gestures) and beating (gestures accompanying rhythm) gestures, are differentiated.

For the purpose of our own research on L2 (see section 4- below), significant extracts of the teacher's and students' gestures in the two teaching classroom environments will be annotated and compared. .We will annotate interactive gestures (gestures addressed to the interlocutor in order to manage the interaction) and aborted gestures (half-made gestures), quite typical of L2 speech interaction (Tellier and Stam, 2010). Head movements, body position, gaze directions and facial expression, will be encoded as well as hand movements.

4. First steps towards the exploitation of the MULTIPHONIA database

4.1 Prosody supporting L2 segmental perception and production

At the segmental level, we are planning on measuring the influence of prosody on the acquisition of the phonemic features. The VTM hypothesizes indeed that the prosodic structure will facilitate phoneme perception and thus will also facilitate phoneme production.

In order to test this hypothesis, we will measure formants' repartition of L2 vowels in different repetition contexts and at different stages of the training. We will extract significant audio sequences that will be annotated manually by three experts using the coding elaborated in Bertrand *et al.* (2008) with the PRAAT software (Boersma, P. & Weenink, D., 2005), as described in section 3. The annotation will then be automatically phonetized and aligned with the audio extracts using the aligner elaborated by Bigi & Hirst (2012), to facilitate automatic vowel detection.

4.2 Gestural impact on prosodic characteristics' learning

Two different phonetics teaching methods imply two different uses of pedagogical gestures. The AM puts the accent on the gestures of the articulators only, implying central or peripheral gestural spaces, whereas the VTM focuses on prosodic guiding gestures, implying peripheral and upper gestural spaces (Figure 3).

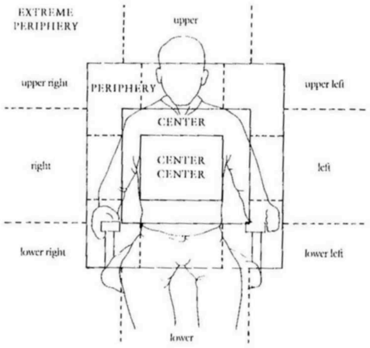

Figure 3: The gestural space (McNeil, 1992)

More specifically, with the VTM, the teacher helps apprehending speech rhythm in correlation with body/gestural rhythm, as well as perceiving prosodically salient syllables and segmenting the speech flow (Billières, 2002). The teacher thus helps developing the perception of 'rhythmic phrases' through the perception of rhythmic prominences (Initial Accent (IA) and Final Accent (FA)). Rhythmic saliences are seen as anchor points for the speech flow segmentation into smaller units (3 to 4 syllables), necessary for ultimate processing in the working memory.

In parallel to the pedagogic work on the verbal continuum, the student is thus encouraged to associate gestures activating the motor memory at play during the repetition of target words or phrases. Indeed, research in cognitive psychology has demonstrated the impact of the motor modality on sentence recall (Cohen & Otterbein, 1992). By the same token, Second Language Acquisition studies have shown that pedagogical gestures have an impact on second language lexical items' recollection (Allen, 1995; Tellier, 2008).

Because prosodic cues help access the lexicon and segment the speech flow in the native language but are poorly explored in a second language (Snijders *et al.*, 2007), we wish to experimentally demonstrate that pedagogical gestures have a facilitating impact on the reproduction and memorization of relevant non native prosodic cues. This will be achieved through the systematic analysis of SL French words or phrases repeated by the English learners with the use of both gestural and proper French accentual patterns throughout the eight weeks' classes in the VTM group. The basic accentual pattern for French in the Accentual Phrase (AP) is Initial Accent and Final Accent (hereafter /IA-FA/; see Di Cristo, 2000, and Jun & Fougeron, 2002, for a description). We hypothesize that gestures will help anchor the /IA-FA/ prosodic pattern, and help correct accurate realization of the target language accentual and segmental systems.

The annotation of gestures – specifically beating metaphorics and interactive gestures - using the coding elaborated in Blache *et al.* (2010) will be done with the ELAN software.

5. Conclusion

This paper presents a MULTImodal database of PHONetics teaching methods in classroom InterActions (MULTIPHONIA), consisting of 96 hours of audio-video classroom recordings of two methods of phonetic correction (the 'traditional' articulatory method, and the Verbo-Tonal Method).

If this database primarily constitutes a rich supply for pedagogical resources, it also provides multimodal resources for SLA researchers interested in various aspects of L2 learning, via the annotation of segmental, prosodic, morphosyntaxic, syntaxic, lexical and gestural levels of L2 interactional speech.

Annotated extracts of this MULTIPHONIA database are going to be shortly available on line. (http://crdo.up.univ-aix.fr/voir_depot.php?lang=fr&id=000780).

6. Acknowledgements

The authors wish to thank the DTICE and especially Mr Bruno Bastard, the audio-visual technician of the DTICE, for his help in the realization of this database

7. References

Alazard, C., Astésano, C., and Billières. M. (2009). "Rôle de la prosodie dans la structuration du discours - Proposition d'une méthodologie d'enseignement de l'oral vers l'écrit en Français Langue Etrangère-". Proceedins of the *Interfance Discours and Prosodie conference 2009* (Paris, France). http://makino.linguist.jussieu.fr/id09/actes_fr.html

Alazard, C., Astésano, C., and Billières. M. (2010). "The Implicit Prosody Hypothesis applied to Foreign Language Learning: From oral abilities to reading skills". Proceedings of the *5th Speech Prosody 2010* (Chicago, USA). http://www.speechprosody2010.illinois.edu/papers/100648.pdf

Allen, L. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal* 79: 521-529.

Bigi, B., and Hirst, D. (2012). "*Speech Phonetization Alignment and Syllabification: A tool for the automatic analysis of Speech Prosody"*. Proceedings of the *6th Speech Prosody 2012* (Shanghai, China).

Billières, M. (1993). 'Théorie et pratique du rythme parolier en phonétique corrective'. *Cahiers du Centre Interdisciplinaire des Sciences du Langage* 9: 3-32.

Billières, M. (2002). Le corps en phonétique corrective. In R. Renard (Ed.), *Apprentissage d'une langue étrangère/seconde. La Phonétique verbo-tonale.* Bruxelles: De Boeck Université, pp. 37-70.

Billières, M. (2005). Les pratiques du verbo-tonal. Retour aux sources. Berré, M (Eds), *Linguistique de la parole et apprentissage des langues. Questions autour de la méthode verbo-tonale de P. Guberina*. Centre International de Phonétique Appliquée, Mons, pp. 67-87.

Blache P., Bertrand R. and Ferré, G. (2009). Creating and Exploiting Multimodal Annotated Corpora: The ToMA. In M. Kipp et al. (Eds): Multimodal Corpora, LNAI 5509, pp. 38-53.

Blache P., Bertrand R., Bigi B., Bruno E., Cela E., Espesser R., Ferré G., Guardiola M., Hirst D., Magro E. -P., Martin J. -C., Meunier C., Morel M. -A., Murisasco E., Nesterenko I., Nocéra P., Pallaud B., Prévot L., Priego-Valverde B., Seinturier J., Tan N., Tellier M. and Rauzy, S. (2010). 'Multimodal Annotation of Conversational Data'. *Proceedings of Linguistic Annotation Workshop 2010* (Uppsala, Sweden), pp. 186-191.

Blanche-Benveniste C. (1987). "Syntaxe, choix du lexique et lieu de bafouillage". In *DRLAV* 36-37

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B. and S. Rauzy. (2008). Le CID – Corpus of Interactional Data. *TAL*, 49 (3), pp105-134.

Boersma, P. and Weenink, D. (2005). Praat: doing phonetics by computer, http://www.praat.org

Borrell, A. (1996). Parallèle entre perception et production: complexité du lien entre reconnaissance et production des unités phonético-phonologique. *La linguistique* 32 (2)

Cohen, R. L. and Otterbein, N. (1992). The mnemonic Effect of speech Gestures: Pantomimic and Non-Pantomimic Gestures compared. *European Journal of Cognitive Psychology*, *4* (2), pp 113-139.

Di Cristo, A. (2000). Vers une modélisation de l'accentuation en français. Deuxième partie : le modèle. *Journal of French Language Studies,* 10. pp. 27-44.

Di Cristo, A. (2004). La prosodie au Carrefour de la phonétique, de la phonologie et de l'articulation formes-fonctions. *Travaux Interdisciplinaires du Laboratoire Parole & Langage* 23, pp. 67-211.

Freed, B. F. (1995). What Makes Us Think that Students Who Study Abroad Become Fluent? In B. F. Freed (Eds), *Second Language Acquisition in a Study Abroad Context*, pp. 123-145.

Freed, B. F., Segalowitz, N., and Dewey, D. (2004). Contexts of learning and second language fluency in French: Comparing regular classrooms, study abroad, and intensive domestic programs. *Second Language Acquisition, 26*, pp. 275-301.

Jun, S. -A. and Fougeron C. (2002). Realizations of accentual phrase in French intonation. *Probus* 14, pp147-172.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kormos, J. (2006). *Speech production and Second Language Acquisition*. New York, London: Routledge.

McNeill, David (2005). *Gesture and Thought.* Chicago: Chicago University Press.

Snijders, T. M., Kooijman, V., Cutler, A., and Hagoort, P. (2007). Neurophysiological evidence of delayed segmentation in a foreign language. *Brain Research, 1178,* pp. 106-113.

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. In Gullberg, M., & de

Bot, K. (Eds.) *Special issue Gestures in language development*. *Gesture*, 8(2), pp. 219-235.

Tellier, M and Stam, G. (2010). Découvrir le pouvoir de ses mains: La gestuelle des futurs enseignants de langue. Proceedings of the conference Specificités et diversité des interaction didactiques: disciplines, finalités, contextes, 2010 (Lyon, France)