MSc Natural Language Processing

2020/2021

UE705 EC1 Supervised Project

Evaluation of a Multilingual Animated Speech System

Students:

Juliana DE FERRAN

Sonita TE

Stephanie MONTEIRO

Supervisor:

Slim OUNI

December 2020

Contents

1 Introduction

In the field of audiovisual speech, talking heads have been developed to study the mechanisms of audiovisual speech communication. This report is part of a project led by Slim Ouni aiming to make a talking head multilingual by using a single universal system to animate all languages: we will be joining the team to evaluate the quality of a talking head animated using a system intended for a different language.

So far, the developers in the Loria research laboratory have offered a multilingual version thanks to a set of three independent systems designed for a single language each. However, researchers led by Sarah Taylor argue that it is possible to animate a talking head through a universal system, regardless of the target language: they believe it is possible to have a single system animating any language. This conjecture supposes that coarticulation is the same across all languages, or at least that it has minimal impact on the intelligibility of the final animation.

But how important is coarticulation when it comes to speech animation? Our goal in this report is to study the effects of coarticulation in several languages in order to determine whether this phenomenon manifests itself differently from one language to another or if it is universal. We want to establish whether it is indeed possible to create one system for all languages, or if the quality of the animation is higher when a language has its own dedicated system that follows a precise set of rules.

In the first section of this report, we introduce the area of phonetics, followed by a description of the International Phonetic Alphabet. The second part is devoted to the definition of the coarticulation phenomenon. Then, we discuss the importance of the talking head, as it is a tool that combines all the modalities of speech that are crucial to intelligibility, which we will see in detail. We also explain the different methods used to make it multilingual. The final section is on coarticulation across languages and dialects: we make an inventory of multiple studies led on the realization of coarticulation in different languages in order to test the hypothesis presented previously.

2 Phonetics

2.1 Definition

Phonetics is the area of linguistics that studies the organization of sounds in natural languages. It describes how sounds are produced (articulatory phonetics), how they are transmitted (acoustic phonetics), and how they are perceived by the ear and converted in the brain (auditory phonetics). We have based our research on knowledge from our undergraduate studies and the book by Hannahs & Davenport (2010).

2.1.1 Articulatory Phonetics

This field of phonetics is interested in the way that speech is produced. There are two major agents in human speech production: phonation and articulation. Phonation refers to the vibration of the vocal folds which are opened and closed rapidly to produce sounds by agitating air particles.

Articulation refers to the movement of active articulators (such as the tongue, lips, glottis, etc) and their interaction with passive ones (such as the teeth, or the roof of the mouth), 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.

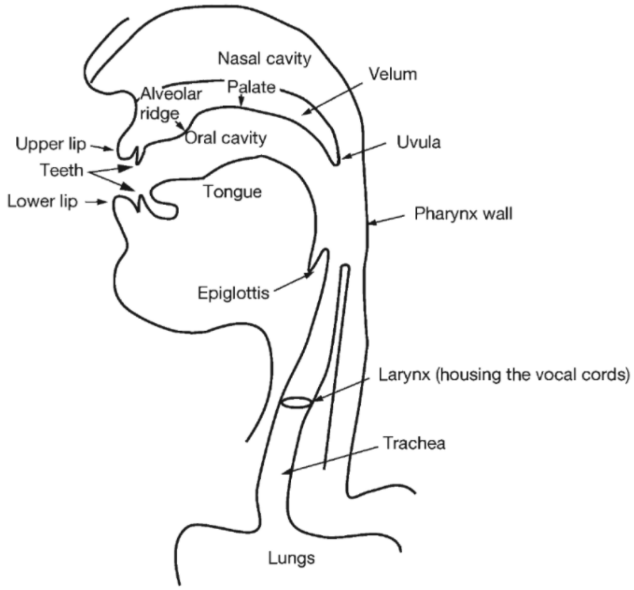


Figure 1: Articulators used to produce sounds of speech (Hannahs & Davenport, 2010)

Different articulators are needed for different sounds, and they perform various movements to create a plethora of sounds. This is what is studied in order to program an animated talking head.

The sounds of the language are generally divided into two classes depending on whether the air flow is disturbed or not: vowels and consonants.

Vowels are characterized by a free passage of air flow through the vocal tract. The classification criteria are the following: the manner of articulation (degree of opening of the jaw), the place of articulation (point at which the articulator approaches in the vocal tract), resonance (passage of air in the nasal cavities or not) and labiality (i.e. protrusion, projection of the lips). Vowels are classified in terms of front, central, or back, which is the horizontal position of the tongue (place of articulation), and closed vs. open, which corresponds to the vertical position of the tongue (manner of articulation).

Consonants are articulated with complete or partial closure of the vocal tract. These kinds of sounds are classified according to the manner of articulation (how the airflow is released from the vocal tract), the articulation place (where the obstruction takes place in the vocal tract), resonance (whether the nasal cavity is associated with the oral cavity), and phonation/voicing (whether the vocal cords vibrate or not).

When it comes to the manner of articulation, here are a few examples:

- Plosives/Stops: Air flow is completely stopped and then released with a slight explosion, which is audible. Examples: /p/, /b/, /t/, /d/, /k/ and /g/.
- Fricatives: Two organs are brought and held sufficiently close together for the escaping airstream to cause local air disturbance or friction, which is audible. Examples: /f/, /v/, /s/, /ʃ/, /z/ and /ʒ/.
- Nasals: There is a complete closure in the mouth, but the velum is lowered, allowing air into the nasal cavity which gives the sound its resonance. Examples: /m/, /n/, /ŋ/ and /ɲ/.

2.1.2 Acoustic Phonetics

Acoustic phonetics deals with waves, amplitudes, and frequency spectrums. In 1877, Thomas Edison was able to record and reproduce sounds: the vibrations were engraved by a needle on a rotating disc, which made the diaphragm vibrate and reproduce the corresponding sound. This model was improved by many successors, until it became what we know nowadays: the representation of sounds in spectrograms (Gelatt, 1955).

The study of vowels' formants in the spectrums makes it possible to know the positioning of the tongue in the oral cavity. The formant of a sound corresponds to one of the energy maxima of the sound spectrum which results from the acoustic resonance of the human vocal tract. The first formant refers to the vertical position of the tongue (i.e. manner of articulation) while the second corresponds to its horizontal position (i.e. place of articulation). When the F1 is at a low value, the tongue is placed high, and when the F2 is at a low value, the back of the tongue is at the back of the oral cavity.

2.1.3 Auditory Phonetics

This branch focuses on speech perception: how pitch, intensity, quality, and duration of information might impact the way that it is parsed. It studies the processing of speech in segments (often corresponding to vowels and consonants) within the context of each other. It also studies prosody, which deals with pitch, duration, and energy of each syllable.

The modality of phonetics that interests us the most in this project is the one that studies the process of articulation. We will look in detail at a specific part of it: coarticulation, which deals with how sounds affect each other during pronunciation.

2.2 IPA & SAMPA

Phonetics heavily relies, and this is what is of most interest to this project, on phonetic symbols, grouped under the International Phonetic Alphabet. The IPA, developed in 1888, is a system based on the Latin alphabet that allows a transcription of sounds and stressed syllables in speech thanks to not only letters but also diacritics. Indeed, the latin alphabet by itself is ill-suited to represent sounds insofar as:

- For example, several characters can represent the sound [ʃ] in English: such as 'ch', or 'sh', but it is transcribed by a single symbol in IPA.
- In the same language, the same sound can be encoded in several ways, as in for example *j̲amais* and *g̲entil* in French. The API uses the unique code [ʒ] for this sound.

- In several languages, the same sound can also be encoded in several ways: for example, 's' in Hungarian and 'sz' in Polish represent the sound [ʃ], symbol assigned by the API in a more homogeneous manner across all languages.

Thus, each distinctive sound (i.e. phoneme), that serves to distinguish one word from another has a paired symbol. These symbols can be found in the IPA chart in the appendix of this report.

In more detail, a phoneme is an abstract representation that groups together all possible concrete realizations called phones (synonymous of the term *sound*), whose characteristics depend on the context and on different factors such as the origin, sex, and age of the speaker. The phoneme is transcribed by symbols placed between slashes while the phones are represented with square parentheses.

To better understand, let's take the example of the phoneme /r/ in the French word *rat*. It can be produced in several ways: rolled [r], burred [ʀ], or even normal [ʁ]. Thus, [r], [ʀ] and [ʁ] are phones of the phoneme /r/ in French. In phonemic transcription, such details are left out.

Not all languages use all of the IPA phonemes. For example, French has 37 phonemes against 44 for English, and 68 for German.

While the IPA is very useful in the study of phonetics, it has limitations when it comes to computational linguistics: there are many special characters that are not available on a regular latin keyboard and that cannot be easily parsed by a machine. This is where the Speech Assessment Methods Phonetic Alphabet (SAMPA) comes in.

SAMPA, which has been expanded since the 1980s in the form of the X-SAMPA, is an encoding of the IPA into ASCII characters that are readable and printable by a machine. Latin symbols from the IPA are kept, but those that cannot be, such as for example the "ə" or the "ʒ" symbols, are transformed into characters readily available on QWERTY and derived keyboards (uppercases, "∼", "@" for example) (Wells, 1997).

Long vowels		Short vowels		Consonants	
SAMPA	IPA	SAMPA	IPA	SAMPA	IPA
a:	aː	a	a	b	b
e:	eː	E	ɛ	d	d
i:	iː	I	ɪ	g	ɡ
o:	oː	O	ɔ	l	l
u:	uː	U	ʊ	x	x
E:	ɛː	Y	ʏ	r	χ, ʁ
2:	øː	9	œ	m	m
y:	yː	@	ə	n	n
		6	ɐ	s	s

Figure 2: A few examples of IPA to SAMPA character conversion (Birkholz, 2013)

In this project, SAMPA was used. The audio given to the program is accompanied by a text, which is segmented and transformed into phonemes so that the talking head may know which articulation comes at what point in the sentence, and its duration.

Another limitation of the IPA alphabet is the fact that it does not deal with coarticulation: it transcribes one phoneme after the other without thought as to how they might affect one another.

3 Coarticulation

Coarticulation is the concept according to which the articulation of phonemes surrounding the one being pronounced might affect its characteristics given the way the articulators move from one sound to the other. Language is not a simple string of phonemes, but rather a concatenation of interactions between them.

A concept close to coarticulation that must not be confused with is assimilation: it speaks only of the way a phoneme is distorted by another, while coarticulation explains the changes in the movements of the articulators which then create changes in pronunciation. In this study, we focus solely on coarticulation, as it is the one that has the most impact on the visual aspect and therefore on the animated head.

Daniloff & Hammarberg (1973) posit that it is the context that defines the way in which phonemes are pronounced:

> "At the articulatory level, segments are not separately and independently articulated. Rather they are coarticulated, and neighboring segments overlap and affect each other in various ways." (Daniloff & Hammarberg, 1973: 239)

Kent & Minifie (1977), in their article where they discuss how to model coarticulation, explain the reason behind this phenomenon:

> "Some overlapping of articulatory movements is inevitable, given that the speech organs are not capable of infinite acceleration. The vocal tract cannot change instantaneously from one configuration to another, and the articulatory transitions between sounds will therefore reveal interactive influences." (Kent & Minifie, 1977: 118)

Coarticulation is due to the fact that the muscles used to pronounce phonemes do not reset themselves to a neutral position after each segment, but move from one to the next in a continuous manner, thus creating a fluid positioning.

As Volenec (2015) explains, we speak of anticipatory coarticulation when a phoneme following the current one impacts the way that it is pronounced, and perseverative coarticulation wherein the remnants of the pronunciation of a previous segment impact the pronunciation of following ones. They are also described as forward/right-to-left, and backward/left-to-right respectively.

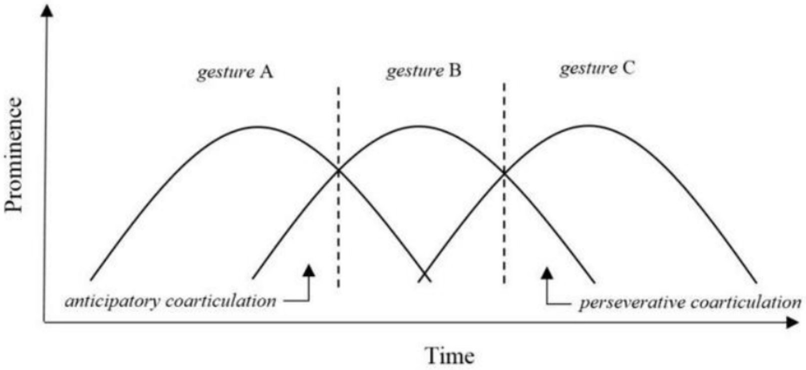


Figure 3: Anticipatory and perseverative coarticulation (Volenec, 2015)

Kent & Minifie also explain that it is not only neighboring phonemes that affect each other, but that it is also sometimes possible to see the articulators prepare themselves to produce a phoneme up to 7 segments before its actual realization in English.

Given the fact that coarticulation is one of the key factors that influence the quality of the visual information transmitted, it is an important component to take into consideration when animating.

4 Speech Animation

In this section, we discuss the value of using a talking head to model speech communication. Next, we outline two methods to make it multilingual, the first of which is to use a set of systems designed for a single language each, while the second consists in creating a system independent of the target language that would therefore work for any input.

4.1 Multimodality of Speech

In 1993, Denes & Pinson developed the concept of the speech chain to describe the different stages of the speech communication connecting the speaker and the listener (see FIGURE 4). These actors each have their own tasks. On one hand, the speaker has to translate his thoughts into speech and produce intelligible sounds. On the other hand, the listener has to hear, interpret, and understand what is being transmitted.

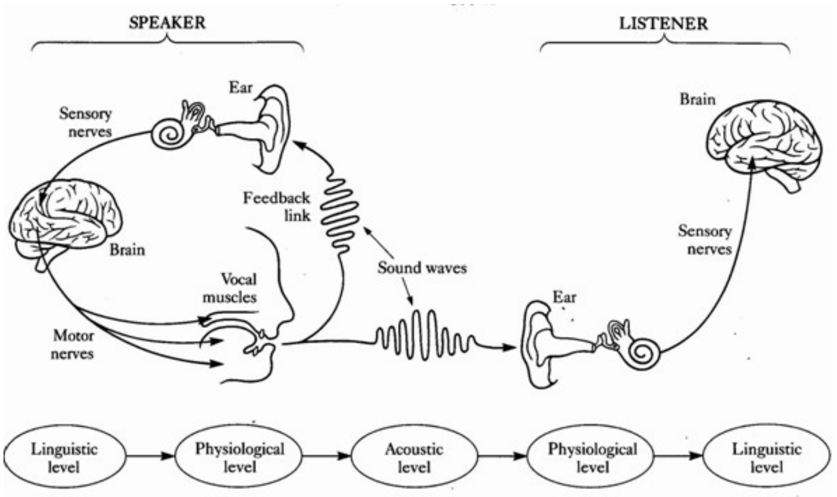


Figure 4: The speech chain by Denes & Pinson (1963)

First of all, the speaker transforms the message they want to convey into linguistic units and orders them according to the grammatical rules of the language they are speaking in (linguistic level). Then, the motor nerves communicate signals from the brain to the muscles that activate articulatory gestures made by the various organs of speech (physiological level). The movements of vocal muscles generate an accompanying sound wave (acoustic level). The listener perceives the sound signal through their ear (physiological level), which their sensory nerves then communicate to the brain in order to interpret it as linguistic units and understand them (linguistic level).

These researchers consider the speech communication to be multimodal: the movements of the different organs of speech correspond to the articulatory modality and the perception of the acoustic signal represents the auditory modality. According to them, these are the only two modalities that help understand speech communication.

Two previous studies, one conducted by McGurk & MacDonald in 1976 and the other by Sumby & Pollack in 1954, demonstrated the effect of a visual support on speech intelligibility.

Sumby & Pollack (1954) showed that the presence of visual information (such as the speaker's lips and facial movements) in addition to audio information improves speech intelligibility in a noisy environment. Note that the size of the vocabulary also has an impact. For example, on average, out of 8 words heard with a speech-to-noise ratio of 30 dB, about 15% of the words were correctly recognized with audio only against over 90% with audio and additional visual cues. They also found that the louder the noise, the more the listener relied on what they saw to better understand the audio.

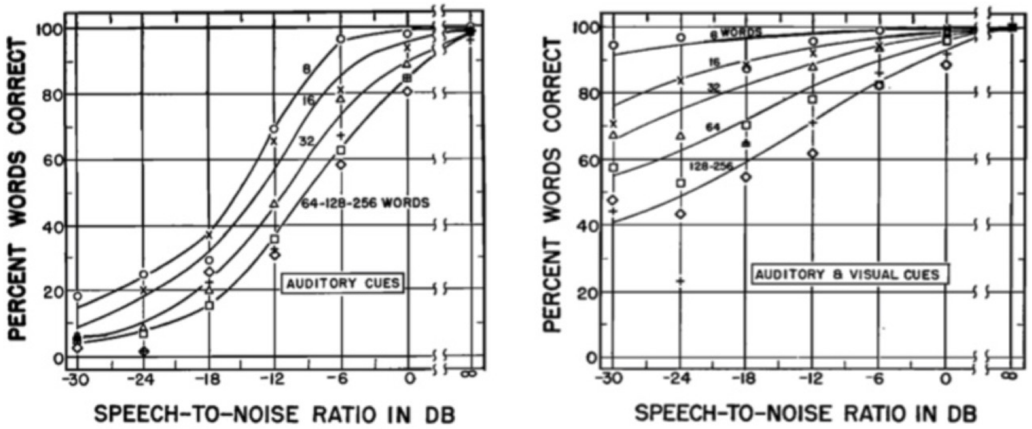


Figure 5: Speech intelligibility as a function of speech-to-noise ratio: in the first case, only with auditory cues and in the second case, with auditory and visual cues (Sumby & Pollack, 1954).

The McGurk effect presented in the article *Hearing Lips and Seeing Voices* reveals that when visual input is imprecise (or outright misleading), speech intelligibility is impacted. Indeed, faced with the association of the audio sequence [baba] with the visual film corresponding to the sequence [gaga], the subjects reported hearing [dada]. To conclude, the visual must be as close as possible to reality to ensure the intelligibility of the listener. (McGurk & MacDonald, 1976).

A talking head is an animated virtual 3D face of a human speaker in sync with audio. This tool includes the two main segments of the human speech system for producing sounds: phonation and articulation. The audio sequence represents the phonation, i.e. the mechanism of vibration through the vocal folds which are opened and closed quickly to produce the sounds. The visual of the head allows to obtain the facial deformation generated by the articulation which refers to the movement of the organs of speech such as the lips, jaw, tongue, etc.

The interest of the talking head is to facilitate the understanding of the listener thanks to the pairing between the visual and the audio. This would be especially useful for hard-of-hearing audiences.

4.2 Multilingual Systems

4.2.1 A System for Each Language

In the field of audiovisual speech, talking heads have been developed by the Multispeech team in the Loria laboratory. They worked on an original method to synchronize the movement of the lips of a 3D face with a given speech audio (Biasutto-Lervat et al., 2019).

The team has worked with three different languages: English, French, and German. Their current belief is that every language follows a very specific set of linguistic rules that cannot be homogenized to create a universal program to fit all languages. This system is therefore composed of three different ones that each animate a single language.

In the current process, an audio file is input into a program along with its phonetic transcription. The program segments the phonenemes according to duration and intensity. That new information is given to the animation program that controls the movements of the talking head.

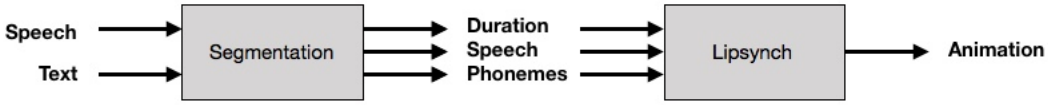


Figure 6: Animation system of a given language

In *Modeling Labial Coarticulation with Bidirectional Gated Recurrent Networks and Transfer Learning* (2019), Biasutto-Lervat et al. describe how RNNs were used to deal with consecutive data by storing past data and current contributions in order to decide on the outputs.

Another project that dealt with creating a talking head to aid in the understanding of audio information was that of Cohen & Massaro (1999), wherein they discussed the strategies to make the animation of speech realistic. They constructed the muscular and bone structure of the face, then added a tongue.

As for the occurrence of coarticulation in human speech to help in animation, Löfqvist's gestural theory (1990) has been described in their research:

> "A speech segment has dominance over the vocal articulators which increases and then decreases over time during articulation. Adjacent segments will have overlapping dominance functions which leads to a blending over time of the articulatory commands related to these segments" (Cohen & Massaro, 1999)

They worked on a system based solely on English and its characteristics: lip protrusion and shaping, muscle contraction, tongue movements pertaining solely to English phonetics. They concluded that the closer the features of the animated face are to reality, the more accurate it seems, meaning that they would be adding teeth, as well as variations to the movements of the jaw and cheeks.

4.2.2 A Universal System

Taylor et al. (2017) posit that one single system could work for any language. Working with Deep Learning and Neural Networks, their team worked on a program that could replace both manual animation of speech and performance capture animation.

They trained their model in a way that made it easy to deploy and to adapt to new characters' physical features, and claimed they could edit and change the animation style in their software. They claim that their work can be used for "any speech content" and with foreign languages.

They work around the problem of coarticulation by training their model to be able to recognize and reproduce the effects of coarticulation when given the context. They work based on the assumption that

> "[...] coarticulation effects are localized, and do not exhibit very long range dependences. For example, how one articulates the end of *prediction* is effectively the same as how one articulates the end of *construction*, and does not depend (too much) on the beginning of either word." (Taylor et al., 2017)

Therefore, they assume that training their model to reproduce sections of articulation instead of looking at the words and sentences as a whole would be enough to circumvent the accuracy problems created by coarticulation.

5 How Does Coarticulation Work...

Coarticulation is quite problematic when trying to reproduce human articulation as closely as possible, as it transforms animation into the task of recognizing phonemes but also understanding how they affect each other in context. We must therefore analyze different studies made on coarticulation and to see in what different languages have different rules.

5.1 ... In Different Dialects of the Same Language?

Tamminga & Zellou (2015) undertake this comparative task at a small scale: they contrast the differences in articulation in two dialects of American English. They studied two small corpora, one from the Philadelphia area and another from Ohio. They focus on observing variation in monosyllabic words by analyzing terms where nasal consonants were preceded by a vowel. They find that by dividing the population according to gender (male vs. female) and age (young vs. old), the data had clear discrepancies between strata.

Older Philadelphia women were noticed to have less nasal coarticulation than men, and the younger generation seemed to be following their lead. On the other hand, when looking at the population in Ohio, they notice that it is young women who are driving the change: there is less nasal coarticulation in young women and there are in the older generation (regardless of gender) and than young men. They conclude that the gradual shift towards the reduction of nasal coarticulation does not follow a predefined form but there is meaningful variation following unanticipated patterns given gender and age.

What is important for our analysis in this study is the fact that, even within the same language, it appears that coarticulation does not have the same characteristics given different dialects, different genders, and different age ranges.

5.2 ... In Different Languages?

Solé (2018) studies voiced stops in Spanish, French, and English. She is interested in how speakers of different languages might have different methods of articulation for the same sound. In this study, the sounds looked at are /b/, /d/, /p/, /t/, and /m/. What interests her are the results for voiced stops, so she compiled a corpus of native speakers of the three languages to see how they position their articulators, how they regulate oral pressure, aerodynamics, and voicing amplitude.

After thorough analysis of the spectrograms, graphs, and the results observed, she posits that the three languages do indeed have differing voicing adjustments, meaning that the same sound is not produced with the same strategy in every language. Indeed, some speakers will use nasal leaks or advance the tongue's root further then others. It is also important to note that within the same language, speakers also utilized different gestures according to regional accents (for example Australian compared to American pronunciation).

As for vocal folds vibrations, English speakers seem to have multiple approaches to articulation (including either a vibration starting during stop closure, or sometimes no vibration at all), while romance languages tend to be homogeneous in their strategy and always have vocal fold vibration.

Through this study, we notice that different languages have different articulation strategies and that they do not follow a universal set of coarticulatory rules.

Torreira & Ernestus (2011) compared the realization of voiceless stops in an intervocalic context in both French and Spanish. Their measurements revealed deferring performances of these sequences in each language. Both consonants and vowels created this heterogeneity in languages: consonants had different durations, types of closure (complete/incomplete), and degrees of voicing. As for vowels the differences were found in the devoicing, duration, as well as the value of the first formant of the vowel.

The results show that French and Spanish speakers use different coarticulatory strategies for the realization of these segments: Spanish speakers had shorter stop closures and more voicing, while French speakers had more cases of complete devoicing and longer vowels.

Néron (2011) was particularly interested in American English, German, and French diction. He looked at the phonetic variation depending on adjacent phonemes, for which the IPA is ineffective. The goal was to identify and describe the acoustic similarities and differences between the three languages in a bid to simplify the teaching of a second language to foreign students. Néron demonstrates that while phonetic symbols are the same for all the languages studied, the associated acoustic characteristics are different. For example, he explains:

> "Both German /u/ and /o/ occupy an acoustic space unfamiliar to an English speaker, on the far end of the high back vowel spectrum. Minimal coarticulation effect adds to the difficulty for English speakers to achieve an accurate rendition of both vowels. In addition of being more back, German /o/ is also much less open than its English counterpart" (Néron, 2011)

We can see that, while the IPA symbol is the same, the manner of articulation and the accompanying movements of it differ: distinct spaces of pronunciation and opening of the mouth in this example, or the vowel /y/, which is pronounced further back in French than it is in German, and most resembles the English /u/ in manner of articulation.

Strange et al. (2007) worked on the acoustic variation of vowels in English, French, and German by looking at the formants F1 and F2 from a loudness perspective using the Bark scale (which measures frequency in Hz). They selected a small group of native speakers of both genders from each dialect and had them recite a pre-set list of terms with the same phonemes in each language. For example, for /i/, German participants were made to articulate the word *Hieba*, while French speakers said *Hibe*, and English speakers *Heeba.*

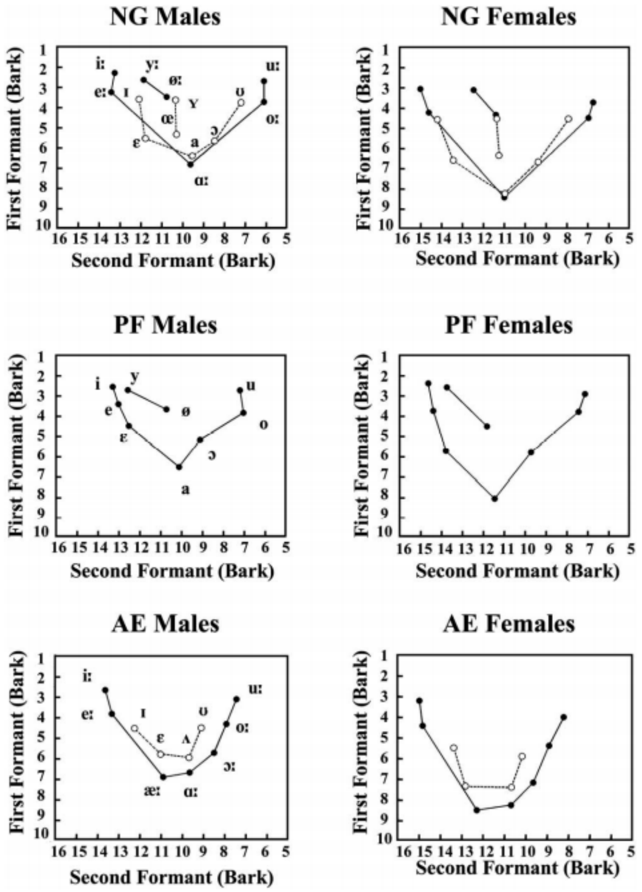


Figure 7: Average mid-syllable F1/F2 values for North German, Parisian French, and New York English values according to gender (Strange et al., 2007)

They noticed a difference in tongue positioning (be it vertical or horizontal) when it came to the articulation of vowels, not only between languages, but also, within languages, accoridng the gender of the speaker. They conclude that mid-syllable F1 and F2 values and the duration of the phonemes were acoustically different in all three languages studied.

This is interesting to us as it helps us to better understand the limitations of the IPA when it comes to coarticulation and phonetic variation: while the IPA symbol might be the same, the way that it is articulated varies from language to language. It furthers the hypothesis that there cannot be a universal talking head system.

6 Conclusion

So is it possible to have a universal program of phonetic animation? We initially asked ourselves how important coarticulation was when it came to the animation of a talking head, and tried to answer that question by looking at how speech is produced by humans, how that process can be faithfully animated, and how coarticulation works in different languages and their dialects.

The studies researched conclude that there are differences in coarticulation not only between languages but also within languages (according to different accents and dialects, and different genders). Thus, we ourselves can conclude that, when it comes to animating a high quality talking head, the movements it creates would not be the same for every language as there are clear differences in phonation, articulation, and acoustic spaces.

Coarticulation can be problematic to the veracity of an animated model as speech is multimodal: we look at the auditory and visual modalities as a pair, meaning that if one of them is not exactly matched to the other, it might give the viewer the feeling that the product is not finished and the quality is low. For example, when watching a movie, if the audio and video are not perfectly synchronized, the viewer might experience feelings of disorientation.

In the case of our model, if the visual representation does not match what is being said, it would not be useful for hard-of-hearing audiences and would not give a good result in animation/video game development. As this model can also be used to teach students a foreign language, it would be counter-intuitive to teach them using a standard that is not of a native-speaker quality.

When animating a language, one single dialect can be chosen, so that its specific are rules followed and the product becomes be acceptable to native speakers viewing it. When it comes to multiples languages in one model, we assume that the quality of the articulation and its exactitude would lower due to the effects of coarticulation.

7 What Comes Next?

In the second part of our research, we will be establishing a map between languages. Our current model is an amalgamation of three systems that work on one language each: English, French, and German. As it stands, the speech and its corresponding transcription of Language A are segmented and treated by the animation program of Language A.

Our upcoming task will be to transform the phonemes in the segmentation of Language A into a transcription that the animation program of Language B can understand.

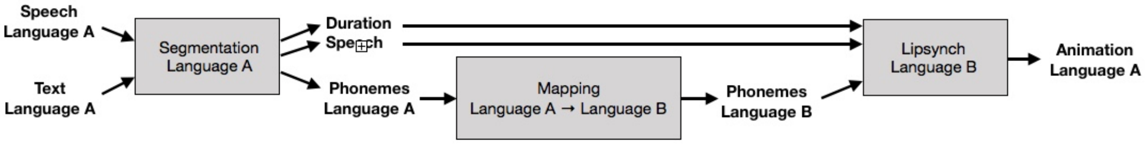


Figure 8: Animation of Language A using the system of Language B

To accomplish this, we will be looking at the phonemes from both languages and mapping them according to how they are articulated. Some phonemes such as /t/ and /b/ could potentially be the same across all three languages, but others (such as /θ/ in English) do not have a direct correspondence in French. We will therefore need to understand how it is articulated in the original language (tongue placement, lip rounding and/or protrusion) and find the phoneme with the same (or as close as possible) characteristics in the other languages.

This correspondence mapped from Language A will then be given to the system of language B so that it can be animated. This is to see whether articulation rules are interchangeable between languages or if results are of the best quality when each language has its own animation system.

We will be evaluating the results by asking native speakers to appraise the quality of the following talking heads: German animated by the French and English systems, French animated by the German and English systems, and English animated by the French and German systems. These results will help us make our own conclusion on the effects of coarticulation and how different it really is in each language.

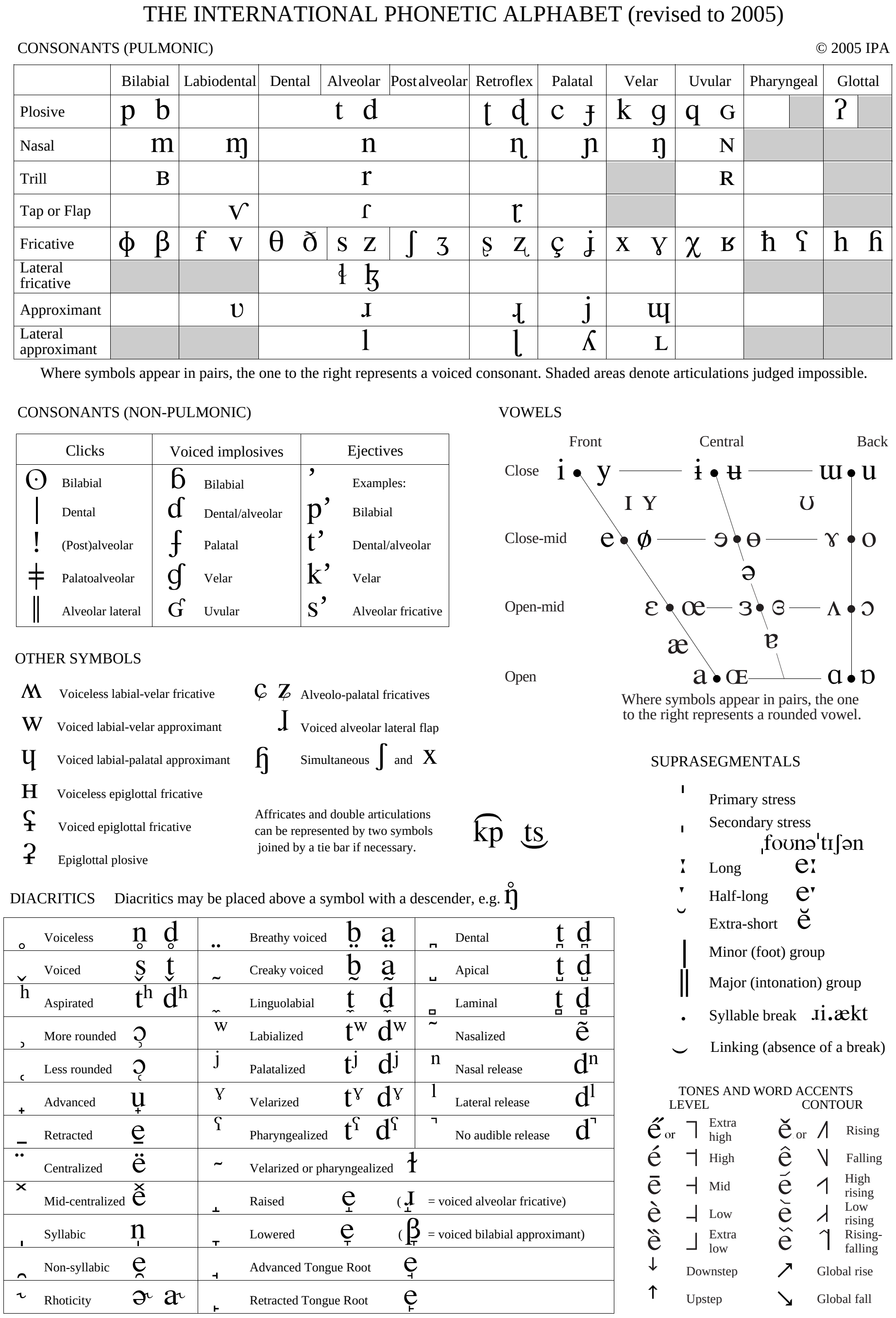
8 Personal Motivation

We chose to work on this project because we felt that it perfectly merged the two fields we are studying: linguistics and computer science, adding the animation of speech to form a well rounded project that could teach us many new and interesting things. As NLP students, we are very interested in speech and image processing, which we will get to work with during the second part of this project.

We have learnt more about phonetics by looking in depth into different concepts and theories of articulation and coarticulation, and how we may subconsciously rely on visual information to help us understand audio. It has become even more clear to us the importance of merging linguistics and computer science and how they depend on each other to advance NLP research.

We would like to thank our supervisor, Slim Ouni, for taking his time to explain what his team has done so far and taking us into the fold so seamlessly. We would also thank the University of Lorraine and the Loria research laboratory for partnering in this manner and thus giving us the opportunity to learn more about NLP and the world of academia.

9 Appendix



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

© 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap		ⱱ		ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ						▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	ʼ	Examples:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

VOWELS

	Front		Central		Back
Close	i • y		ɨ • ʉ		ɯ • u
		ɪ ʏ		ʊ	
Close-mid	e • ø		ɘ • ɵ		ɤ • o
			ə		
Open-mid		ɛ • œ	ɜ • ɞ		ʌ • ɔ
		æ	ɐ		
Open		a • ɶ			ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricatives

ɺ Voiced alveolar lateral flap

ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. k͡p t͡s

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
˞	Rhoticity	ɚ a˞	◌̙	Retracted Tongue Root	e̙			

SUPRASEGMENTALS

ˈ Primary stress

ˌ Secondary stress

ˌfoʊnəˈtɪʃən

ː Long eː

ˑ Half-long eˑ

◌̆ Extra-short ĕ

| Minor (foot) group

‖ Major (intonation) group

. Syllable break ɹi.ækt

‿ Linking (absence of a break)

TONES AND WORD ACCENTS

LEVEL			CONTOUR		
e̋ or	˥	Extra high	ě or	˩˥	Rising
é	˦	High	ê	˥˩	Falling
ē	˧	Mid	e᷄	˦˥	High rising
è	˨	Low	e᷅	˩˨	Low rising
ȅ	˩	Extra low	e᷈	˧˦˧	Rising-falling
ꜜ		Downstep	↗		Global rise
ꜛ		Upstep	↘		Global fall

Appendix 1: IPA chart (2005)

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] Birkholz, P. (2013, 04). Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis. *Processing. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.*, *8*, e60603. doi: 10.1371/journal.pone.0060603

[3] Cohen, M., & Massaro, D. (1999, 03). Modeling Coarticulation in Synthetic Visual Speech. *In: Thalmann N.M., Thalmann D. (eds) Models and Techniques in Computer Animation. Computer Animation Series. Springer, Tokyo.*. doi: 10.1007/978-4-431-66911-1_13

[4] 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

[5] Denes, P. B., & Pinson, E. N. (1963). The Speech Chain: the Physics and Biology of Spoken Language. New York, N.Y: W.H.Freeman & Co Ltd; 2nd Revised edition, April 1993.

[6] Gelatt, R. (1955). *The Fabulous Phonograph: from Tin Foil to High Fidelity.* Philadelphia: J. B. Lippincott Company, 1955.

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

[8] Kent, F. D., R. D. Minifie. (1977). Coarticulation in Recent Speech Production Models. *Journal of Phonetics*, *5*(2), 115 - 133. doi: https://doi.org/10.1016/S0095-4470(19)31123-4

[9] Löfqvist, A. (1990). Speech as Audible Gestures. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 289–322). Dordrecht, Netherlands: NATO ASI Series (Series D: Behavioural and Social Sciences), vol 55. Springer, Dordrecht.

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

[11] Néron, M. (2011). Coarticulation: Aspects and Effects on American English, German, and French Diction. *Journal of Singing; Jacksonville Vol. 67, N° 3, (Jan/Feb 2011)*, 313-325.

[12] Solé, M.-J. (2018). Articulatory Adjustments in Initial Voiced Stops in Spanish, French and English. *Journal of Phonetics; Vol. 66, January 2018*, *66*, 217 - 241. doi: https://doi.org/10.1016/j.wocn.2017.10.002

[13] Strange, W., Weber, A., Levy, E., Shafiro, V., Hisagi, M., & Nishi, K. (2007, 09). Acoustic Variability Within and across German, French, and American English vowels: Phonetic context effects. *The Journal of the Acoustical Society of America*, *122(2)*, 1111-29. doi: 10.7916/d8-arq2-7m09

[14] Sumby, W., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, *26(2)*, 212-215.

[15] Tamminga, M., & Zellou, G. (2015, 08). Cross-dialectal Differences in Nasal Coarticulation in American English. Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS XVIII, Glasgow).

[16] 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

[17] Torreira, F., & Ernestus, M. (2011, 01). Realization of Voiceless Stops and Vowels in Conversational French and Spanish. *Laboratory Phonology*, *2*, 331-353. doi: 10.1515/labphon.2011.012

[18] Volenec, V. (2015, 01). Coarticulation. In (p. 47-86). Nova Science Publishers, Inc, Phonetics, Chapter: 2.

[19] Wells, J. (1997). *SAMPA computer readable phonetic alphabet.* In Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.