

MSc Natural Language Processing

2020/2021

UE805 EC1 Supervised Project

Evaluation of a Multilingual Animated Speech System

Students:

Juliana DE FERRAN

Sonita TE

Stephanie MONTEIRO

Supervisor:

Slim OUNI

Reviewer:

Denis JOUVET

June 2021

Contents

List of Tables

List of Figures

Listings

1 Acknowledgements

We would like to thank our supervisor Slim Ouni for letting us work with him and his team. We would also like to deeply thank Théo Biasutto-Lervat for his help with understanding the softwares (both the one for animation and the one used to create the surveys), debugging the problems we discovered in the established software, and taking the time to support us this semester.

We also thank Loria and the University of Lorraine for giving us the opportunity to take part in this project. We would also like to thank our colleagues and professors who stepped up to help us find as many participants as possible.

2 Introduction

In the field of audiovisual speech, animated virtual 3D faces of human speakers synced with audio (called talking heads) have been developed to model audiovisual speech communication and therefore study its mechanism. This paper is part of a project led by researcher Slim Ouni which consists of creating and evaluating a multilingual talking head. This study originates from the fact that researchers such as Taylor et al. posit that it is possible to have a high quality audiovisual speech animation using a single system that animates any language. Our purpose is to use a previously existing animation software developed by Biasutto-Lervat et al. at Loria laboratory in Lorraine (France) in order to simulate a universal system.

Each of the three initial systems animated the talking head's articulation of one single language: French, English, and German, and each worked with the set of rules of that language. Our procedure to make the system multilingual was to map the phonemes between the languages of study and then to use that to animate Language A with the original audio of that language but with the articulation of another one (Language B) using the existing platform.

For the evaluation, videos showing the articulation of the talking head producing sentences of any given language were generated using the monolingual system for the reference language and multilingual one for the others. Then, these videos were evaluated by native speakers and, finally, comes the analysis of the results.

Our goal in this evaluation was to test the thesis that a monolingual system such as the one originally available at Loria is more accurate in its representation than a multilingual system that could

animate any language regardless of its constraints as we have demonstrated in our previous report by studying the phenomena of coarticulation in the scientific literature (Daniloff & Hammarberg, 1973). Indeed, since the fact that neighboring sounds affect each other in the articulatory process, the quality of the visual information transmitted is also influenced. This process may differ in each language and as such is one of the most important points to consider when establishing whether a multilingual system is viable.

Therefore, if the animation rendered using a mapping is perceived to be better, then coarticulation does not have an important role in this context, which means that we can create an universal system that would work for any language instead of a unique system for each one. On the other hand, if the results show that the videos created with the monolingual system are clearly judged to be better by native speakers, a single language system remains the best one and proves that coarticulation is significant.

In this report, we will first describe the phonesets we dealt with before discussing our mapping methodology. Then, we will tackle the generation of videos produced with the interface of our initial system and a created mapping code. Finally, we will look at how we established the questionnaires for the three languages that allowed native speakers to rate the articulation of the animated face, our criteria for choosing the participants, and how we parsed our data to establish a ranking of the systems.

3 Phonesets

We dealt with three different phonesets:

- The English phoneset (see table 1 in Appendices) with 24 consonants and 15 vowels, i.e. 39 phones in total.
- The French phoneset (see table 2 in Appendices) with 20 consonants and 10 vowels, i.e. 30 phones in total.
- The German phoneset (see table 3 in Appendices) with 22 consonants and 20 vowels, i.e. 42 phones in total.

We incorporated the correspondences with the International Phonetic Alphabet (IPA) in order to have a reference because the three systems do not use the same symbols.

On one hand, the English system symbols correspond to the ARPABET. Each phoneme of General American English is represented with a distinct sequence of ASCII characters. In this case, the system is in 2-letters notation.

A symbol that we encountered that was not from a phonetic alphabet is the one for silence: "SIL" in English, and "sil" in French and German, which also had to be mapped.

On the other hand, the French and German system symbols come from the X-SAMPA alphabet. This is an encoding of the IPA into ASCII characters that are readable and printable by a machine (see figure 8 in Appendices). In both phonesets, some symbols are not X-SAMPA ones or are not generally used in the language concerned. From a linguistic point of view, it is not consistent to, on the one hand, take the elements of a certain alphabet and, on the other hand, to modify them or add external ones when the original X-SAMPA contains all the phonemes that exist in a given language.

We had to work with them because we couldn't modify the existing software that used these phonesets. In the two paragraphs below, we have listed these strange components present.

French In French, the nasal vowels of the phones we study are 'an', 'in', 'on' while in X-SAMPA, they are respectively 'a~', 'U~/' and 'o~'.

German In German, there are two exceptions: 'r' and 'a~'. The 'r' symbol of X-SAMPA corresponds to /r/ in IPA but in standard German, /ʁ/ is used. In the German X-SAMPA symbol set, the phoneme 'a~' is not usually present.

We noticed other inconsistencies such as different symbols in our phonesets that refer to the same IPA symbol. For example, the 'R' in French and the 'r' in German correspond to the symbol IPA /ʁ/ while in English we have the same symbol 'R' as in French but it matches with /ɹ/.

From these phonesets, we did what we call a mapping. This process is clearly explained in the next section.

4 Mapping

In this part, we describe the mapping step. In our context, the method consists of associating a phone of Language A to one that most closely resembles it in Language B in the matter of articulatory

characteristics. We specify "articulatory" because we focus only on the visual representation of each sound (especially at the level of the lips) and not on the acoustics.

The first reason for this concerns the material used and the second, the purpose of our experiment. Indeed, as we work with a talking head, the most important information comes from the lips. In addition, in order to create a multilingual system, the animation will be composed of the audio of the base Language A associated with the articulation produced by the system of Language B, and as such the audio file will be the original one generated by the system of that language and will not be impacted by the mapping.

A study conducted by McGurk & MacDonald in 1976 demonstrated the importance of the visual information to the understanding of the listener. The McGurk effect proves that when visual input leads the listener astray, speech intelligibility is impacted. For this, the researchers carried out a perception experiment. They created a video associating the audio sequence [baba] with the visual input of the sequence [gaga]. Faced with this, the participants reported hearing [dada]. Therefore, we had to choose the closest counterparts in Language B in order to ensure the intelligibility of Language A.

Thus, in the next part, we will explain our steps to establish an accurate mapping. As we dealt with three languages, we have done 6 mappings altogether: English to French, English to German, French to English, French to German, German to English, and German to French.

4.1 Articulation

First of all, we must define the concept of articulation, based on the book by Hannahs & Davenport [3]. Articulation corresponds to the movement of active articulators (also often called lower articulators, such as the bottom lip, the tongue, the glottis) and their interaction with passive ones, also often called articulators (such as the upper lip, the upper teeth and the alveolar, palatal, velar, and uvular region), allowing the air expelled by the lungs to flow through a more or less restricted airway and to reverberate in the oral and/or nasal cavities in different ways, therefore producing different sounds.

Our three languages distinguish two classes of sounds depending on whether the air flow is disturbed or not: vowels and consonants.

On the one hand, vowels are characterized by a free passage of air flow through the vocal tract. There are four classification criteria:

1. The manner of articulation, which refers to the degree of opening of the jaw. The terms 'close', 'open' and their derivatives are used in relation to the aperture of the mouth.
2. The place of articulation, which corresponds to the approach point of the active articulator in the vocal tract. Therefore, the horizontal position of the tongue is defined with the terms 'front', 'central', 'back' and their derivatives.
3. Resonance, which allows to distinguish whether the air passes in the nasal cavities or not.
4. Labiality, which indicates rounding or stretching of the lips.

Parallel to the vowels, there is also a class called diphthongs, which corresponds to a type of vowel whose point of articulation and timbre vary during its emission. We find them in particular in English and German.

On the other hand, consonants are articulated with complete or partial closure of the vocal tract. There are also four classification criteria:

1. The manner of articulation, which refers to how the airflow is released from the vocal tract.
2. The place of articulation, which corresponds to where the obstruction takes place in the vocal tract.
3. Resonance, as defined above.
4. Voicing, which indicates whether the vocal cords vibrate.

Regarding consonants, there are several manners of articulation, we present here those used in at least one of our languages of study:

- Plosives: air flow is completely stopped and then released with a slight audible explosion. E.g. /p/, /b/, /t/, /d/, /g/, /k/...
- Nasal: as with plosives, there is complete closure, but the velum is lowered so the air passes into the nasal cavity. E/g. /m/, /n/, /ŋ/, /ɲ/...

- Fricatives: two articulators (one active and the other passive) are brought and held sufficiently close together for the escaping airstream to cause local air disturbance and friction, which is audible. E.g. /f/, /v/, /s/, /z/, /ʃ/, /ʒ/...
- Approximant: like fricatives, but without causing audible friction. Note that an approximant is said to be lateral when the air flow passes through the sides. E.g. /j/, /l/ ...
- Affricate: This kind of sound consists of a plosive and then a fricative, which have the same place of articulation. E.g. /tʃ/, /dʒ/ ...

There are also particular sounds called semi-vowels or semi-consonants because they are derived from respective vowels and also behave like consonants. For example, French has three semi-consonants: /w/, /h/, and /j/. Each are respectively related to the vowels: /u/, /y/ and /i/, and they are also classified as approximants.

In our case, we are especially interested in the parameters that can be seen, that is to say: the manner of articulation and labiality for vowels and the manner of articulation and the point of articulation for consonants.

First, we begin the mapping by pairing a phone in Language A to its counterpart in Language B when there is total correspondence in the IPA. Note that for vowels, when there is an elongation for a phoneme in Language A and the same phoneme exists in Language B but without elongation, we use it as a counterpart. So, we do not really take into account the elongation because it remains the same articulation although it is prolonged and the initial segmentation takes into account the duration of a phone in a sentence. The phones considered as similar are colored in gray in the mapping tables (see tables 4, 5 and 6 in Appendices).

Following that, we also had to deal with Language A phones that have no direct equivalent in Language B and therefore select the phone that came closest to it on an articulatory level. To help us establish this kind of correspondence, we were inspired by the IPA chart with SAMPA correspondence to create an articulatory classification of the three phonesets of the system (see figures 10, 9 and 11 in Appendices). In the following subsections, we go through them in order to explain our reasoning.

4.2 Mapping: English to ...

4.2.1 French

Vowels

- 'IH', 'UH': Both vowels are near-close, we opted for 'i' and 'u' in French because they are two close vowels.
- 'ER': The only phoneme that is close to it is 'swa' in French because this is also a central vowel and they are close regarding the aperture of the mouth.
- 'AE': 'a' is the phoneme of the French system that is the closest because it is located in the same area of the classification table.
- 'EY', 'OY', 'OW', 'AY', 'AW': For diphthongs, we kept the first part, i.e. the first vowel, as a counterpart. We obtained, respectively, 'e', 'o', 'o', 'a', and 'a' in French.

Consonants

- 'TH', 'DH': As these phones are fricative dental, we decided to use respectively 'f' and 'v' which are fricative labiodental as counterparts for French.
- 'HH': For this phone, we chose the vowel 'swa' in French because they both have a neutral position of the lips.
- 'NG': This phone is a nasal velar consonant and as we have an oral (plosive) velar one in French (i.e. 'G'), we selected it.
- 'R': This sound is in free variation with the French 'R', that is to say that when they appear in the same environment, they are interchangeable for a native speaker who will understand the meaning regardless.
- 'CH', 'JH': For affricates, we affected the sounds that come second, i.e. the fricative. So, in this case, these are respectively 'S' and 'Z' in French.

4.2.2 German

Vowels

- 'ER': The only phoneme that is close to it is '@' in German because this is also a central vowel and they are also similar regarding the aperture of the mouth.
- 'AE': 'a' is the phoneme of the two languages that is the closest because it is located in the same area of the classification table.
- 'AA': For this sound, we focus on the aperture of the mouth to choose the right counterpart. As this is an open vowel, we opted for 'a' as counterpart in German.
- 'EY', 'OY', 'OW': For diphthongs, we kept the first vowel as a counterpart. So, we obtained respectively 'E', 'O' and 'O' in German.

Consonants

- 'TH', 'DH': As these phones are fricative dental, we decided to use respectively 'f' and 'v' which are fricative labiodental as counterparts in German.
- 'R': This sound is in free variation with the German 'r', that is to say that when they appear in the same environment, the meaning does not change and it is acceptable to a native speaker.
- 'CH', 'JH': For affricates, we affected the sounds that comes in second, i.e. the fricative. So, in this case, these are respectively 'S' and 'Z' in German.

4.3 Mapping: French to ...

4.3.1 English

Vowels

- 'i', 'u': We decided to choose 'IH' and 'UH' for the mapping because there is no big difference with the articulation of these sounds since they are shorter than 'IY' and 'UW' and so closest to the French sounds.

- 'y': For this sound, we did not take into account the place of articulation because we cannot see the tongue inside the oral cavity. However, the aperture of the mouth and the labiality help us to select the English vowel 'UW' because it is also close and rounding.

- 'in', 'an', 'on': For these three nasal vowels, we pronounced these sounds ourselves to see the movement of the lips to be as accurate as possible. So, we determined that 'AH', 'AO' and 'UH' are respectively the best candidates.

Consonants

- 'J': This phoneme is a kind of combination of the nasal 'N' and the approximant 'Y' for English. Therefore, we decided to keep a phoneme with the same manner of articulation, that is to say 'N'.

- 'R': This sound is in free variation with the English 'R', that is to say that when they appear in the same environment, the meaning does not change and this is acceptable to a native speaker.

- 'H': In order to find a counterpart for this sound, we considered it as a semi-vowel. Consequently, we know that this sound comes from the vowel 'y'. Therefore, they have the same counterpart 'UW'.

4.3.2 German

Vowels

- 'i', 'u', 'y': We decided to choose respectively 'I', 'U' and 'Y' because there is no big difference with the articulation of these sounds and they are less long than 'i:', 'u:' and 'y:' and so closest to the French pronunciation.

- 'in', 'on': For 'in', we used the oral equivalent '9' as a counterpart. Regarding 'on', we pronounced this sound ourselves to see the movement of the lips to be as accurate as possible and we estimated 'o:' to be the best candidate.

Consonants

- 'J': This phoneme is a combination of the nasal 'n' and the approximant 'j' in German. Therefore, we decided to keep a phoneme that has the same manner of articulation, that is to say 'n'.
- 'w': In order to find a counterpart for this sound, we considered it as a semi-vowel. Consequently, we know that this sound comes from the vowel 'u', but we noticed that 'w' appeared to be longer than 'u', so we opted for the German phone 'u:'.
- 'H': We also considered this sound as a semi-vowel. Consequently, we know that it comes from the vowel 'y'. Therefore, they have the same counterpart 'Y'.

4.4 Mapping: German to ...

4.4.1 English

Vowels

- '2:' : For this sound, we were based on two parameters, the manner of articulation and the rounding of the lips. As it is close-mid and rounding, we needed to search a rounding English vowel which is preferably almost close. 'UH' was the best candidate that matched with our requirement.
- '6': Among the central phonemes, we favored 'AH' as a counterpart rather than 'ER' due to the stretching of the lips during pronunciation of '6'.
- '9': To find a counterpart, we only took into account its aperture of the mouth. We therefore selected the closest English phone which is also open-mid, i.e. 'ER'.
- 'E', 'E:': We chose 'EH' because it is the only phone which is front and roughly in the middle.
- 'OY': For diphthongs, we kept the first vowel ('O' in German) as a counterpart. So, we obtained 'AO' for English.
- 'Y', 'y:': As 'Y' and 'y:' are close to 'U' and 'u:' both in terms of aperture and rounding of the lips, we have chosen the same equivalents, i.e., respectively, 'UH' and 'UW'.
- 'a', 'a:': For these two phones, we took 'AE' which is the closest one for both regarding the articulatory classification in German and English.

- 'a~': This phone is open and near-back, so we selected 'AA' which is also open and which has roughly the same place of articulation, back.
- 'o:': We picked 'AO' as counterpart because it is the only English phone which is back and circa in the middle.

Consonants

- '?': For this phone, we chose the vowel 'AH' in English because '?' is a glottal sound, so we needed a sound for which the lips are in a neutral position.
- 'C': This is a palatal fricative. We decided to select 'K' because it's located in the region next to the palatal one: the velar region.
- 'x', 'r': These sounds are in free variation with the English 'R', that is to say that when they appear in the same environment, the meaning does not change and this is acceptable to a native speaker.

4.4.2 French

Vowels

- '2:': We could have kept the French phone 'swa' as an equivalent but, although it looked similar, it was not rounded enough. Therefore, we decided that 'o' was the best candidate as it is also close-mid.
- 'I', 'U', 'Y': These three vowels are near-close, we opted for 'i', 'u', and 'y' respectively in French because they are all close vowels.
- 'aI', 'aU', 'OY': For diphthongs, we kept the first vowel as a counterpart. So, we obtained respectively 'a', 'a' and 'o' for French.

Consonants

- '?': For this phone, we chose the vowel 'swa' in French because '?' is a glottal sound, so we needed a sound for which the lips are in a neutral position.

- 'C': This is a palatal fricative. We decided to select 'k' because it's located in the region next to the palatal one: the velar region.
- 'x': This sound is in free variation with the French 'R', that is to say that when they appear in the same environment, the meaning does not change and this is acceptable to a native speaker.
- 'N': This phoneme seems to be close to the oral equivalent (i.e. plosive) 'g', so we opted for it.

5 Video Generation

In this part, we will describe the different steps, as described in figure 12, of the generation of videos for the evaluation of our multilingual system.

5.1 Step 1

First, we created 20 sentences per language that cover all the sounds in each of them (see tables 8, 7 and 9 in Appendices). We favored short sentences, on the one hand in relation to the duration of the survey and on the other hand because the longer the sentences, the longer it takes to generate the videos.

The existing software allowed us to generate a segmentation file for each of those sentences that sliced them phone by phone and displayed the duration of each phone. This could then be input into the articulation generator to be translated into articulatory movements of the talking head by the animation generator. This program uses the audio created for each sentence by text-to-speech synthesis and then pairs it with the visual of the talking head.

This is the process that is usually followed in a monolingual system. We obtained videos using the segmentation file of Language A associated with the audio of the Language A so that participants could rate their quality in our survey.

The next step was to create animations of Language A using the articulation of Language B, so that we could test our theory on their quality compared to the quality of a system specifically tailored for the constraints of a given language.

5.2 Step 2

Once we obtained the segmentation file of Language A, we used a code described in the following section to apply the mapping.

5.2.1 Code

We have created a code in Python programming language (page 36) which transforms a segmentation file of Language A into a segmentation file of Language B as can be seen in figure 13 in Appendices. For this, we have used six dictionaries to map phones in one language and their counterparts in another language.

More precisely, each line of the input consists of the duration of the sound, the beginning and the end, as well as the symbol used in our system. So, in the function called do_mapping, we keep the durations as they are but replace the phone of Language A with its counterpart in Language B by using the dictionaries we previously established.

To run the code more easily, we have used the argparse module and created a function called get_cli_args which consists in adding arguments to the parser. Therefore, in the command prompt, we have to enter: –from Language A –to Language B inputs. First, we specify the two languages in order to select the mapping dictionary and then the file on which we want to execute the code. In the main function, we avoid the case where the user wants to establish a mapping between the same language because it is useless. When the parameters are correct, we display the name of the file and the message "done" when the execution is finished.

5.2.2 Procedure

Once the mapping was done, we took the segmentation file of Language B and generated the articulation trajectories of it. Then, the interface paired them with the audio of Language A in order to create an animation of a sentence in Language A.

We used this whole process for each reference language: English, French and German. We obtained a total of 180 videos, i.e. 60 per language.

These were the videos we presented to our native speakers in the survey we used to evaluate the quality of the monolingual and multilingual talking heads.

6 Evaluation

6.1 Methodology

We conducted three online surveys on user preferences regarding the articulation of a talking head on an existing survey platform (TTSEVAL) advised by our supervisor.

We were particular in our choice of participants and looked for native speakers for each survey. By native speaker, we understand a person who has spoken that language since birth or for most of their lives and is comfortable in it. We also accepted bilingual native speakers.

Our minimum number of participants was 10 people, but the goal was to have around 15 to 20 in order to have a bigger pool and a more representative set of results. We shared the links of the evaluations with our acquaintances, our classmates and our language teachers who also kindly passed them on.

We made decisions on how to approach the subjective evaluation: we needed to give participants a proper introduction to the functioning of the platform and what was expected of them so as to make sure that the questionnaires were understandable to all and that the devices they were using to complete the survey were appropriate. We also gave them tips on to best analyze the quality of the videos, such as memorizing the sentence and performing the lip movement along with the animation to compare the talking head to reality.

We also made sure to ask them precise questions so they had no doubts about what was expected of them. There were two parts in the evaluation surveys. First, we asked participants to evaluate the quality of videos one by one. The following block asked them to compare three videos of the same sentence generated by all three systems to see which one resembled reality the most and which one was the furthest from it.

6.1.1 Rating Independent Videos

For the three surveys, we halved our set of 60 videos. In Part 1, we required participants to independently score the articulation of the talking head in 30 videos. Note that the videos were shown in a random order. We used a 5-answers Likert scale. It allowed participants to express their degree of satisfaction (from very bad to very good) using a slider. We added an example of the survey pages for this part in the Appendices section (see Figure 14).

6.1.2 Comparison of Videos

In this part, we used the remaining 30 videos and chose to present three side by side videos to the participants. The first video of each batch was generated with the basic monolingual system and the other two with our previously introduced mapping mechanism which acts as a multilingual system. The goal was to compare the three articulations through three questions. The first tested whether they saw a difference between the three videos as they sometimes seemed to be similar or had minimal difference. The others required them to determine which video featured the articulation closest to and furthest from reality. In addition to these queries, we have also given them the option of leaving a comment so that they can express their opinion about the viewing. We added an overview of this second part in the Appendices section (see Figure 15).

6.2 Results and Analysis

As the technical difficulties we encountered with the softwares set us back significantly, we only had 10 days to share the surveys and unfortunately did not have as many participants as initially expected.

We took the risk of keeping invalid answers such as missing values (3 cases in French where a question was inadvertently skipped by the interface) or, specifically for English, partially completed surveys. Indeed, while making the survey, there was a technical problem and as such part of the second half was not saved. Unfortunately, the links were sent out before we rectified the mistake and we had to send out a second survey with the missing questions, which was only filled out by some of the initial participants.

We also paid attention to the number of plays and pauses to verify that the person performed the study properly. When we found anomalies (video was not played, time between questions was too short), we agreed to remove all of the participant's answers from our pool. Keeping these parameters in mind, we collected the responses of 13 participants in English (including 4 who did only the first part), 17 in French and only 2 in German.

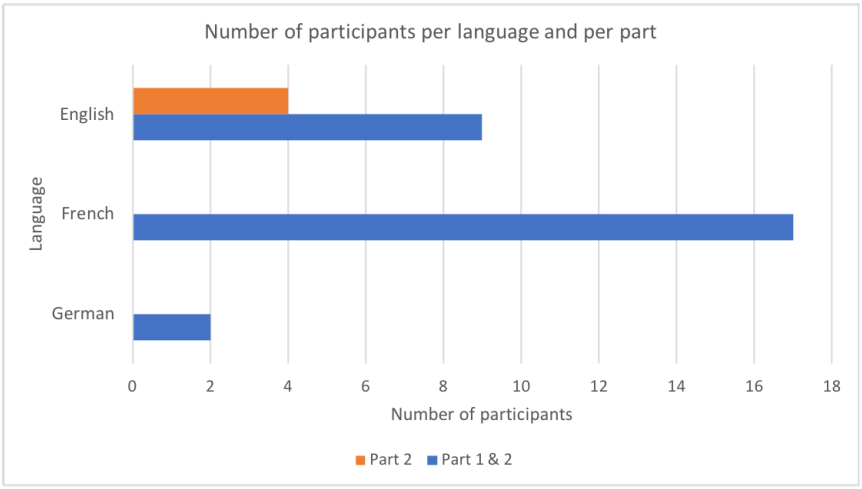


Figure 1: Number of participants per survey

The results shown in the three following graphs are the average of the results given by all of the participants of that language. To reiterate, the articulation shown in the videos in the first half are rated from 1 to 5, and the second part asked participants which one was closest and which one was furthest from reality, with the possibility to choose none, one, two, or all videos in each case. We have offered these answer choices because some videos seem to have no differences between them.

English gives us heterogeneous results, as viewers do not have as clear of a preference: the video made with the English system is only ranked the best 5 times out of 10, without as big of a margin as the one we see in the French results. This can probably be explained by the fact that our participants had different backgrounds (diverse dialects and accents) and therefore had differing opinions on "reality" and what was closest/furthest from it.

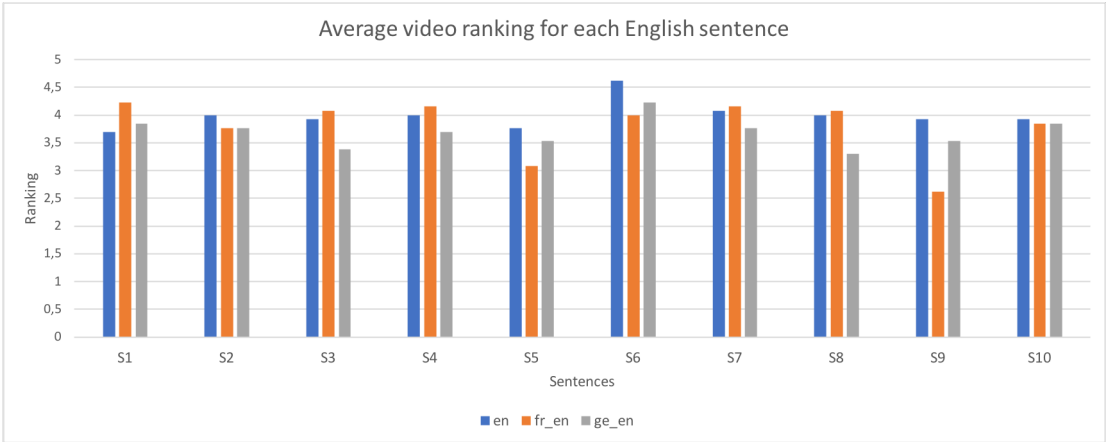


Figure 2: English part 1 results

As we can see from Figure 3, French speakers vastly preferred the videos made using the French articulatory movements, as 9 sentences out of 10 have the highest average in that case.

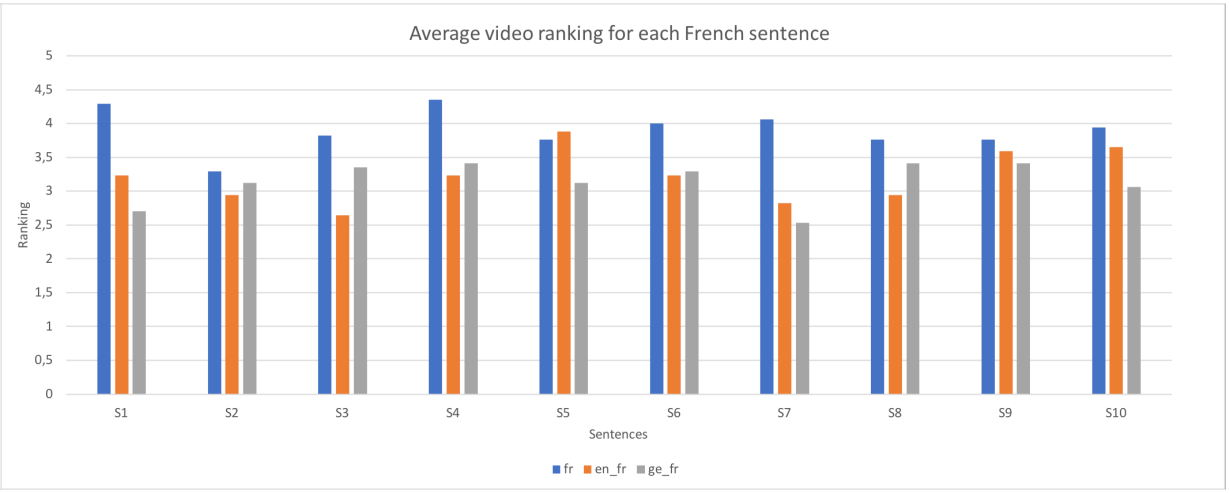


Figure 3: French part 1 results

Unfortunately, the German data cannot be seriously analyzed due to the lack of participants, but we include the graph of the results to show the tendency, where only 3 of the videos saw the German system chosen as superior.

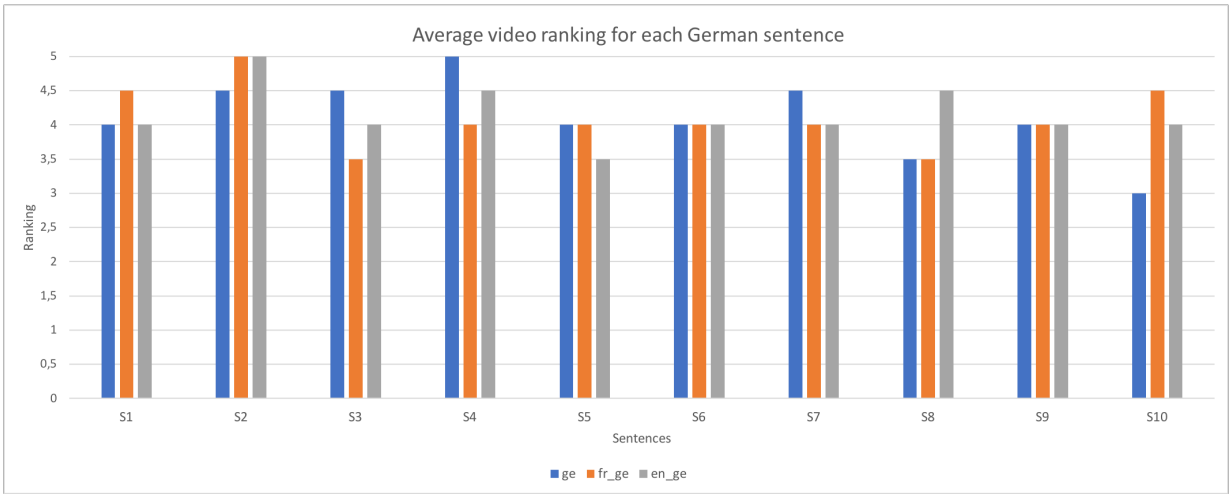


Figure 4: German part 1 results

In the second part of the survey where participants were asked to compare the quality of the articulation across the three systems, we must note that some chose "all" or "none" as their favorite, and these averages must be looked at when considering the individual videos' results.

33.6% of English speakers enjoy the talking head with the French articulation (video 2) compared to 30.9% who preferred the monolingual English video. The German articulation was the one they appreciated as the one most lacking in quality (31% of participants see it as the furthest from reality). Video 1, which was supposed to be the best one if we consider our hypothesis that monolingual systems are superior, was seen as the furthest one from reality by 19% of the participants. That is to say that, as mentioned before, we had a broad range of participants and differing opinions.

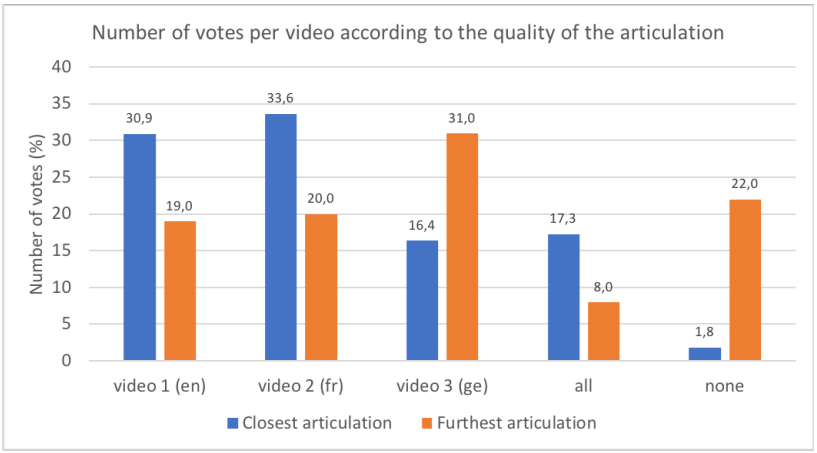


Figure 5: English part 2 results

French speakers, on the other hand, have a clear favorite as video 1 (made with the French audio and articulation) is the overall best with 38.1% of participants choosing it as the closest to reality. Still, it is important to note that 19.2% of participants chose the articulation of video 1 as the one furthest from reality. As was the case in the English survey, 35.9% of participants see the video created using the German lips movement as the one furthest from reality.

We also note the fact that the monolingual video is the one least chosen to be the furthest from reality at 19.2%, while the English system generated one was 23.7% and German 35.9%.

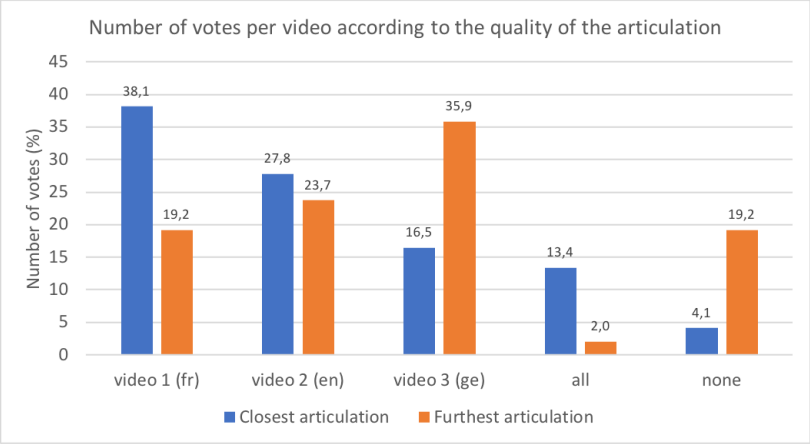


Figure 6: French part 2 results

The German data, while unusable due to the lack of participants, also shows us a general tendency with 30% choosing the monolingual video as the closest to reality.

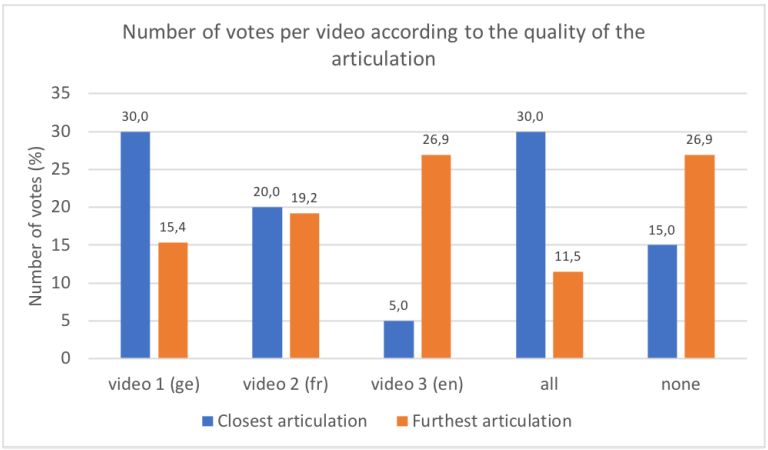


Figure 7: German part 2 results

We let participants write their comments on the videos after taking the survey; here we quote the ones that are of interest to this study:

One participant thinks that "[all talking heads seem] similar but I hear the diction is entirely different from one another." This comment perfectly illustrates the McGurk effect, where the listener is swayed by what is visually presented to them. As a reminder, while the articulation is different, we used the same exact audio for all three videos.

We also note comments on the quality of the animation: "Sometimes the articulation is slower than the audio" and "some [videos] seemed to have a realistic pronunciation but [the visual was] a

little out of sync with the sound". Our hypothesis is that this can be due to coarticulation: the overlap between phones might be happening earlier or later than is usual for that language and thus there seems to be a lag for the native speaker.

Others mention the fact that "at the beginning the [talking head] often has his lips in a strange position but it doesn't really impact the articulation" and "the man's lip from video 2 and video 3 are barely open when speaking which makes this look so unreal", which leads us to believe that humans can pick up the slightest change and as such a generalized system for all languages would fail to take into account minute details such as this.

Another participant comments that "some sentences were quite short, so it was difficult to judge if the whole pronunciation was realistic", but that unfortunately stemmed from the fact that the videos had to be short so that the software could handle their weight.

7 Conclusion

To conclude, this study on an animated speech system was elaborated so that we could evaluate whether coarticulation and other characteristics of a language are proprietary to a language and therefore impact the quality of the animation of a multilingual talking head. Our hypothesis was that a multilingual system like that of Taylor et al. would not be of the best quality compared to a monolingual system that works specifically for a precise set of rules.

In this report, we detailed our methodology for the mapping of the phonesets and then explained how we created the videos and surveys that allowed us to acquire the results of the evaluations to be analyzed to see whether the multilingual system is viable when it comes to audiovisual synthesis.

Audiovisual mainly relies on human preferences, and thus a subjective evaluation is the only way to evaluate this kind of multilingual system. We had the advantage of having the existing TTSEVAL platform to create the surveys, and it has important features (compute the number of play, pause, time) that allowed us to sort the responses quickly.

We can see from the results that the monolingual systems render articulation the best according to French native speakers, even though a big percentage of participants seemed to find that videos created by a multilingual system were suitable (e.g. English speakers rating the video created with the French system highest).

We observed a clear preference from the French native speakers and a slight penchant from the English, and as such we establish that a set of monolingual systems remains the better alternative. The quality of the animation of a multilingual speech system, while not terrible, would not allow this software to be used for things that call for extreme precision, such as lip rendering for hard-of-hearing individuals for example.

8 Discussion

We believe our results of the evaluation can be useful to the next researchers wishing to undertake a similar study or further this one by completing what we still have missing. For further research, once could reach out to native German speakers and evaluate their opinion on these videos. Another interesting angle would be to take coarticulation into consideration when creating a multilingual system in order to see if the average given by native speakers rises.

In case this research is to be continued, the documents needed were sent to our supervisor Slim Ouni:

- our mapping code with a readme file
- a notice explaining our mapping choices and containing the final version of the tables
- the segmentation files (180 in total)
- all videos generated by the three systems (180 in total)
- the evaluation documents containing the results of the survey from each participant (22 for the French speaking participants, 26 for the English speaking participants, and 3 for the German speaking participants, with a total of 51 files)
- a spreadsheet with our data and graphs

We were interested in audiovisual studies and seeing how the system works from the Text-to-Speech synthesis to the animation generator and everything in between. We got to work with a very interesting side of Natural Language Processing and also delved into the world of research.

One unfortunate aspect was the fact that a lot of the existing software had to be debugged, and as such we were set back while Theo Biasutto-Lervat worked on that. While we have managed

to meet the deadline, the quality is not what we aimed for when we started this second part of the project back in January. We explain our setbacks below.

While working on the mapping of the phonemes used in the system, we noticed inconsistencies that hindered our efforts: the symbols used in the segmentation process were inaccurate in a way that could cause the animation to look imprecise: some languages had phones that were doubled, meaning that the same sound could be interpreted in two ways, others had phones that were not in use.

For this reason, we had to do the mapping multiple times as there existed multiple phone sets that were built upon each other. We finally decided to work with SAMPA characters instead of the ones programmed into the system in order to be sure of the precision of the mapping while we waited to see if the problem could be fixed in time.

We were told that we would not be able to complete the evaluation in time and to focus our report on the mapping so that the Multispeech team could spring from our work when they did the evaluation, but we decided we would try to see the project through as we were very interested in the results. Once the bugs in the software were corrected, we replaced the SAMPA characters with the ones in the systems so that the animation program could parse them.

We also note the fact that some systems used X-SAMPA and others ARPABET. This disconnect added an extra step to our mapping that could have been avoided if all systems were based on the same foundation.

As for the evaluation, the native speakers we had contacted for each language had a very limited time to fill out the survey and we also had trouble finding more willing participants. French speakers were not necessarily an issue, and we had a certain number of native English speakers, but we only managed to find two German speakers to join us. There are other approaches available (such as crowd-sourcing) but they require time we did not have.

It is unfortunate that we do not get to fully work on the evaluation of the system, as this was the core of the project: the part that most interested us and could have taught us a lot. As it is, we enjoyed doing the research but feel like we were deprived of the main goal of the project due to circumstances out of our control.

9 Appendices

System	IPA	Example	Transcription
P	p	pie	**P** AY
B	b	buy	**B** AY
K	k	clean	**K** L IY N
G	ɡ	guy	**G** AY
F	f	fan	**F** AE N
V	v	visits	**V** IH Z IH T S
T	t	trap	**T** R AE P
D	d	dress	**D** R EH S
TH	θ	through	**TH** R UW
DH	ð	rhythm	R IH **DH** AH M
S	s	spy	**S** P AY
Z	z	nose	N OW **Z**
SH	ʃ	fruition	F R UW IH **SH** AH N
ZH	ʒ	pleasure	P L EH **ZH** ER
HH	h	high	**HH** AY
M	m	meet	**M** IY T
N	n	when	W EH **N**
NG	ŋ	morning	M AO R N IH **NG**
R	ɹ	try	T **R** AY
L	l	letter	**L** EH T ER
W	w	wine	**W** AY N
Y	j	yes	**Y** EH S
CH	tʃ	catch	K AE **CH**
JH	dʒ	giant	**JH** AY AH N T

System	IPA	Example	Transcription
IY	iː	leave	L **IY** V
IH	ɪ	mirror	M **IH** R ER
UW	uː	new	N Y **UW**
UH	ʊ	foot	F **UH** T
EY	eɪ	face	F **EY** S
EH	e	merry	M e R IY
ER	ɜː, ə^{r}	nurse, never	N **ER** S, N EH V **ER**
AO	ɔː	horse	HH **AO** R S
AE	æ	rabbit	R **AE** B IH T
AH	ʌ, ə	button, bottle	B **AH** T **AH** N, B AA T **AH** L
AA	ɑː, ɒ	start, follower	S T **AA** R T, F **AA** L OW ER
OY	ɔɪ	choice	CH **OY** S
OW	əʊ	goat	G **OW** T
AY	aɪ	white	W **AY** T
AW	aʊ	mouth	M **AW** TH

Table 1: English System symbols and the corresponding symbols of the International Phonetic Alphabet (IPA)

System	IPA	Example	Transcription
p	p	pain	**p** in
b	b	bois	**b** w a
t	t	temps	**t** an
d	d	dimanche	**d** i m an S
k	k	quartier	**k** a R t i e
g	g	graine	**g** R e n
f	f	fête	**f** e t
v	v	vent	**v** an
s	s	soleil	**s** o l e j
z	z	bêtise	b e t i **z**
S	ʃ	blanche	b l an **S**
Z	ʒ	jardin	**Z** a R d in
m	m	matin	**m** a t in
n	n	uniforme	y **n** i f o R m
J	ɲ	montagne	m on t a **J**
l	l	bocal	b o k a **l**
R	ʁ	robot	**R** o b o
w	w	oiseau	**w** a z o
H	ɥ	puissant	p **ɥ** i s an
j	j	fille	f i **j**

System	IPA	Example	Transcription
i	i	musique	m y z **i** k
e	e, ɛ	boulanger, mère	b u l an Z **e**, m **e** R
a	a, ɑ	magasin, château	m **a** g **a** z in , S **a** t o
o	o, ɔ	forêt, homme	f **o** R E, **o** m
u	u	coupe	k **u** p
y	y	rue	R **y**
swa	ə, ø, œ	petite, feu, fleur	p **swa** t i t, f **swa**, f l **swa** R
in	ɛ̃, œ̃	chien, un	S j **in**, **in**
an	ɑ̃	parent	p a R **an**
on	ɔ̃	poisson	p w a s **on**

Table 2: French System symbols and the corresponding symbols of the International Phonetic Alphabet (IPA)

System	IPA	Example	Transcription
p	p	Post	**p** O s t
b	b	besonders	**b** @ z O n d 6 s
t	t	Tür	**t** y: 6
d	d	Deich	**d** aI C
k	k	kalt	**k** a l t
g	g	morgens	m O 6 **g** @ n s
?	ʔ	erinnere	E 6 **?** I n @ r @
f	f	fast	**f** a s t
v	v	Wort	**v** O 6 t
s	s	Klasse	k l a s @
z	z	Wiese	v i: **z** @
S	ʃ	stumm	**S** t U m
Z	ʒ	Genie	**Z** E n i:
C	ç	Plätzchen	p l E t s **C** @ n
x	x	Sprachen	S p r a: **x** @ n
h	h	Hand	**h** a n t
j	j	Jahr	**j** a: R
m	m	Monat	**m** o: n a t
n	n	nicht	y **n** I C t
N	ŋ	Dinge	d I **N** @
l	l	Labor	**l** a b o: 6
r	ʁ	Raum	**r** aU m

System	IPA	Example	Transcription
I	ɪ	Friseur	f r **I** z 2: 6
E	ɛ	entdecken	**E** n t d **E** k @ n
a	a	rasch	R **a** S
O	ɔ	dennoch	d E n **O** x
U	ʊ	Munde	m **U** n d @
Y	ʏ	hübsche	h **Y** p S @
9	œ	zwölf	ts v **9** l f
i:	i:	Ziegen	ts **i:** g @ n
e:	e:	Leoparden	L **e:** o: p a R d @ n
E:	ɛ:	Zähne	ts **E**: n @
a:	a:	Haare	h **a**: R @
o:	o:	Brot	f b R **o**: t
u:	u:	Supermarkt	z **u**: p 6 m a R k t
y:	y:	über	**y**: b 6
2:	ø:	fröhlich	f R **2**: l I C
aI	aɪ	kleinen	k l **aI** n @ n
aU	aʊ	Urlaub	u: 6 l **aU** p
OY	ɔʏ	Träumer	t R OY m 6
@	ə	offen	O f **@** n
6	ɐ	später	S p E: t **6**
a~	ɑ̃	Tante	t a~t @

Table 3: German System symbols and the corresponding symbols of the International Phonetic Alphabet (IPA)

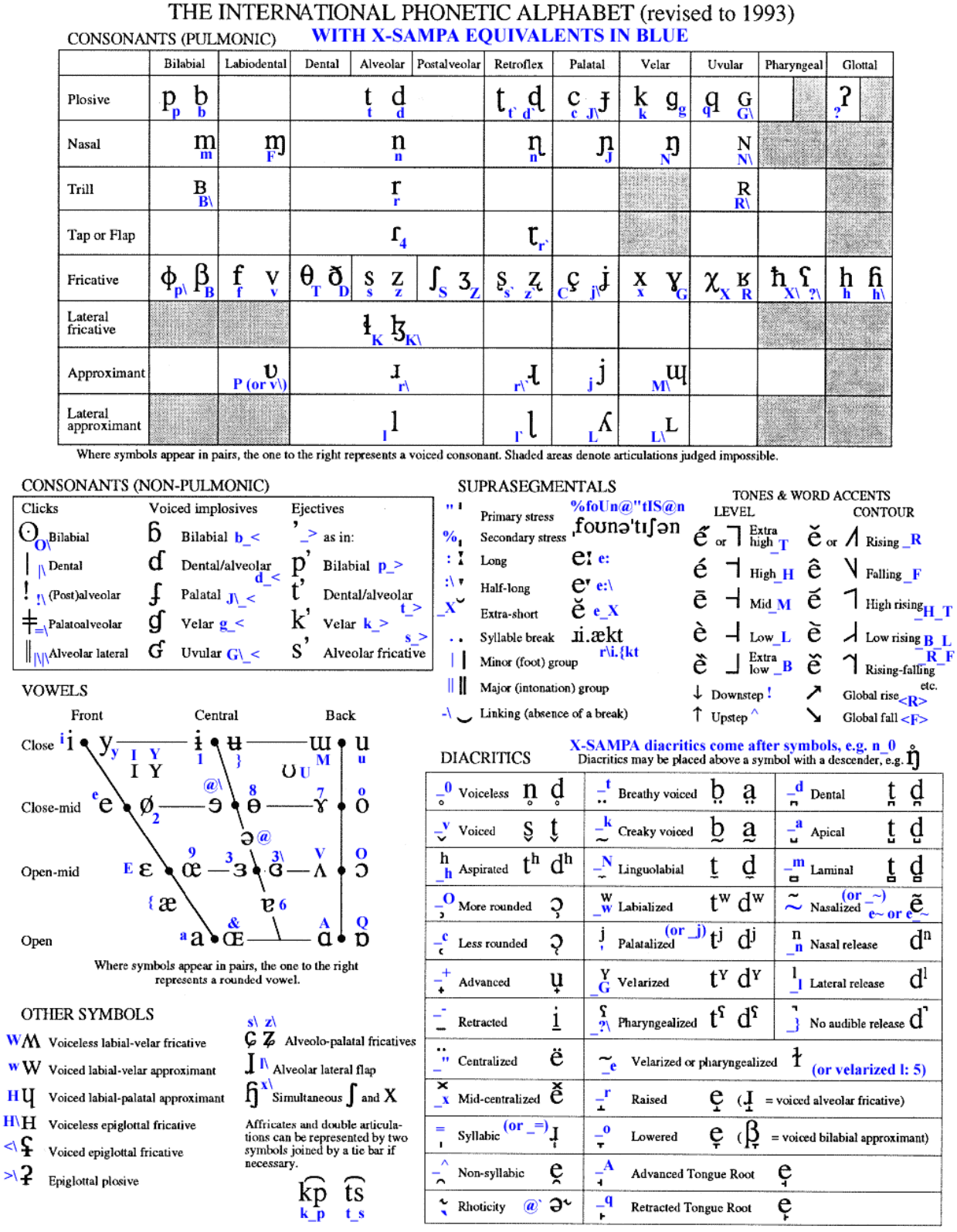


Figure 8: IPA chart with SAMPA correspondence

		Place of articulation															
		Bilabial		Labiodental		Dental		Alveolar		Postalveolar		Palatal		Velar		Glottal	
Manner of articulation	Plosive	P	B					T	D					K	G		
	Nasal		M						N						NG		
	Fricative			F	V	TH	DH	S	Z	SH	ZH					HH	
	Approximant								R				Y		W*		
	Lateral approximant								L								
	Affricate									CH	JH						

*W= labial-velar

(a) English consonants

		Place of articulation									
		Front		Near-front		Central		Near-back		Back	
Aperture	Close	IY									UW
	Near-close			IH					UH		
	Close-mid	EH									
	Mid					AH					
	Open-mid						ER			AH	AO
	Near-open	AE									
	Open									AA	AA

+ Diphthongs: EY, OY, OW, AY, AW

(b) English vowels

Legend:
~ nasalization
: elongation

unvoiced	voiced

unrounded	rounded

Figure 9: Articulatory classification of English system phones

		Place of articulation													
		Bilabial		Labiodental		Alveolar		Postalveolar		Palatal		Velar		Uvular	
Manner of articulation	Plosive	p	b			t	d					k	g		
	Nasal		m				n				J				
	Fricative			f	v	s	z	S	Z						R
	Approximant										H* j		w*		
	Lateral approximant						l								

*w = labial-velar; H = labial-palatal

(a) French consonants

		Place of articulation					
		Front		Central		Back	
Aperture	Close	i	y				u
	Close-mid	e	swa				o
	Mid			swa			
	Open-mid	e in	swa in				o on
	Open	a				a an	

(b) French vowels

Legend:
~ nasalization
: elongation

unvoiced	voiced

unrounded	rounded

Figure 10: Articulatory classification of French system phones

		Place of articulation															
		Bilabial		Labiodental		Alveolar		Postalveolar		Palatal		Velar		Uvular		Glottal	
Manner of articulation	Plosive	p	b			t	d					k	g			?	
	Nasal		m				n						N				
	Fricative			f	v	s	z	S	Z	C		x			r	h	
	Approximant										j						
	Lateral approximant						l										

(a) German consonants

		Place of articulation									
		Front		Near-front		Central		Near-back		Back	
Aperture	Close	i:	y:								u:
	Near-close			I	Y				U		
	Close-mid	e:	2:								o:
	Mid					@					
	Open-mid	E E:	9								O
	Near-open					6					
	Open	a a:							a~		

+ Diphthongs: aI, aU, OY

(b) German vowels

Legend:
~ nasalization
: elongation

unvoiced	voiced

unrounded	rounded

Figure 11: Articulatory classification of German system phones

	EN → FR	
Vowels	IY	i
	IH	i
	UW	u
	UH	u
	EY	e
	EH	e
	ER	swa
	AO	o
	AE	a
	AH	swa
	AA	a
	OY	o
	OW	o
	AY	a
	AW	a
Consonants	P	p
	B	b
	K	k
	G	g
	F	f
	V	v
	T	t
	D	d
	TH	f
	DH	v
	S	s
	Z	z
	SH	S
	ZH	Z
	HH	swa
	M	m
	N	n
	NG	g
	R	R
	L	l
	W	w
	Y	j
	CH	S
	JH	Z
Silence	SIL	sil

	EN → GE	
Vowels	IY	i:
	IH	I
	UW	u:
	UH	U
	EY	e:
	EH	e:
	ER	@
	AO	O
	AE	a
	AH	@
	AA	a
	OY	O
	OW	O
	AY	aI
	AW	aU
Consonants	P	p
	B	b
	K	k
	G	g
	F	f
	V	v
	T	t
	D	d
	TH	f
	DH	v
	S	s
	Z	z
	SH	S
	ZH	Z
	HH	h
	M	m
	N	n
	NG	N
	R	r
	L	l
	W	w
	Y	j
	CH	S
	JH	Z
Silence	SIL	sil

Table 4: Mapping with English as reference language

	FR → EN	
Vowels	a	AE
	e	EH
	i	IH
	o	AO
	u	UH
	y	UW
	an	AO
	swa	AH
	in	AH
	on	UH
Consonants	p	P
	b	B
	t	T
	d	D
	k	K
	g	G
	f	F
	v	V
	s	S
	z	Z
	S	SH
	Z	ZH
	m	M
	n	N
	J	N
	l	L
	R	R
	w	W
	H	UW
	j	Y
Silence	sil	SIL

	FR → GE	
Vowels	a	a
	e	E
	i	I
	o	o:
	u	U
	y	Y
	an	a~
	swa	@
	in	9
	on	o:
Consonants	p	p
	b	b
	t	t
	d	d
	k	k
	g	g
	f	f
	v	v
	s	s
	z	z
	S	S
	Z	Z
	m	m
	n	n
	J	n
	l	l
	R	r
	w	u:
	H	Y
	j	j
Silence	sil	sil

Table 5: Mapping with French as reference language

	GE → EN	
Vowels	2:	UH
	6	AH
	9	ER
	@	AH
	E	EH
	E:	EH
	I	IH
	O	AO
	OY	AO
	U	UH
	Y	UH
	a	AE
	a:	AE
	aI	AY
	aU	AW
	a~	AA
	e:	EH
	i:	IY
	o:	AO
	u:	UW
	y:	UW
Consonants	p	P
	b	B
	t	T
	d	D
	k	K
	g	G
	?	AH
	f	F
	v	V
	s	S
	z	Z
	S	SH
	Z	ZH
	C	K
	x	R
	m	M
	n	N
	N	NG
	l	L
	r	R
	j	Y
Silence	sil	SIL

	GE → FR	
Vowels	2:	o
	6	swa
	9	swa
	@	swa
	E	e
	E:	e
	I	i
	O	o
	OY	o
	U	u
	Y	y
	a	a
	a:	a
	aI	a
	aU	a
	a~	an
	e:	e
	i:	i
	o:	o
	u:	u
	y:	y
Consonants	p	p
	b	b
	t	t
	d	d
	k	k
	g	g
	?	swa
	f	f
	v	v
	s	s
	z	z
	S	S
	Z	Z
	C	k
	x	R
	m	m
	n	n
	N	g
	l	l
	r	R
	j	j
Silence	sil	sil

Table 6: Mapping with German as reference language

1	He will buy her a bottle of wine when they marry.
2	She tried to spy his thigh in the mirror through the coir curtains.
3	The nurse breathed in a hoarse sigh.
4	He'll try to catch a rabbit while singing a new tune.
5	Her thoughts were lost on his merry letter.
6	My new cat has a pedigree.
7	The goat is in a hurry, but its foot is in a trap.
8	The tie-dye dress of her choice had been ruined by the pie he had 9 enthused over.
9	The rhythm of the horse influenced the strut of the courier.
10	She would leave enough flour for it to sink into.
11	His nose was like a button on his square face.
12	This serious tour guide was a mediocre choice for historic visits.
13	Can you believe the lies that come out of his mouth?
14	He has a cruel follower from California, a fan consumed by emotion.
15	The morning dew reflected the giant white sky.
16	Yes, it had been a pleasure to meet the guy and start the day.
17	A high whine came from the lute in his palm.
18	They hired Zeus to clean the parking lot.
19	They argued over the moral of the Oxford comma.
20	His endeavors never came to fruition.

Table 7: English sentences

1	Le sorcier habite dans un château.
2	Les chiens font des bêtises.
3	Ce robot est très puissant.
4	Le poisson tourne dans son bocal.
5	Les oiseaux mangent leurs graines.
6	Cet homme fait un jogging tous les matins.
7	Le temps est orageux.
8	Mes parents marchent dans les rues du quartier.
9	Ma mère coupe des fleurs blanches dans son jardin.
10	Le parking du magasin est rempli.
11	Pauline se promène dans la forêt.
12	La petite fille ramasse des châtaignes.
13	Le boulanger faisait cuire du pain au feu de bois.
14	Hugo aime bien se reposer au soleil.
15	Le vent souffle derrière les montagnes.
16	Léa chante avec la chorale tous les dimanches.
17	La fête de la musique est célébrée en juin.
18	Alice enfile l'uniforme de son école.
19	Notre nouvelle tente de camping est immense.
20	Les musiciens accordent leurs instruments.

Table 8: French sentences

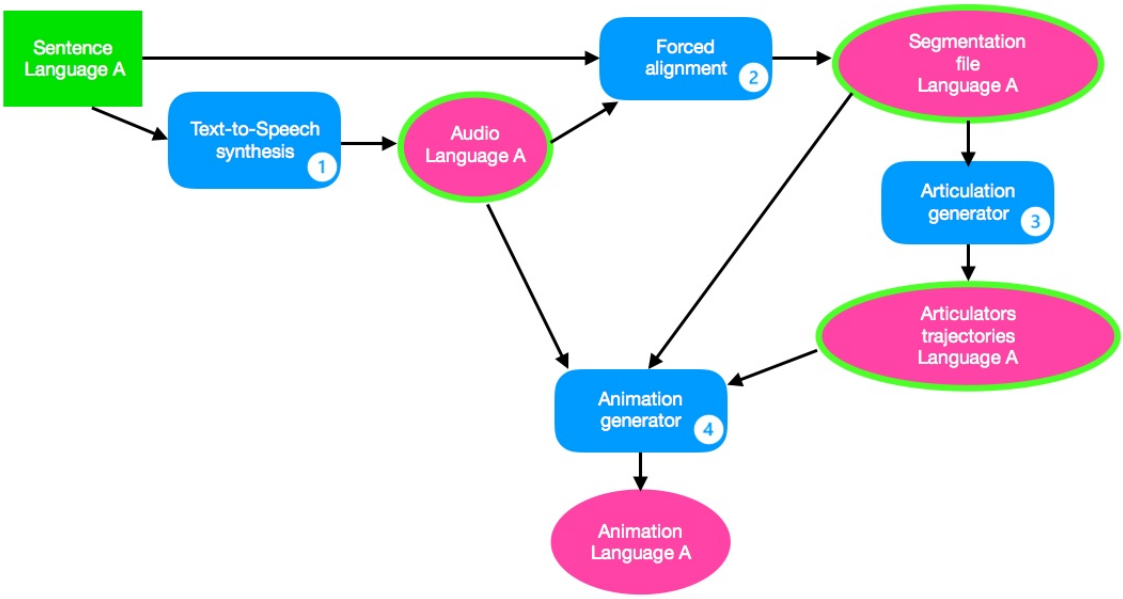
1	Franz jagt im komplett verwahrlosten Taxi quer durch Bayern.
2	Victor jagt zwölf Boxkämpfer quer über den großen Sylter Deich.
3	In diesem Augenblick kam die Post.
4	Die Klasse blieb stumm nach dieser kleinen Rede.
5	Die Plätzchen waren so hart, dass sie sich fast die Zähne ausbissen.
6	Sie gingen rasch hinaus und sprachen kein Wort.
7	Anfang November wurde es sehr kalt.
8	Er sang, besonders morgens: dennoch erschient er nicht fröhlich zu sein.
9	Ein Monat später, versuchte er es noch ein mal.
10	In Allgemeinem, erinnere ich mich nicht an meiner Träumer.
11	Ich kaufe im Supermarkt Äpfel, Brot und Schokoladenkuchen.
12	Die Ziegen tollen über die Wiese und entdecken viele Dinge.
13	Mein Onkel fällt der Kamm aus der Hand.
14	Das hübsche Holzkamel hat mir meine Tante aus dem Urlaub mitgebracht.
15	Dieses deutsche Genie lebt in seinem Labor.
16	Sein gesprenkeltes Fell tarnt den Leoparden im dichten Dschungel.
17	Lena hat beim Friseur ihre Haare kurz geschnitten.
18	Ich betrat den Raum, weil die Tür offen war.
19	Diese Wörter sind in aller Munde.
20	Das Bulletin erscheint mit zwei Ausgaben im Jahr.

Table 9: German sentences

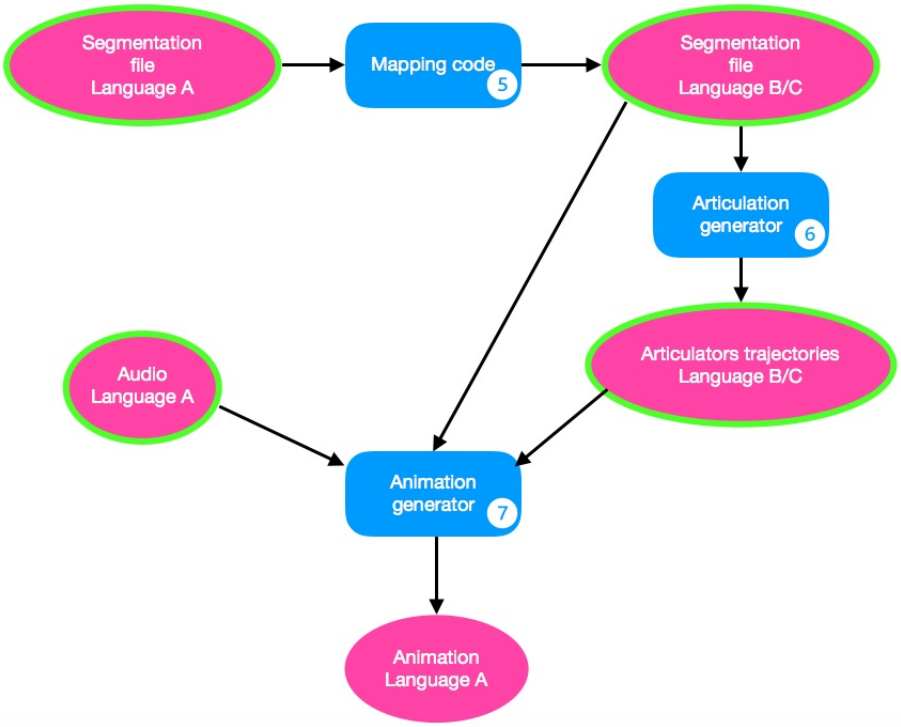
```
"""
    Authors: Juliana De Ferran, Sonita Te, Stephanie Monteiro
    Date: June 2021
    Goal: As part of our supervised project, this code is used to transform a segmentation
    file of a language A into a segmentation file of a language B by mapping the phonemes of
     language A to the phonemes of a language B.
"""

# !/usr/bin/env python
# coding: utf-8

import argparse

# mapping dictionaries
map_en_to_fr = {'SIL': 'sil', 'AA': 'a', 'AE': 'a', 'AH': 'swa', 'AO': 'o', 'AW': 'a', 'AY':
     'a', 'B': 'b', 'CH': 'S', 'D': 'd', 'DH': 'v', 'EH': 'e', 'ER': 'swa', 'EY': 'e', 'F':
    'f', 'G': 'g', 'HH': 'swa', 'IH': 'i', 'IY': 'i', 'JH': 'Z', 'K': 'k', 'L': 'l', 'M': 'm
    ', 'N': 'n', 'NG': 'g', 'OW': 'o', 'OY': 'o', 'P': 'p', 'R': 'R', 'S': 's', 'SH': 'S', '
    T': 't', 'TH': 'f', 'UH': 'u', 'UW': 'u', 'V': 'v', 'W': 'w', 'Y': 'j', 'Z': 'z', 'ZH':
    'Z'}
map_en_to_ge = {'SIL': 'sil', 'AA': 'a', 'AE': 'a', 'AH': '@', 'AO': 'O', 'AW': 'aU', 'AY':
    'aI', 'B': 'b', 'CH': 'S', 'D': 'd', 'DH': 'v', 'EH': 'e:', 'ER': '@', 'EY': 'e:', 'F':
    'f', 'G': 'g', 'HH': 'h', 'IH': 'I', 'IY': 'i:', 'JH': 'Z', 'K': 'k', 'L': 'l', 'M': 'm'
    , 'N': 'n', 'NG': 'N', 'OW': 'O', 'OY': 'O', 'P': 'p', 'R': 'r', 'S': 's', 'SH': 'S', 'T
    ': 't', 'TH': 'f', 'UH': 'U', 'UW': 'u:', 'V': 'v', 'W': 'U', 'Y': 'j', 'Z': 'z', 'ZH':
    'Z'}
map_fr_to_en = {'sil': 'SIL', 'H': 'UW', 'J': 'N', 'R': 'R', 'S': 'SH', 'Z': 'ZH', 'a': 'AE'
    , 'b': 'B', 'd': 'D','e': 'EH', 'f': 'F', 'g': 'G', 'i': 'IH', 'j': 'Y', 'k': 'K', 'l':
    'L', 'm': 'M', 'n': 'N', 'o': 'AO', 'p': 'P', 's': 'S', 't': 'T', 'u': 'UH', 'v': 'V', '
    w': 'W', 'y': 'UW', 'z': 'Z', 'an': 'AO', 'swa': 'AH', 'in': 'AH', 'on': 'UH'}
map_fr_to_ge = {'sil': 'sil', 'H': 'Y', 'J': 'n', 'R': 'r', 'S': 'S', 'Z': 'Z', 'a': 'a', 'b
    ': 'b', 'd': 'd', 'e': 'E', 'f': 'f', 'g': 'g', 'i': 'I', 'j': 'j', 'k': 'k', 'l': 'l',
    'm': 'm', 'n': 'n', 'o': 'o:', 'p': 'p', 's': 's', 't': 't', 'u': 'U', 'v': 'v', 'w': 'u
    :', 'y': 'Y', 'z': 'z', 'an': 'a~', 'swa': '@', 'in': '9', 'on': 'o:'}
map_ge_to_en = {'sil': 'SIL', '2:': 'UH', '6': 'AH', '9': 'ER', '?': 'AH', '@': 'AH', 'C': '
    K', 'E': 'EH', 'E:': 'EH', 'I': 'IH', 'N': 'NG', 'O': 'AO', 'OY': 'AO', 'S': 'SH', 'U':
    'UH', 'Y': 'UH', 'Z': 'ZH', 'a': 'AE', 'a:': 'AE', 'aI': 'AY', 'aU': 'AW', 'a~': 'AA', '
    b': 'B', 'd': 'D', 'e:': 'EH', 'f': 'F', 'g': 'G', 'h': 'HH', 'i:': 'IY', 'j': 'Y', 'k':
     'K', 'l': 'L', 'm': 'M', 'n': 'N', 'o:': 'AO', 'p': 'P', 'r': 'R', 's': 'S', 't': 'T',
    'u:': 'UW', 'v': 'V', 'x': 'R', 'y:': 'UW', 'z': 'Z'}
map_ge_to_fr = {'sil': 'sil', '2:': 'o', '6': 'swa', '9': 'swa', '?': 'swa', '@': 'swa', 'C'
    : 'k', 'E': 'e', 'E:': 'e', 'I': 'i', 'N': 'g', 'O': 'o', 'OY': 'o', 'S': 'S', 'U': 'u',
     'Y': 'y', 'Z': 'Z', 'a': 'a', 'a:': 'a', 'aI': 'a', 'aU': 'a', 'a~': 'an', 'b': 'b', 'd
    ': 'd', 'e:': 'e', 'f': 'f', 'g': 'g', 'h': 'swa', 'i:': 'i', 'j': 'j', 'k': 'k', 'l': '
    l', 'm': 'm', 'n': 'n', 'o:': 'o', 'p': 'p', 'r': 'R', 's': 's', 't': 't', 'u:': 'u', 'v
    ': 'v', 'x': 'R', 'y:': 'y', 'z': 'z'}

# store the name of the three languages of study
EN_KEY = "en"
FR_KEY = "fr"
DE_KEY = "ge"

# dictionary containing itself 3 dictionaries for which the key is the reference language
    and the value corresponds to two dictionaries where the key is the language chosen to
    establish the mapping with respect to the reference language and the value is the
    associated mapping dictionary
MAPPING = {
    EN_KEY: {
        FR_KEY: map_en_to_fr,
        DE_KEY: map_en_to_ge
    },
    FR_KEY: {
        EN_KEY: map_fr_to_en,
        DE_KEY: map_fr_to_ge
    },
    DE_KEY: {
```

```
        FR_KEY: map_ge_to_fr,
        EN_KEY: map_ge_to_en
    }
}

def do_mapping(fopen, mapping_dict):
    """ transform a segmentation file of a language A into one of a language B using a dictionary to map phonemes
    input: a segmentation file of a language A and a mapping dictionary
    output: a segmentation file of a language B """

    # open a new file for writing whose name contains the name of the input file
    with open('translated_' + fopen.name, 'w') as fwrite:
        # iterate over lines of the input file
        for line in fopen:
            # transform the line into a list containing the start time, the end time and the associated phoneme
            line = line.strip().split()
            # store the new value of the phoneme created using a mapping dictionary
            new_ph = mapping_dict[line[2]] if line[2] in mapping_dict else line[2]
            # add a new line containing the original duration and the new phoneme in the output file
            fwrite.write(f"{line[0]} {line[1]} {new_ph}\n")

def get_cli_args():
    """create an argument parser, parse and return the arguments"""

    # create the argument parser object
    parser = argparse.ArgumentParser(description="Mapping from a language A to a language B")
    # add optional arguments
    parser.add_argument("--from", dest="LA", help='Language A', required=True, choices=[EN_KEY, FR_KEY, DE_KEY])
    parser.add_argument("--to", dest="LB", help='Language B', required=True, choices=[EN_KEY, FR_KEY, DE_KEY])
    # add positional argument
    parser.add_argument("inputs", help="path to segmentation input files", nargs='+', type=argparse.FileType('r'))
    # parse arguments
    args = parser.parse_args()
    return (args)

def main():
    # store the arguments
    args = get_cli_args()
    # display a message if the same language is chosen for the mapping
    if args.LA == args.LB:
        print("Dude, are you really trying to map a language to the same language ? Don't you think it's useless ? *insert troll_face.jpg*")
    else:
        # retrieve the mapping dictionary using the input arguments
        mapping_dict = MAPPING[args.LA][args.LB]
        # iterate over the selected input files
        for input_seg in args.inputs:
            # display a message to inform the input file on which the program is running
            print(f"Processing {input_seg.name}: ")
            # apply the mapping
            do_mapping(input_seg, mapping_dict)
            # display a message to announce the end of the execution on the file
            print("done")

if __name__ == "__main__":
    main()
```

Listing 1: code



(a) Step 1



(b) Step 2

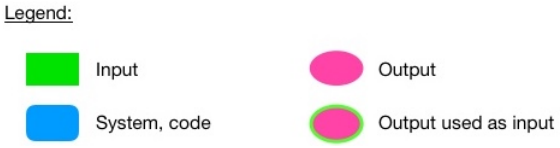


Figure 12: Videos generation

0 680 SIL	0 680 sil
680 770 M	680 770 m
770 900 AY	770 900 a
900 1010 N	900 1010 n
1010 1150 UW	1010 1150 u
1150 1280 K	1150 1280 k
1280 1460 AE	1280 1460 a
1460 1500 T	1460 1500 t
1500 1610 HH	1500 1610 swa
1610 1680 AE	1610 1680 a
1680 1750 Z	1680 1750 z
1750 1820 AH	1750 1820 swa
1820 1930 P	1820 1930 p
1930 1990 EH	1930 1990 e
1990 2040 D	1990 2040 d
2040 2090 AH	2040 2090 swa
2090 2170 G	2090 2170 g
2170 2260 R	2170 2260 R
2260 2400 IY	2260 2400 i
2400 2910 SIL	2400 2910 sil

Mapping code

Figure 13: Mapping code application

1 - Rate the talking head articulation on the scale below:

very bad very good

Valider (Retour à l'étape)

Figure 14: Evaluation part 1

Video 1 Video 2 Video 3

1 - Do you find a difference between the three articulations?

- 1 - **Yes**
- 2 - **No**

2 - In your opinion, which articulation is closest to reality?

- 1 - **Video 1**
- 2 - **Video 2**
- 3 - **Video 3**
- 4 - **Videos 1 & 2**
- 5 - **Videos 1 & 3**
- 6 - **Videos 2 & 3**
- 7 - **All**
- 8 - **None**

3 - In your opinion, which articulation is the furthest from reality?

- 1 - **Video 1**
- 2 - **Video 2**
- 3 - **Video 3**
- 4 - **Videos 1 & 2**
- 5 - **Videos 1 & 3**
- 6 - **Videos 2 & 3**
- 7 - **All**
- 8 - **None**

4 - Comments (if you have no comments, put 0):

Réponse

Valider (Retour à l'étape)

Figure 15: Evaluation part 2

References

[1] Biasutto-Lervat, T., Dahmani, S., & Ouni, S. (2019, September). Modeling Labial Coarticulation with Bidirectional Gated Recurrent Networks and Transfer Learning. In *INTERSPEECH 2019 - 20th Annual Conference of the International Speech Communication Association.* Graz, Austria.

[2] Daniloff, R., & Hammarberg, R. (1973). On Defining Coarticulation. *Journal of Phonetics*, *1*(3), 239 - 248. doi: https://doi.org/10.1016/S0095-4470(19)31388-9

[3] Hannahs, S., & Davenport, M. (2010). *Introducing Phonetics & Phonology.* Third edition, NY, USA: Routledge, 2010. doi: 10.4324/9781351042789

[4] McGurk, H., & MacDonald, J. (1976, 12). Hearing Lips and Seeing Voices. *Nature*, *264*, 746-748.

[5] Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A., ... Matthews, I. (2017, 07). A Deep Learning Approach For Generalized Speech Animation. *Congress of Phonetic Sciences (ICPhS XVIII, Glasgow)., ACM Transactions on Graphics*, *36(4)*, 1-11. doi: 10.1145/3072959.3073699