Linguistic Portfolios

Volume 11 Article 3

2022

A TUTORIAL ON FORMANT-BASED SPEECH SYNTHESIS FOR THE DOCUMENTATION OF CRITICALLY ENDANGERED LANGUAGES

Ettien Koffi
St. Cloud State University

Mark Petzold
St. Cloud State University

Linguistic Portfolios – ISSN 2472-5102 –Volume 11 – 2022| 26

A TUTORIAL ON FORMANT-BASED SPEECH SYNTHESIS FOR THE DOCUMENTATION OF CRITICALLY ENDANGERED LANGUAGES

ETTIEN KOFFI AND MARK PETZOLD[1]

ABSTRACT

Smaller languages, that is, those spoken by 5,000 people or less are dying at an alarming rate (Krauss 1992). Many are disappearing without having been studied acoustically. The methodology discussed in this paper can help build formant-based speech synthesis systems for the documentation and revitalization of these languages. Developing Text-to-Speech (TTS) functionalities for use in smart devices can breathe a new life into dying languages (Crystal 2000). In the first tutorial on this topic, Koffi (2020) explained how the Arpabet transcription system can be expanded for use in African languages and beyond. In the present tutorial, Author 1 and Author 2 lay the foundations for formant-based speech synthesis patterned after Klatt (1980) and Klatt and Klatt (1990). Betine, (ISO: 639-3-eot), a critically endangered language in Côte d'Ivoire, West Africa, is used to illustrate the processes involved in building a speech synthesis from the ground up for moribund languages. The steps include constructing a language model, a speaker model, a software model, an intonation model, extracting relevant acoustic phonetic data, and coding them. Ancillary topics such as text normalization, downsampling, and bandwidth calculations are also discussed.

Keywords: Speech Synthesis, Formant Synthesis, Formant Extraction, Speech Coding, Sampling Rate, Endangered Languages, Formant Bandwidths, Betine, Language Model, Speaker Model, Intonation Model.

1.0 Introduction

According to Rabiner and Schafer (2011:907), humans have been trying to build machines that speak for more than 200 years. The early experiments were abject failures, but the drive never waned. Incremental improvements have resulted in present-day high-performing systems such as Alexa, Google Voice, and Siri. Our goal in this paper is to harness the breakthroughs in speech synthesis and use them for the documentation and revitalization of critically endangered and under-documented languages. The current tutorial builds upon Koffi (2020) and explains the steps that one can follow to build a Text-to-Speech (TTS) synthesis for dying and lesser-known languages. The tutorial is divided into five parts. The first explains what goes into a language model, the second focuses on the speaker model, the third highlights the software model, the fourth extracts relevant acoustic phonetic measurements, and the fifth is devoted to the intonation model without which a naturally sounding synthesized speech is hard to achieve. Betine, ISO 639-3-eot, a moribund language spoken in Côte d'Ivoire, West Africa, is used to illustrate and exemplify the steps needed to build a speech synthesis system from the ground up.

2.0 Statement of the Problem

Linguists, anthropologists, and language activists have used and continue to use a variety of methods to document critically endangered languages. Koffi (2021) reviewed these methods and highlighted their strengths and weaknesses. The strengths are many because they

[1] **Authorship responsibilities:** This paper is the result of an inaugural *Speech Signal Processing* course that Authors 1 and 2 taught in Spring 2021. Most of the students were engineers with little to no background in linguistics and a handful of linguistics majors. The linguistic analyses and explanations are by Author 1. Coding and script demonstrations are by Author 2.

help preserve aspects of languages that will otherwise die without leaving a trace in the annals of linguistics. In spite of their usefulness, current methods share one glaring flaw, that is, they document endangered languages statically. By this, we mean that after the IPA transcriptions, the recordings, the annotations, the videos, the palatographs, the articulatory tracings, etc., are done, one cannot use the results to generate novel utterances in the language. Our contention is that speech synthesis can and should be added to existing methodologies so that endangered languages can be documented dynamically. This means that, because of documentation via synthesis, the functionalities of critically endangered languages can be expanded to include TTS. This aligns perfectly with Crystal's (2000:141) Postulate 6 which is stated as follows:

> *An endangered language will progress if its speakers can make use of electronic technology*.

We believe that moribund languages can be saved from extinction if they are digitalized for use in smart phones which are now pervasive even in the remotest corners of the globe. If the capabilities of severely endangered languages are augmented to include TTS, language learning apps, and indigenous gaming apps, some of them will not only survive but they will thrive again. This tutorial aims at laying the foundations for what makes speech synthesis possible. The topics covered include making a language model, choosing a speaker model, selecting a software model, coming to grips with the intonation model, and basic speech coding. Other ancillary topics that are discussed include bandwidth measurements and downsampling.

2.1 Building a Language Model

Every speech synthesis system needs a "language model." The goal of the language model, according to Rabiner and Schafer (2011:967-968, 970), is "to design a grammar to represent and include every legal sentence in the task language, while eliminating every non-legal sentence from consideration." This is a tall order that requires an in-depth analysis of the language under consideration. The know-how required to build a language model for speech synthesis is beyond the grasp of a single individual. English,[2] French, Spanish, Portuguese, Russian, Mandarin, Japanese, Korean, etc. have robust TTS systems because thousands of research hours and millions of dollars have been devoted to building rich and expansive databases. For under-documented or critically endangered languages, the language model is inexistant or woefully inadequate for speech synthesis. For this reason, a language model must be built from the ground up. This requires interdisciplinary collaboration. Author 1 is an acoustic phonetician. Author 2 is a signal processing engineer. They have teamed up to teach a course on how to build speech synthesis systems for under-documented languages. They have chosen Betine (ISO: 639-3-eot) as their experimental language. It is one of 10 critically endangered languages in Côte d'Ivoire, West Africa. *Ethnologue* (2019:125) gives Betine an Expanded Graded Intergeneration Disruption Scale (EGIDS) of 8a, which means that it is a **moribund language**: "the only remaining active users of the language are members of the grandparent generation and older." The last known monolingual speakers of the language were a man known only by his first name Angoran and another one named Émile Aiko Etchoua. The former passed away probably in 1967 or shortly thereafter (Perrot 2008:14) and the latter sometime after 1983 (Hérault 1983:403, Perrot 2008:18-19). Betine is not the only language in this undesirable state. *Ethnologue* (2019:35) estimates that 2,923 languages, that is, almost half of the world's languages are on the brink of extinction.

[2] TTS is an ongoing process even in English. Rabiner and Schafer (2011:925) note that there are 1.7 million surnames in the US which need attention.

In building a language model, one provides some basic genealogical information about the language under consideration. As far as its linguistic ancestry is concerned, Betine belongs to the Niger-Congo family of languages, to the sub-family called Kwa, and to the Akan sub-subfamily. Within Akan, it belongs to the Nzi-Tano subgroup. It is further classified into the smaller subfamily of Lagoon languages. For under-documented languages such as Betine, the language model should contain information about how the speakers define themselves. The following is a cosmological narrative about the origins of the Betine:

A Story about the origin of the Beti People[3]

We know that most of the people who live in Côte d'Ivoire came from Ghana. Our ancestors have told us that the Betibe came from subterranean waters. They came from the depths of subterranean waters and established themselves on firm ground. The person who led them from the underground waters to live on dry land is called Ngbandji Ohouman. After they sprang up from the waters, they lived in a place called Monobaka, which was a big village on a large island. Two people ruled that village. One was called Wopou Siguin, and the other, Wopou Nigbeni. Wopou Siguin and Wopou Nigbeni ruled the village until the time when the Agni arrived from Ghana. They came and found the Eotile and waged war against them. The war raged on until they defeated the Eotile. Finally, the Agni chased them out of their land. The Eotile were chased out of their village because the two brothers, Wopou Siguin and Wopou Nigbeni, no longer saw eye to eye because of a woman. The two fell in love with the same woman. For this reason, when the Agni were waging war against Wopou Siguin's troops, his brother did not come to his rescue. So, the Agni defeated Wopou Siguin. Wopou Siguin and his warriors fled and were scattered about. Some went to Ghana, some settled in the town of Adiake, some went to live in a suburb now nicknamed "France." The people who went to live there, are presently in the village of Vitré.

This narrative helps to uncover the history of war, occupation, and colonialism that have contributed to the demise of the Beti people and their language. Except for the mythical subterranean story that cannot be verified historically, the wars that the Anyi waged against the Beti people took place in 1725 (Perrot 2008:27). These events were well documented by French missionaries and merchants who were already living among the Beti people as early as 1701 (Perrot 2008:37).

2.2 Review of the Literature on Betine

In building a language model for an under-resourced language, one should endeavor to find everything that has been written about the structures of the language itself. We have searched high and low in order to find materials that we can use to build a good language model for Betine. To the best of our knowledge, there are only five sources. The first is Perrot's 2008 book entitled: *Les Éotilé de Côte d'Ivoire aux XVIII^e^ et XIX^e^ Siècles: Pouvoir Lignager et Religion*. It is an anthropological study of the political and religious system of the Betine people. They are officially known in Côte d'Ivoire as *Éotilé*, a name most likely given to them by the Anyi who defeated them (Perrot 2008:14). However, they refer to themselves as **Betibe** and to their language as **Betine**. Out of respect for them, we use the word Betine even though the Ivorian government knows them as "Éotilé."

[3] This story about the origin of the Beti people was recorded on Monday, May 2001 at Vitré 2, village in the county of Grand Bassam, which is the county seat. The storyteller was Honorable Aspa Emmanuel, the cane bearer of the king. He is also the spokesman of the Order of the Bloussoué. The version of the story used in this tutorial was read by Louise. The audio was sent on March 23, 2021, by Dr. Antoine Foba.

As far as linguistic description is concerned, the second publication of some significance is a 21-page paper by Hérault (1981:403-424). It includes a very succinct account of the sound system, the morphological system, the grammatical system, and a list of 100 words. The third is Gropou (2002:151-160). He pretty much replicated the previous study, focusing this time mostly on the morphophonological system of Betine. The fourth source is also by Gropou (2006:261-273). He collected, transcribed, and annotated three stories. The fifth source is Foba's (2009) dissertation on the syntactic structure of Betine. It examines some issues in theoretical syntax in light of Betine. It is an important source because it has some 200 words in the appendix. These five sources are a drop in the bucket when it comes to getting a language ready for speech synthesis. According to Rabiner and Schafer (2011:929), a minimum of 60 minutes of recording (approximately 1000 sentences) is required for building a basic TTS system for any language. We still have a long way to go because the story that we have contains only 18 sentences. To the best of our knowledge, we and our students are the first to undertake an acoustic phonetic investigation of Betine speech sounds. As a result of the *Speech Signal Processing* course that we taught, we have accumulated a very large set of acoustic phonetic measurements. More is needed for the complete digitalization of Betine (see 3.0 for a sample of the digitalized speech).

2.3 Text Normalization

Once a language model has been created, the next step in the development of a TTS system is called "text normalization" (Rabiner and Schafer 2011:909-912, 929). Texts need to be normalized for a variety of reasons. The most common ones are abbreviations, units of measurements, dates, punctuation, numerals, etc. When they appear in a text, they need to be spelled out exactly as they should be pronounced. For example, when a word such as "Dr." occurs in a text, one must decide whether the abbreviation should be read by the TTS system as "doctor" or "drive." A date written as 05/04/21 must be spelled out clearly as May fifth, two thousand twenty-one. Text normalization is also required for audio recordings, especially if they contain "hesitations, disfluencies, breaths, laughs, false starts or restarts, and other spontaneous speech effects," (Silverman et al. 1992:868). All unwanted "noises" and long pauses are edited out. In the end, the version of the text that is used for speech synthesis is clean, crisp, and clear.

The *Story of the Origin of the Beti People* was normalized before being assigned to students as their final course project. The original text contains 18 sentences (see Appendix) but it was divided into smaller fragments of five to eight words. Sentences 1 to 10 were assigned to students. Authors 1 and 2 devoted their attention to Sentence 11 from which they extracted a number of correlates (to be discussed in detail in 3.0):

Sentence 11: [wò lé àmú wɔ̀ kʉ̀lɛ̀ dɔ̀ lɛ̌ càlɛ̌ ɔ́ wɔ́ bà̱lɛ̀ àmú]

/Ils/avec/eux/ils/battre+ACC./guerre/finir/PART./ils/chasser+ACC./eux/

"Ils firent la guerre jusqu'à ce qu'ils(les Eotilé) soient vaincus et chassés à la fin."

They waged war against the Eotile until they defeated them. Finally, they chased them out.

This sentence is used to illustrate the process of building an acoustic phonetic model for Betine.

2.4 Speaker Model in Speech Synthesis

In building a TTS system for a language, one must decide on a "speaker model" before getting further into the project. The speaker model can be a single native speaker whose voice

is selected as the ideal representative of the speech community. All things being equal, a model that is based on multiple speakers is better than one that is based on a single speaker because, as Rabiner and Schafer (2011:929) note, "larger recorded databases provide more variations of each of the speech units, and generally this leads to more natural sounding synthetic speech." Even so, on page 952, on the discussion of speaker models, they note that speech synthesis can be based on the voice/accent of a single speaker. Two well-known examples are often cited in this respect. Klatt synthesized his own voice for the device that Physicist Stephen Hawking used to talk (https://bit.ly/32Lnsry). Until recently, Siri, the voice in Apple's smart devices was based on Susan Bennett's voice (https://bit.ly/3dNwLxt), (https://bit.ly/3nj0Ubc). In the case of critically endangered languages, one may not have the luxury of multiple voices because the language is known only by a very small number of people. Butcher's (2013:63) acoustic phonetic fieldwork observations underscore how hard it can be to find a speaker model:

> For most indigenous peoples, language is inextricably interwoven with culture and knowledge. It is viewed as a constant and unchanging entity which is owned by the community and whose custodianship is entrusted to certain elders of that community. Consequently, those who are regarded by the community (and perhaps the linguist) as the 'best speakers' may not necessarily be the best speakers for the purposes of phonetic research. They may be breathless, toothless, lacking volume, with slurred articulation and poor understanding of the nature of the task. It may well be appropriate to spend some time making recordings of such speakers, however, in terms of your relationship with the community before discreetly seeking out younger, clearer speakers you will undoubtedly need, especially for tasks such as palatography and aerodynamic recording.

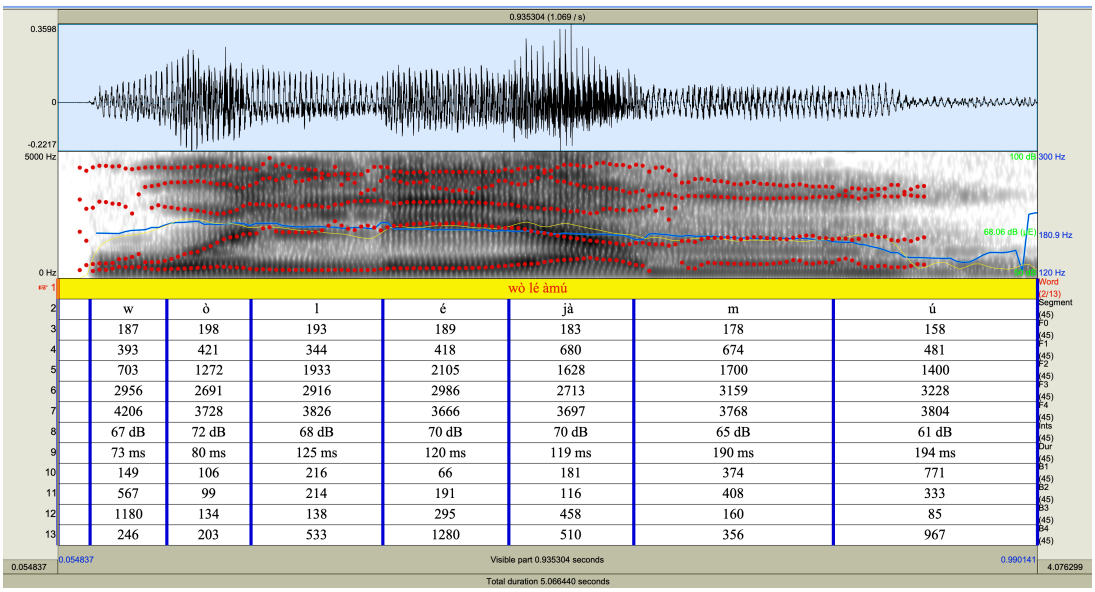
For Betine, we have selected **Louise Ahatetamala** as the speaker model. She holds a Ph.D. degree in linguistics from the Université Félix Houphouet Boigny (UFHB) in Abidjan, Côte d'Ivoire. In addition to Louise, we are also collaborating with two other people: Dr. Foba and Adelaide Amenan. They worked together on the IPA transcription and the translation of the Betine text into French (see Appendix).

3.0 Feature Extraction for Speech Digitalization

The ultimate goal of acoustic feature extraction is speech digitalization. It is the process whereby analog speech sounds are turned into numerical vectors so that they can be used in speech synthesis. In Klatt's approach to speech synthesis that we are following, each phoneme is represented by 11 numerical vectors (see Table 1). Since our goal is syllable-based formant synthesis, the vectors that make up each segment are concatenated into syllable vectors. In some cases, a syllable will consist of only 11 vectors, in others, there will be 22 vectors because the most common types of syllables in Betine are CV (Consonant Vowel).[4] Feature extraction is the next logical step after text normalization. Speech digitalization is tedious and time-consuming, but it is an indispensable aspect of speech synthesis. For English, Klatt (1980) and Klatt and Klatt (1990) extracted 11 main parameters which are F0, F1, F2, F3, F4, intensity, duration, and four bandwidths (B1, B2, B3, and B4). These correlates must be extracted for **every** speech sound in the language, no matter how often it occurs, because as Klatt (1980:987) puts it, "a large set of CV syllables" is needed for speech synthesis to work well. Spectrographs

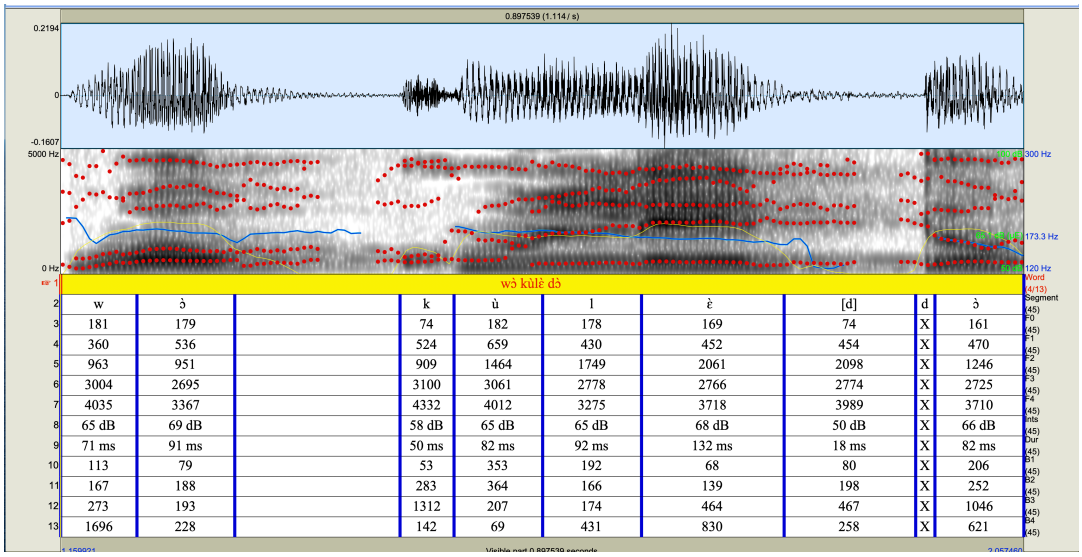
[4] This paper focuses on formant-based synthesis. We are concurrently researching the feasibility of concatenated syllable synthesis. However, this approach is not discussed in this paper.

1 to 4 display all the 11 correlates extracted for every single sound in Sentence 11. Praat, Version 6.1.42, is the software used to extract all relevant parameters.



	w	ò	l	é	jà	m	ú
2	187	198	193	189	183	178	158
3	393	421	344	418	680	674	481
4	703	1272	1933	2105	1628	1700	1400
5	2956	2691	2916	2986	2713	3159	3228
6	4206	3728	3826	3666	3697	3768	3804
7	67 dB	72 dB	68 dB	70 dB	70 dB	65 dB	61 dB
8	73 ms	80 ms	125 ms	120 ms	119 ms	190 ms	194 ms
9	149	106	216	66	181	374	771
10	567	99	214	191	116	408	333
11	1180	134	138	295	458	160	85
12	246	203	533	1280	510	356	967

Figure 1: Spectrogram 1 of Sentence 11



	w	ò		k	ù	l	è	[d]	d	ò
2	181	179		74	182	178	169	74	X	161
3	360	536		524	659	430	452	454	X	470
4	963	951		909	1464	1749	2061	2098	X	1246
5	3004	2695		3100	3061	2778	2766	2774	X	2725
6	4035	3367		4332	4012	3275	3718	3989	X	3710
7	65 dB	69 dB		58 dB	65 dB	65 dB	68 dB	50 dB	X	66 dB
8	71 ms	91 ms		50 ms	82 ms	92 ms	132 ms	18 ms	X	82 ms
9	113	79		53	353	192	68	80	X	206
10	167	188		283	364	166	139	198	X	252
11	273	193		1312	207	174	464	467	X	1046
12	1696	228		142	69	431	830	258	X	621

Figure 2: Spectrogram 2 of Sentence 11

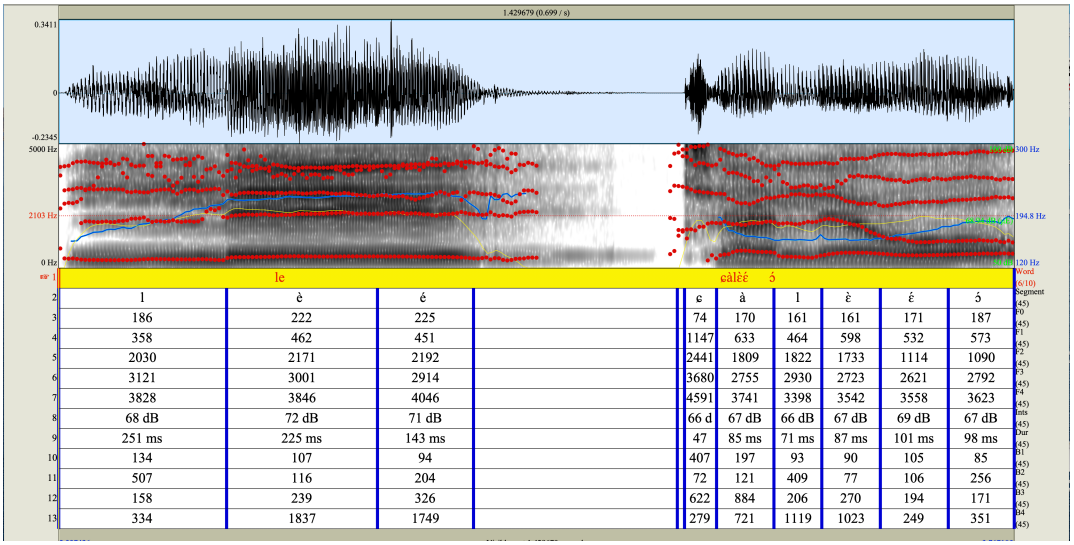


Figure 3: Spectrogram 3 of Sentence 11

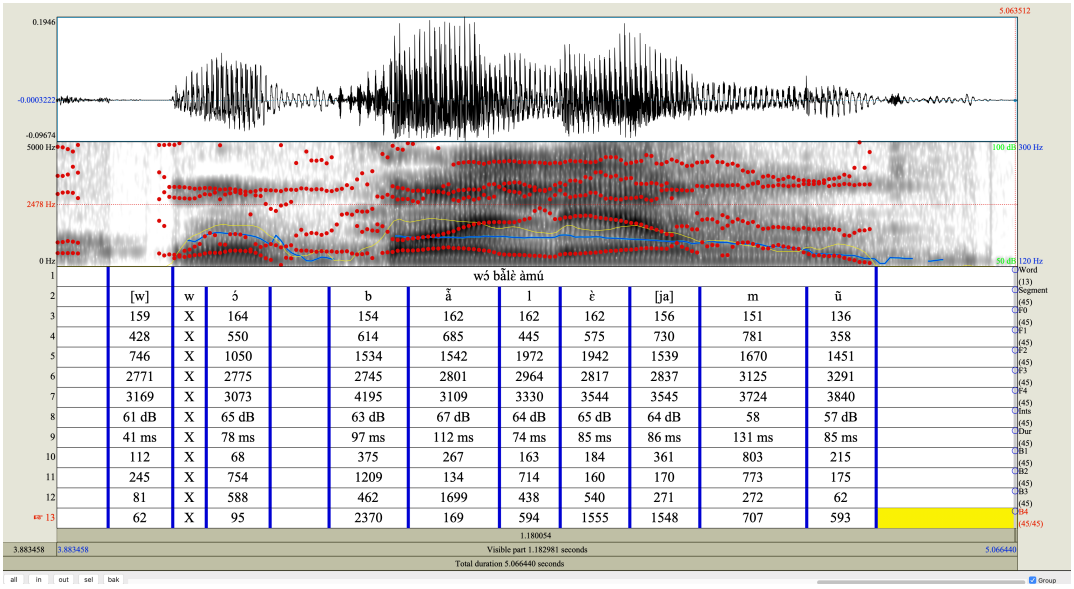


Figure 4: Spectrogram 4 of Sentence 11

Once the relevant features are extracted, they are tabulated, as displayed in Table 1. This presentation mirrors the one found in Klatt (1980:987).

N0	Segment	F0	F1	F2	F3	F4	Ints	Dur	B1	B2	B3	B4
1.	w	187	393	703	2956	4206	67	73	149	106	1180	246
2.	ò	198	421	1272	2691	3728	67	80	106	99	134	203
3.	l	193	344	1933	2916	3826	68	125	216	214	138	533
4.	é	189	418	2105	2986	3666	70	120	66	191	295	1280
5.	[j]à	183	680	1628	2713	3697	70	119	181	116	458	510
6.	m	178	674	1700	3159	3768	65	190	374	408	160	356
7.	ú	158	481	1400	3228	3804	61	194	771	333	85	967
8.	w[5]	181	360	963	3004	4035	65	71	113	167	273	1696
9.	ɔ̀	179	536	951	2695	3367	69	91	79	188	193	228

[5] We will not discuss nasals and nasalized vowels in this paper. A companion paper written by our student, Scarlet Dusosky (2022), deals with synthesizing nasal vowels. We want to simply point out that nasal segments present a number of challenges in synthesized speech (Rabiner and Schafer 2011:644). Klatt (1980) and Klatt and Klatt (1990) give nasals considerable attention in their papers.

10.	k	60[6]	534	909	3100	4332	58	50	53	282	1312	142
11.	ù	182	659	1464	3061	4012	65	82	353	364	207	69
12.	l	178	430	1749	2778	3275	65	92	192	166	174	431
13.	è	169	452	2061	2766	3718	68	132	68	130	464	830
14.	d	60	454	2098	2774	3989	50	18	80	198	467	258
15.	ɔ	161	470	1246	2725	3710	66	82	206	252	1046	621
16.	l	186	358	2030	3121	3828	68	251	134	507	158	334
17.	è	222	462	2171	3001	3846	72	225	107	116	239	1837
18.	ɛ́	225	451	2192	2914	4046	71	143	94	204	326	1749
19.	ɐ	60	1147	2441	3680	4591	66	44	407	72	622	279
20.	à	170	633	1809	2755	3741	67	85	197	121	884	721
21.	l	161	464	1822	2930	3398	66	71	93	409	206	1119
22.	ɛ̌	161	598	1733	2723	3542	67	87	90	77	270	1023
23.	ɔ́	187	573	1090	2792	3623	67	98	85	256	171	351
24.	w	159	428	746	2771	3169	61	41	112	245	81	62
25.	ɔ́	164	550	1050	2775	3073	65	78	68	754	588	95
26.	b	154	614	1534	2745	4195	63	97	375	1209	462	2370
27.	à	162	658	1542	2801	3109	67	112	267	134	1699	169
28.	l	162	445	1972	2964	3330	64	74	163	714	438	594
29.	è	162	575	1942	2817	3544	65	85	184	160	540	1555
30.	[j]à	156	730	1539	2837	3545	64	86	361	170	271	1548
31.	m	151	781	1670	3125	3724	58	131	803	773	272	707
32.	ú	136	358	1451	3291	3840	57	85	215	175	62	593

Table 1: Parametric Segmental Measurements

Measuring the same segments in multiple phonological environments helps to understand its range of variations. From Klatt (1980), Klatt and Klatt (1990), and from general knowledge of acoustic phonetics that has accumulated for nearly 100 years of research, the following generalizations apply to speech synthesis in all languages:

1. Voiceless segments can be given an F0 of 60 Hz because Fry (1979:68) notes that the lowest F0 that human beings can produce is 60 Hz.
2. The shortest segment should be at least 30 msec and the longest should not be more than 500 msec (Klatt and Klatt 1990:844). However, in general, one should be suspicious of any segment lasting more than 300 msec. More often than not, durations of more than 300 msec are signs of 'over-articulated," (Rabiner and Schafer 2011:929).
3. Voiced segments may be devoiced. Such is the case of [d] in # 14 whose F0 is listed by Praat as "undefined." As a rule of thumb, all "undefined" segments should be given an F0 of 60 Hz.

Now that we have extracted some features of Betine, let's pretend that we are writing a **pseudocode** for the syllable [wò]. First, we convert the IPA transcription [wò] into an Arpabet transcription, which is, /#W OW#/.[7] The syllable is represented by the numerical vectors [F0 187, F1 393, B1 149, F2 703, B2 106, F3 2956, B3 1180, F4 4206, B4 246, Ints 67, Dur 73**;** F0 198**%2**,[8] F1 421, B1 106, F2 1272, B2 99, F3 2691, B3 134, F4 3728, B4 203, Ints 67, Dur 80]. Ideally, a well-functioning speech synthesis system will pronounce the concatenated sequences of codes as [wò]. By the time the system is ready to launch, all the licit syllables of

[6] When Praat renders an "undefined" for pitch measurements, the value of 60 Hz should be used.
[7] In keeping with Rabiner and Schafer (2011:930), Arpabet transcriptions are inside of slashes. The symbol "#" signifies the beginning and end of each utterance. We are treating [wò] as both a syllable and a word.
[8] The code "%" after a vowel is a tone marker. "1" indicates a high tone, "2" signifies a low tone (See Koffi 2020:136-137).

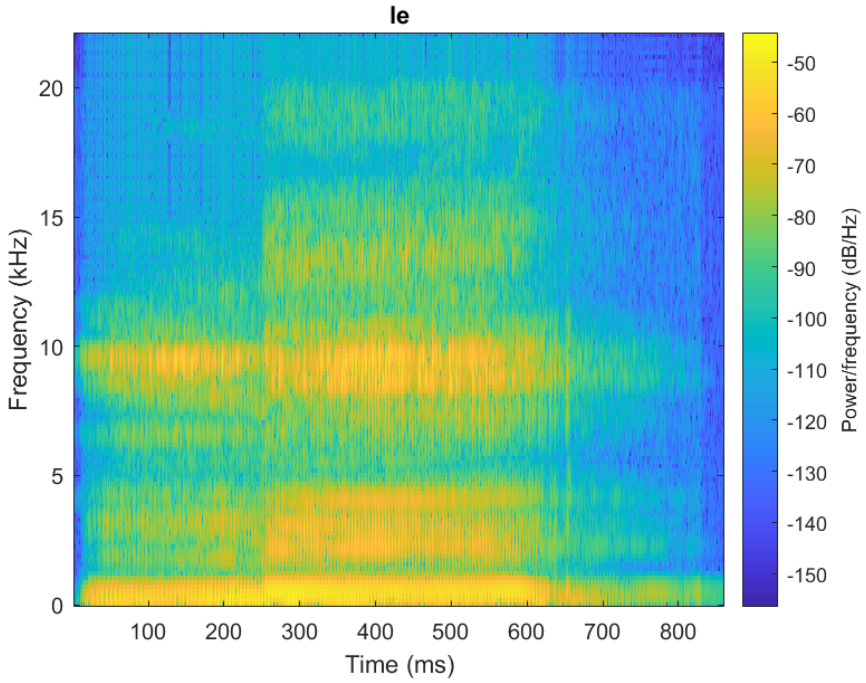
the language will have been concatenated this way. With this method, all the existing words and yet-to-be-coined words can be pronounced.

4.0 F0/Pitch Extraction from Praat and MATLAB

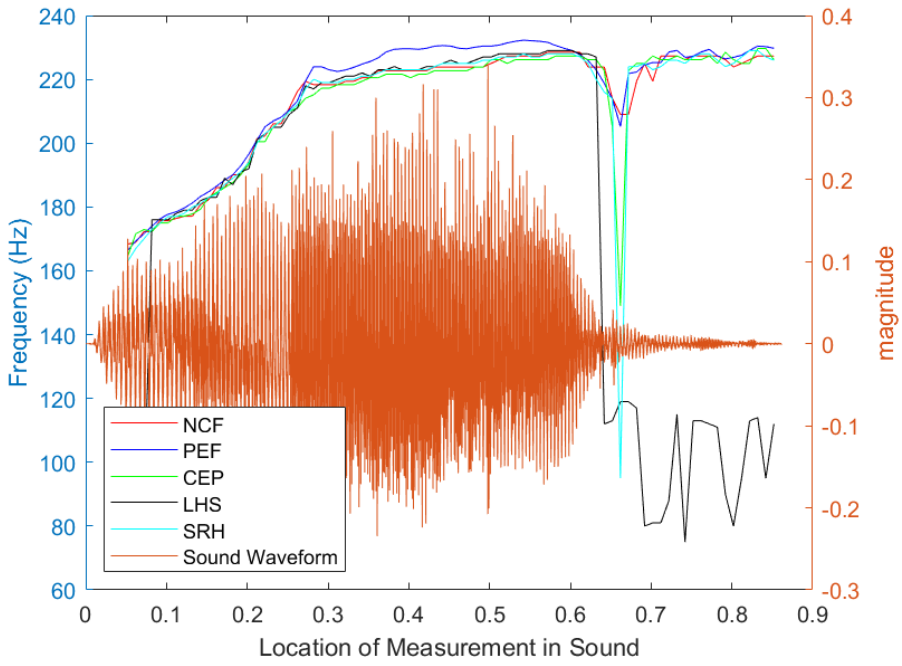
In extracting data for speech synthesis, it is very important to extract F0 measurements accurately, otherwise speech synthesis will not work well. Multiple measurements should be extracted from the same segments. This way, in the end, we will end up with three sets of measurements for each segment, in accordance with Klatt (1980:975). The three sets are the minimum pitch value, the maximum pitch value, and the arithmetic mean (sometimes, the mode).

Normally, one and the same software is used to extract all parameters. However, since this tutorial is based on a course that Authors 1 and 2 taught, we used two different software. Praat was used to extract all the measurements. Subsequently, MATLAB (2020) was used for coding. The goal was to determine whether or not the two software systems would yield the same results. We will return to this when we discuss the "software model." For now, we will only say that extracting F0 measurements is easy in Praat. Author 2 wrote the following code for extracting pitch values in MATLAB. For this demonstration, the mean F0 value of the vowel **[ĕ]** in the word [lĕ] from Sentence 11 is extracted using five different pitch extraction algorithms:

```
clc
close all
[x,fs]=audioread('Sentence11Phrase3.wav');  %read in sound file
sound(x,fs)  %listen to sound file
segmentlen=100;  %segment length of 100 samples
noverlap=90;  %number of samples of overlap for the spectrogram
NFFT=256;   %number of frequency points of the FFT
spectrogram(x,segmentlen, noverlap, NFFT, fs, 'yaxis');
title('le')  %run the spectrogram
```



```
dt=1/fs;  %calculate time between samples
I0=4781;
figure(2)
%calculate pitch using the 5 methods available in MATLAB to compare.
[f0NCF,idxNCF]=pitch(x,fs, "Range",[75,300],'Method','NCF');
[f0PEF,idxPEF]=pitch(x,fs, "Range",[75,300],'Method','PEF');
[f0CEP,idxCEP]=pitch(x,fs, "Range",[75,300],'Method','CEP');
[f0LHS,idxLHS]=pitch(x,fs, "Range",[75,300],'Method','LHS');
[f0SRH,idxSRH]=pitch(x,fs, "Range",[75,300],'Method','SRH');
yyaxis left
plot(dt*idxNCF,f0NCF, 'r-', dt*idxPEF, f0PEF, 'b-', dt*idxCEP, f0CEP, 'g-',
dt*idxLHS, f0LHS, 'k-', dt*idxSRH, f0SRH, 'c-')
hold on
ylabel('Frequency (Hz)')
yyaxis right
t = dt*[0:length(x)-1];
plot(t,x)  %plot the sound wave
legend('NCF','PEF','CEP','LHS','SRH','Sound
Waveform','Location','southwest')
xlabel('Location of Measurement in Sound')
ylabel('magnitude')
```




```
Istart = max(find(dt*idxNCF <= 0.262));
Iend = min(find(dt*idxNCF >= 0.632));
disp(mean(f0NCF(Istart:Iend)));
```

```
223.5016
```

```
Istart = max(find(dt*idxPEF <= 0.262));
Iend = min(find(dt*idxPEF >= 0.632));
```

```
disp(mean(f0PEF(Istart:Iend)));
```

```
227.5428
```

```
Istart = max(find(dt*idxCEP <= 0.262));
Iend = min(find(dt*idxCEP >= 0.632));
disp(mean(f0CEP(Istart:Iend)));
```

```
222.3673
```

```
Istart = max(find(dt*idxLHS <= 0.262));
Iend = min(find(dt*idxLHS >= 0.632));
disp(mean(f0LHS(Istart:Iend)));
```

```
221.4359
```

```
Istart = max(find(dt*idxSRH <= 0.262));
Iend = min(find(dt*idxSRH >= 0.632));
disp(mean(f0SRH(Istart:Iend)));
```

```
223.3846
```

The extracted measurements are summarized as follows:

Pitch Computation Method in MATLAB	Mean Pitch Frequency Computed
Normalized Correlation Function (default)	223.5 Hz
Pitch Estimation Function	227.5 Hz
Cepstrum Pitch Determination	222.4 Hz
Log Harmonic Summation	221.4 Hz
Summation of Residual Harmonics	223.4 Hz

Table 2: Results of the various pitch detection methods available in MATLAB.

The mean F0 of **[ě]** rendered by Praat is 222 Hz. As shown in Table 2, the pitch detection methods in MATLAB give results very close to this, except for the Pitch Estimation Function. In MATLAB, just as in Praat, the user must define the starting and stopping points for the pitch measurement, and this can lead to some error in measurement. The pitch extraction methods used are the following:

1. Normalized Correlation Function (NCF)
2. Pitch Estimation Filter (PEF)
3. Cepstrum Pitch Determination (CEP)
4. Log Harmonic Summation (LHS)
5. Summation of Residual Harmonics (SRH)

MATLAB offers the following disclaimer: "The different methods for estimating pitch provide trade-offs in terms of noise robustness, accuracy, optimal lag, and computation expense." Upon a closer examination, we see that NCF and PEF have the same fall and rise profiles, but LHS, CEP, and SRH differ rather significantly. Since they all render the same measurements, we could ignore these minute differences. However, because F0 is extremely important in speech synthesis, preference should be given to the default method NCF when using MATLAB because it matches Praat the most.

4.1 Formant Extraction in Praat and MATLAB

At the core of Klatt's model of speech synthesis are formants. In other words, the importance of formant extraction cannot be overstated. Klatt (1980) and Klatt and Klatt (1990) extracted F1, F2, F3, and F4 measurements. Occasionally, F5 was included. Formants are worth extracting because they correlate very well with the articulatory characteristics of individual speech sounds. F1 correlates with mouth aperture, F2 with horizontal tongue movements, F3 with lip positions, F4 with the size of one's head and/or laryngeal cavities. There is no clear articulatory correlate for F5. This explains why it is often not measured.

After one has settled on the number of formants to extract, the next important decision one has to make is the sampling rate. Modern audio devices record speech at 44100 Hz. This is an extremely high sampling rate that is likened to compact disk (CD) quality. Oddly enough, this sampling frequency is not good for speech synthesis. It is recommended that audio files be downsampled to 10000 Hz, 8000 Hz, or 5000 Hz. Downsampling recordings this way does not cause speech signals to lose any of their essential attributes (Klatt 1980:975). To prove this point, a file containing the vowel **[ě]** in the word [lě] from Sentence 11 was resampled in Praat at three different rates before measurements were extracted. The results are displayed in Table 3:

Segment	**F1**	**F2**	**F3**	**F4**	**F5**
44100 Hz	456	2179	2968	3937	4147
10000 Hz	455	2175	2961	3799	4147
8000 Hz	448	1923	2399	3080	3773

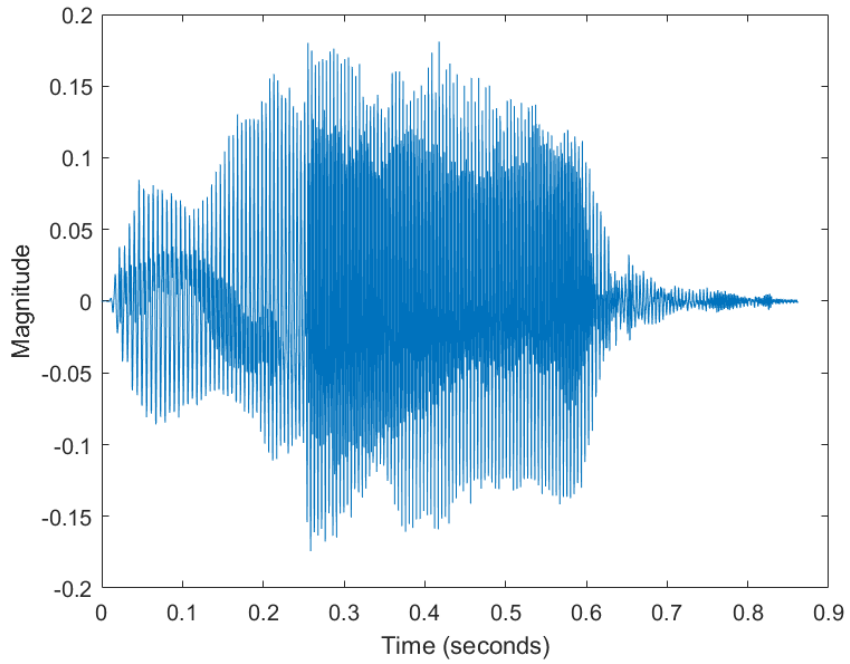
Table 3: Sampling and Downsampling Rates

We see clearly that downsampling from 44100 Hz to 10000 Hz yields the same results. This statement may seem to be demonstrably false because there are arithmetic differences between the two frequency rates. However, the statement must be interpreted from the standpoint of auditory perception. Formant frequency is perceived by the naked ear on a logarithmic scale, not on an arithmetic/linear scale. Nearly 100 years of psychoacoustic research and experimentation have established reliable Just Noticeable Difference (JND) thresholds at which formant measurements become auditorily meaningful. As a general rule of thumb, Rabiner and Juang (1993:152) note that the JND is 3-5% of the formant under consideration. On the F1 formant bandwidth, a difference of ≤ 60 Hz is imperceptible to the naked ear. For F2, the JND is ≤ 200 Hz. In the F3 domain, the JND is ≤ 400 Hz. On the F4 bandwidth, the JND is ≤ 600 Hz. For F5, the JND is ≤ 800 Hz. Sampling at 8000 Hz does not yield the best results because the differences begin to be perceptually salient in F2, F3, and F4. Klatt (1980:981,988) used a 10000 Hz sampling rate. It is noteworthy that the default resampling frequency in Praat is set at 10000 Hz.

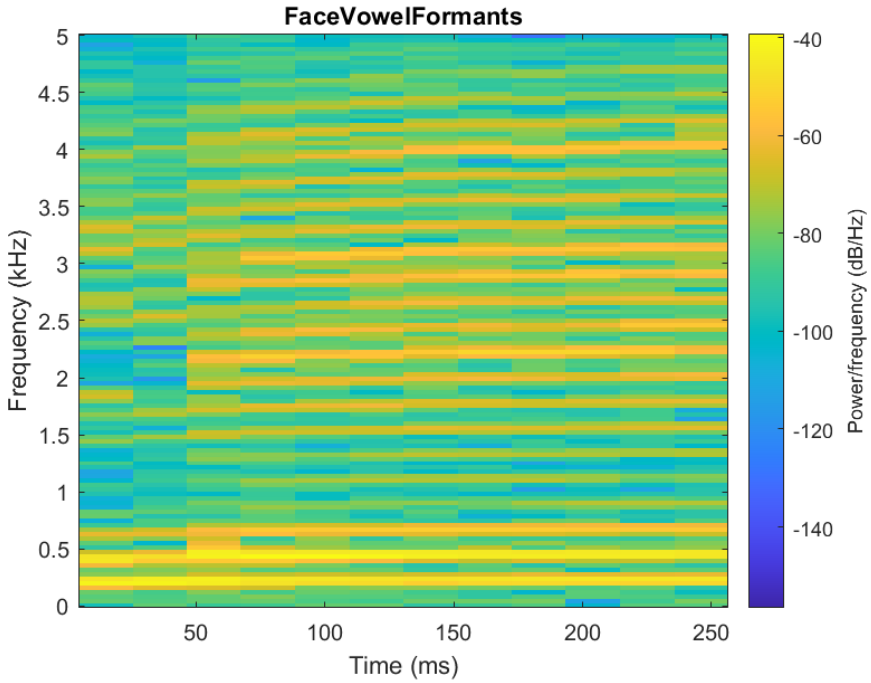
4.2 Downsampling in Praat and MATLAB

It follows from the previous section that files must be downsampled before formant data are extracted. To do so, the following steps need to be implemented in Praat. First, the **<Convert>** tab must be selected. Then one clicks on the **<Resample>** tab. When the dialog window opens, one accepts the preset value of 10000 Hz. Care should be taken **not** to change the preset value of **<Precision (Samples)> of 50 Hz**. Alternatively, one can write a script in MATLAB to downsample files automatically. The script below describes the coding steps used by Author 2 to extract formant measurements.

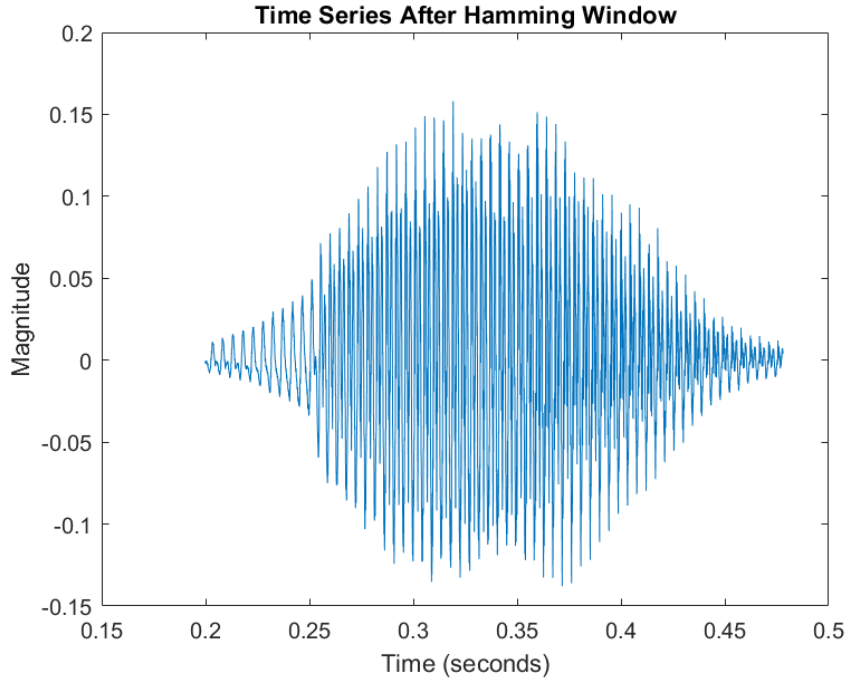
```
clc;
close all
[y,Fs]=audioread("Sentence11Phrase3_10000.wav");
figure(1)
dt = 1/Fs;
t = dt*[0:length(y)-1];  %calculate the time vector
plot(t,y)
xlabel('Time (seconds)')
ylabel('Magnitude')
```




```
I0=2000;
Iend=4781;  %hand pick the starting and ending points for the analysis
x=y(I0:Iend);  %create sequence for analysis
sound(x,Fs)  %play the sound
segmentlen=300;  %length of the segment
noverlap=90; %number of samples of overlap
NFFT=256;  %length of the FFT for the spectrogram
spectrogram(x, segmentlen, noverlap, NFFT, Fs, 'yaxis')
title('FaceVowelFormants')
```




```
[f0,loc]=pitch(x,Fs);  %calculate pitch using the default method (NFC)
x1=x.*hamming(length(x));  %apply a hamming window to the time series
plot(t(I0: Iend),x1)
title('Time Series After Hamming Window')
xlabel('Time (seconds)')
ylabel('Magnitude')
```




```
preemph=[1 0.63];  %preemphasis filter coefficients
x2=filter(1,preemph,x1);  %filter x1 through the preemphasis filter
lpcseglentgth = 50;  %segment length for the linear prediction filter
```

```
 lpcoverlap = 40;  %overlap for the linear prediction filter
 kbegin=1;  %beginning point of the analysis
 kend=kbegin+lpcseglentgth;  %ending point of the analysis
 mm=1;  %array row increment
 while kend <= length(x2)  %while loop, stops when the calculated end point
kend is greater than the length of the time series
     x3 = x2(kbegin:kend);  %easy time series variable
     [A(mm,:), g] = lpc(x3,10);  %the linear prediction coefficeint
calculation of order 10, outputs a row of A, and g the variance of the
prediction
     kbegin=kbegin + (lpcseglentgth-lpcoverlap);  %increment the beginning
point of the calculation
     kend=kbegin + lpcseglentgth;  %increment the ending point of the
calcuation
     rts=roots(A(mm,:));  %roots of the linear prediction coefficients are
related to the formants of the sound segment
     rts1=rts(imag(rts)>=0);  %since they are complex conjugate pairs, we
will take only the roots with positive imaginary parts
     angz=atan2(imag(rts1), real(rts1));  %calculate the angle of the complex
roots
     [frqs, indicies]=sort(angz.*(Fs/(2*pi)));  %sort the angles, convert
them from radians to Hertz
     bw= -(Fs/pi)*log(abs(rts1(indicies)));  %calculate the bandwith of the
formants
     r0=abs(rts1);
     fx1 = (1+r0.^2)./(2*r0).*cos(angz)+(1-r0.^2)./(2*r0).*sin(angz);%this is
an alternative way of caclulating formants and bandwidths
     fx2 = (1+r0.^2)./(2*r0).*cos(angz)-(1-r0.^2)./(2*r0).*sin(angz);
     frqhat = (Fs/(2*pi))*acos(cos(angz).*(r0.^2+1)./(2*r0));
     bwhat = (Fs/(2*pi))*abs(acos(fx1)-acos(fx2));
     frqhatarray{mm}=frqhat(bwhat <500); %These are the alternative formants
and bandwidths
     nn=1;
     for kk=1:length(frqs)  %pull out the valid formants
         if(frqs(kk) >90 && bw(kk) <400);
             formants(nn)=frqs(kk);
             nn=nn+1;
         end
     end
     if length(formants) < 4  %zero pad if the number of formants found is
less than 4
         formants=[formants zeros(1,4-length(formants))];
     end

     formantsarray(mm,:) = formants(1:4);

     mm=mm+1;  %increment the row on A and perform the lpc analysis on the
next segment
 end
```

```
formantsarray(formantsarray==0) = NaN;
dispformant(1) = mean(formantsarray(:,1),'omitnan');
dispformant(2) = mean(formantsarray(:,2),'omitnan');
dispformant(3) = mean(formantsarray(:,3),'omitnan');
dispformant(4) = mean(formantsarray(:,4),'omitnan');
fprintf('%6.2f Hz\n', dispformant)
```

```
436.61 Hz
2759.45 Hz
3558.70 Hz
4074.17 Hz
```

MATLAB includes much of the above code in an example for the Linear Prediction Coefficient function as an example of how to use the function. The pre-emphasis filter is included to attenuate lower frequencies and amplify higher frequencies, making the higher frequency formants easier to detect. In digital filtering, filters act on frequencies between 0 Hz and ½ of the sampling frequency (Oppenheim and Schafer 2010:160). Because of this filter, the sampled data must have a sampling frequency between 8000 Hz and 10000 Hz.

The main function used in the above code is the Linear Prediction Coefficient (*lpc*) function, which calculates a *p*th order linear predictor. A linear predictor attempts to predict the current value based on past values. The order selected in the above code is for a 10^{th} order linear predictor. The order needs to be at least double the number of formants to calculate plus two. Formants are the angles of the complex roots of the linear prediction equation calculated using the *roots* function in MATLAB. MATLAB function *atan2* calculates the angle from the complex number. The bandwidths of the formants are calculated in one of two ways, as outlined in Snell and Milinazzo (1993:129). The two ways give similar results.

The "while loop" in the code is used to calculate the linear predication coefficients over segments of the waveform. The segments are allowed to overlap by the value of *lpcoverlap*. The formants for each segment are then averaged over all segments. Since some formants are not calculated for all segments, zeros in the array are replaced with the MATLAB symbol NaN (Not a Number), which can then be ignored in the *mean* function.

4.3 In Search of a "Software Model"

Nowadays, there are many software packages that can be used for various aspects of the speech synthesis process. It behooves the researcher(s) to choose the one that offers many benefits. Items to consider are price and reliability of results. All things being equal, tools that are free and reliable are the best. For the purposes of this demonstration, we are comparing Praat and MATLAB. There are other coding languages such as Python, C+++, etc. that can be used to build a TTS system. Praat and MATLAB are compared simply because these are the two tools that we used to teach our course. The first and most important question in formant-based synthesis is "which software to use in extracting measurements?" This is not a trivial matter because software packages do not yield the same results. An error analysis is performed to help choose the most suitable tool. The error analysis formula is stated as follows:

$$\text{Errors} = \frac{\text{Accepted Values} - \text{Experimental Values}}{\text{Accepted Values}} \times 100$$

We take measurements rendered to by Praat to be the "accepted values" and those by MATLAB to be the "experimental values." The measurements concern the vowel **[ĕ]** in the word [lĕ] extracted and measured in both Praat and MATLAB at a sampling frequency of 10000 Hz, as displayed in Table 4:

Segment [ĕ]	**F0**	**F1**	**F2**	**F3**	**F4**
Praat (10000 Hz)	222 Hz	448 Hz	1923 Hz	2399 Hz	3080 Hz
MATLAB (10000 Hz)	222 Hz	480 Hz	2159 Hz	2910 Hz	4080 Hz
Difference	**0 Hz**	**32 Hz**	**236 Hz**	**511 Hz**	**1000 Hz**
Error Rate	0%[9]	7.14%	12.27%	21.30%	32.46%
Formant JNDs	**1 Hz**	**60 Hz**	**200 Hz**	**400 Hz**	**600 Hz**

Table 4: Error Rate in Sampling Analysis

A cursory examination reveals that the error rates concerning F0 and F1 are insignificant because they fall below their respective JNDs. However, the differences in F2, F3, and F4 are auditorily salient because they are above their respective JNDs.

What do these differences mean for selecting a software model? It simply means that software packages do not render the same results for all correlates. Some correlates will be the same, but others will vary. For formant extraction, our preference is for Praat for three reasons. First, it is a dedicated software for speech analysis. This means that its algorithms are fine-tuned and geared towards acoustic phonetic analyses. Secondly, it is free and widely available. Thirdly, MATLAB is an expensive proprietary software package. We used the school version to teach our course. Since MATLAB is not free, it goes without saying that many people endeavoring to document and synthesize dying or lesser-known languages will not have access to it.

4.4 Bandwidth Issues in Speech Synthesis

In human-to-human verbal interactions, formant bandwidths are not critically important for intelligibility. There are several mitigating factors such as intonation, loudness, and the tempo of delivery that eclipse formant bandwidth issues. Our students had a lot of questions about bandwidths. We were not always successful in answering their questions about formant bandwidths. In fact, the blank stares on their faces spoke volumes. We did the best we could because bandwidths defy simple definitions. Rabiner and Juang (1993:153) contend that formant bandwidths are not easy to explain because they involve several loosely interrelated phenomena. They put it simply as follows: "Unlike the frequency JNDs, the bandwidth JNDs do not show clear dependence on either the bandwidth itself or the formant frequency." So, instead of a simplistic definition, let's consider five general characteristics:

1. All things being equal, smaller formant bandwidths are better than larger ones because "narrow-formant vowel might be more resistant to noise. … A narrow-formant vowel might thus be a less severe masker. There is therefore ample reason to suspect that formant bandwidth might affect identification of vowels *in competition*. … A vowel's formants are *visually* more prominent when their bandwidths are narrow rather than

[9] Collectively, the error rates that students in the course reported for F0 is less than 2%.

wide," (Cheveigné 1999: 2093, 2096). Also, listening tests indicate that listeners prefer smaller formant bandwidths than larger ones, (Dunn 1961:1745).

2. Klatt (1980:980) correlates formant bandwidths with supralaryngeal articulation, "Formant bandwidths are a function of energy losses due to heat conduction, viscosity, cavity-wall motions, radiation of sound from the lips, and the real part of glottal source of irregularities in the glottal source spectrum. Results indicate that bandwidths vary by a factor of 2 or more as a function of the particular phonetic segment being spoken."
3. There is a correlation between bandwidth with intensity: "The primary perceptual effect of a bandwidth change is an increase or decrease in the effective intensity of a formant energy concentration, … the intensity value in a formant is inversely proportional to its bandwidth," (Klatt 1980: 982). This means that "smaller bandwidths have greater intensity and larger bandwidths have smaller intensity," (Klatt 1980: 982).
4. Women's formant bandwidths vary unpredictably: "While the formant frequencies associated with the different vowels are quite well known and can be measured with considerable accuracy in any given case, it has been found much more difficult to make accurate measurements of bandwidth. … The situation is worse for a woman's voice, and most investigators have confined their attention to male voices," (Dunn 1961:1737).
5. Klatt (1980:975, 986-987) provides three sets of measurements for each bandwidth: The smallest measurements, the largest, and the typical, that is, the mode. They are respectively as follows: B1 40 Hz, 500 Hz, and 50 Hz; B2: 40 Hz, 500 Hz, and 70 Hz; B3: 40 Hz, 500 Hz, and 110 Hz; B4: 100 Hz, 500 Hz, and 250 Hz; B5: 150 Hz, 700 Hz, and 200 Hz.[10] From these measurements, we note that no English sound has a minimum bandwidth that is smaller than 30 Hz and none that is larger than 800 Hz.

The speaker model that one chooses will affect reported bandwidth measurements. If a female speaker is chosen, as is the case for Betine, one can expect larger than normal bandwidth measurements. Rabiner and Juang (1993:152) have given us some thresholds to let us know if a speaker's bandwidths fall outside of normative values. They indicate that bandwidths should fall between 20 and 40% of their respective formant values. Let's evaluate our speaker model's formant in light of this JND. Since smaller is better, the calculations are made with 20% for the vowel **[ĕ]** in the word [lĕ] produced by Louise.

Segment	**F1**	**B1**	**F2**	**B2**	**F3**	**B3**	**F4**	**B4**	**F5**	**B5**
44100 Hz	456	159	2179	258	2968	300	3937	2001	4147	135
10000 Hz	**455**	**153**	**2175**	**252**	**2961**	**282**	**3799**	**3799**	**4147**	**133**
8000 Hz	448	62	1923	352	2399	192	3080	87	3773	131
Louise 20%	455	**91**	2175	**435**	2961	**592**	3799	**759**	4147	**829**

Table 5: Formants and Corresponding Bandwidths

Three sampling frequencies are used, but the bandwidth calculations are based on a sampling rate of 10000 Hz. Right away, we see that some of the speaker model's bandwidth fall outside of normal ranges for some formants but not for others. For speech synthesis, the smallest bandwidth is preferable to the biggest. So, for example, for F1, 91 Hz is better than 153 Hz, for F2, 252 Hz is preferrable to 435 Hz, and so on and so forth. When all the bandwidths are calculated, compared, and contrasted, we follow Klatt (1980) and choose the typical value for that speech sound. The "typical value" can be the arithmetic mean of all the values for any

[10] These are presumably based on men's bandwidths. However, Klatt (1980) claims that they work for both males and females.

given sound or the mode, whichever is the smaller. However, there is a caveat. It should not be less than 40 Hz and more than 500 Hz. For example, in English, Klatt (1980: 986-987) does not list any segment with a bandwidth less than 40 Hz and no segment with bandwidth of over 500 Hz. The general assumption is that bandwidths that are less than 40 Hz are too narrow and can sound like a pure tone. By the same token, most segments do not have bandwidths that are higher than 500 Hz because if they do, they may end up sounding more like white noise. These guidelines give us some rough ideas of ranges of what might be acceptable for speech synthesis in most languages.

5.0 Building an Intonation Model

Some formant-based synthesized speech sound robotic and monotone if intonation is not given the attention it deserves. The lack of an acoustic phonetic framework for intonation analysis has compelled some speech scientists to resort mainly to the ToBI (tone and break indices) approach. In this framework, symbols such as H*, L*, H- and L-, H%, L% are combined in various ways to capture the subtleties of intonation patterns (see Ladd 2008:91). Unfortunately, this approach generates annotations that are not directly quantifiable. For this reason, Author 1 has been championing a model of intonational analysis that is based on Fry's (1958) seminal study of auditory perception of pitch. His research illuminates how humans perceive pitch. He recruited 41 participants and conducted several lexical pitch perception experiments. We highlight only the aspects of his findings that are important for the study of intonation. Words were produced with a reference pitch of 97 Hz (male voice). Thereafter, Fry incrementally increased pitch by 3 Hz on certain syllables and not others. Some participants perceived the syllables in question as having a higher pitch. However, other participants could not tell which syllable had a higher pitch and which did not. In another experiment, he increased pitch by 5 Hz. All 41 participants without exception correctly identified the syllable with the higher pitch. Subsequently, Fry varied pitch by increments of 10 Hz, 15 Hz, 20 Hz, etc. up to 60 Hz. Surprisingly, he obtained the same results as when he varied pitch only by 5 Hz. He concluded that "Increase in the size of the frequency step appears to produce no mark trend in the results." Since Fry's original experiments, other experiments, including t' Hart (1981:812) and Liu (2013:3018), have confirmed that 5 Hz is the optimal threshold at which lexical pitch is perceived unambiguously. Furthermore, Fry (1958:141) made this very important observation:

> In the intonation patterns heard from most English speakers, changes in pitch of more than one octave are infrequent and are not often met within successive syllables, even from the most excitable speakers.

What does this mean? It means that drastic changes in intonation within the same intonation phrase (IP) are unusual. In other words, pitch movements are likely to remain within the same octave level. This begs the question as to what an octave is. For a succinct answer, octave levels based on the Critical Band Theory (CBT) are displayed in Tables 6 and 7. For a fuller explanation, see Koffi (2017:147-165) or Koffi (2018:110-131).

N0	Pitch Registers	Lower Limits	Center Frequency	Upper Limits
1.	Extra low	71	80	88
2.	Low	89	100	113
3.	Mid	114	125	141
4.	High	142	160	176
5.	Extra high	177	200	225

Table 6: Critical Bands for Men

N0	Pitch Registers	Lower Limits	Center Frequency	Upper Limits
1.	Extra low	106	120	132
2.	Low	133	150	169
3.	Mid	170	185	211
4.	High	212	240	265
5.	Extra high	266	300	337

Table 7: Critical Bands for Women

Fry (1958:141) also states that it is customary for speakers to produce their utterances in "one key and for musical modulation to take place between groups." He goes on to clarify that speakers are unaware of the key at which they produce their utterances. It is the researcher's job to find what that key is. When a speaker model has been chosen, it is incumbent on the researcher to understand at what octave level he/she produces various utterances. This understanding is part of building a speaker model. Additional concepts needed to describe intonation patterns accurately are the following: **plateaus**, **rises**, **falls**, **intonation nucleus**, and **melodic components,** and **melodic ratio**.[11]

Let's apply these terms to Sentence 11 to understand the intonation patterns of Louise, our speaker model. At a macrolevel, declarative utterances are characterized by a **declination**, that is, a gradual fall in pitch or in sonority from the beginning of the utterance to its end. The highest intonation unit may contain rhythmic groups which are discernible by their patterns of initial rises and terminal falls. The tree diagram of Sentence 11, as displayed in Figure 5, reflects this hierarchical organization. The higher IP dominates two lower IPs, which also dominate smaller **rhythmic groups**, represented as IP'. In Figure 5, there are two clauses (i.e., two main verbs) but four intonation units.

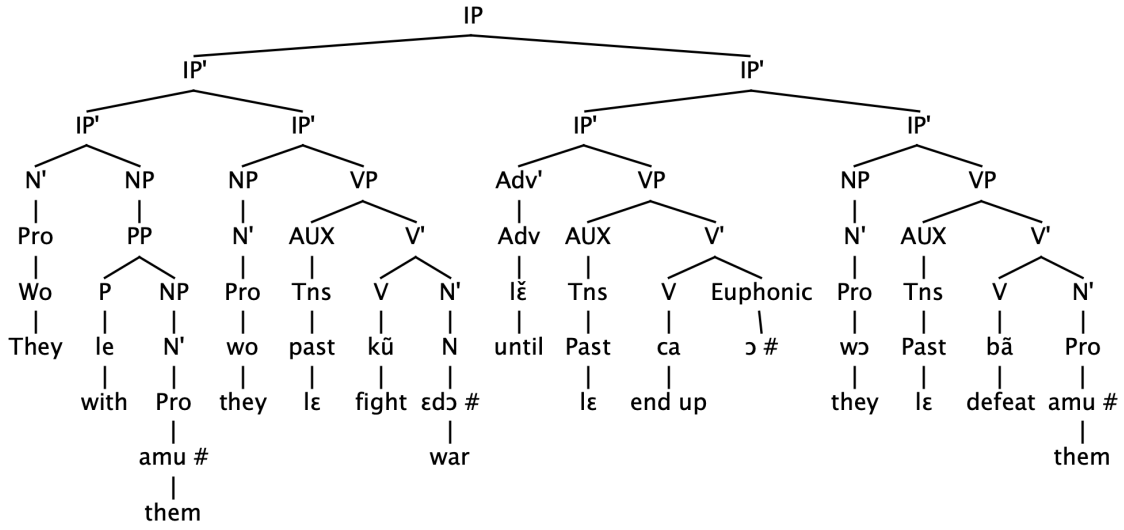


Figure 5: Tree Diagram of Sentence 11

In intonation studies, experts prefer using the phrase "Intonation Phrase" instead of the familiar word "sentence" because intonation does not always correlate with the grammatical concept of "sentence." The four rhythmic groups are as follows:

[11] Author 1 is currently experimenting with a simplified prosodic annotation system that combines ToBI and insights from Critical Bands.

1. Rhythmic Group 1: [wò lé àmú]
2. Rhythmic Group 2: [wɔ̀ kṵ̀lɛ̀ dɔ̀]
3. Rhythmic Group 3: [lɛ̌ càlɛ̌ ɔ́]
4. Rhythmic Group 4: [wɔ́ bà̰lɛ̀ àmú]

The four spectrograms below correspond to the four rhythmic groups. The intonation pattern in each group is characterized by a pitch tract (blue line) and by an intensity tract (yellow line). The breaks in the blue line represent instances when voiceless segments are produced. The color contrast used by Praat helps to visualize what goes on inside of each rhythmic group.

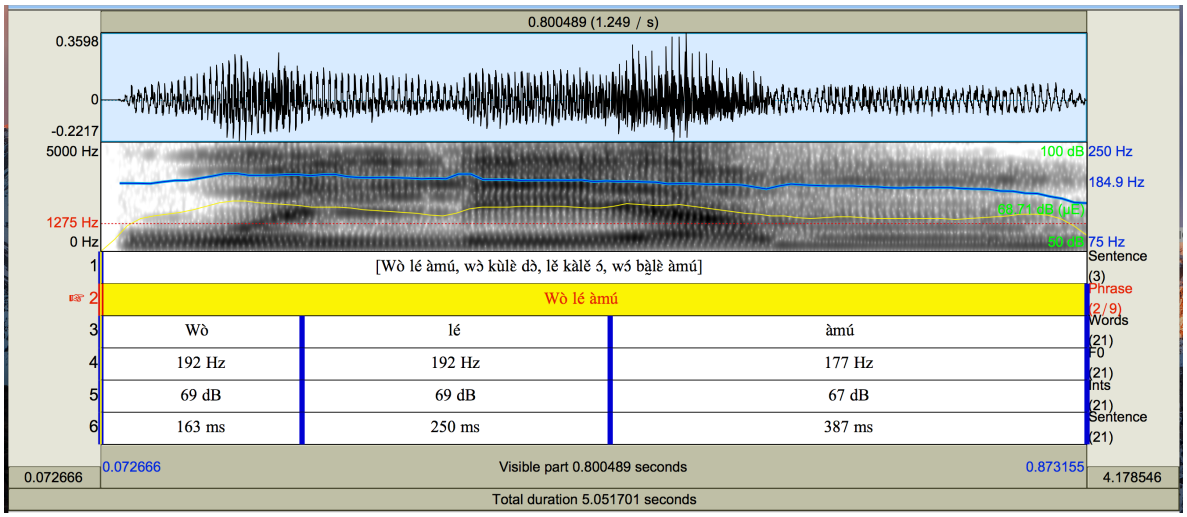


Figure 6: Intonation of Sentence 11-Phrase1

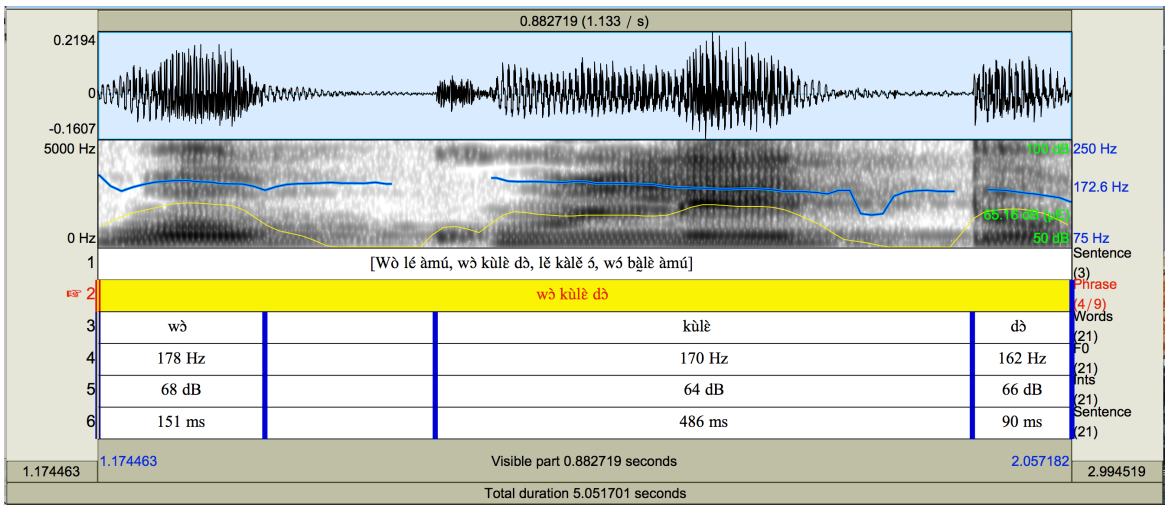


Figure 7: Intonation of Sentence 11- Phrase2

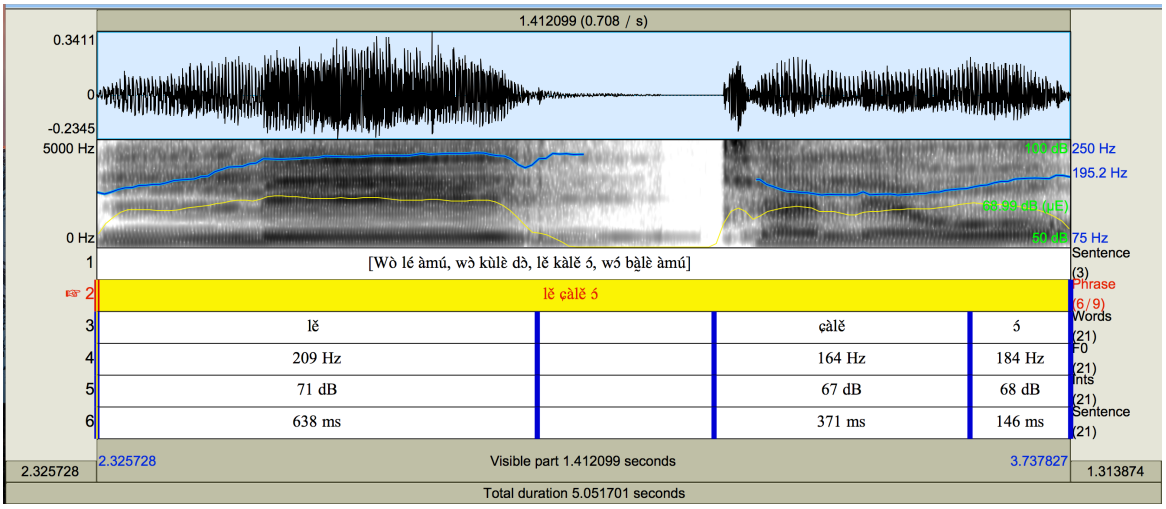


Figure 8: Intonation of Sentence 11- Phrase3

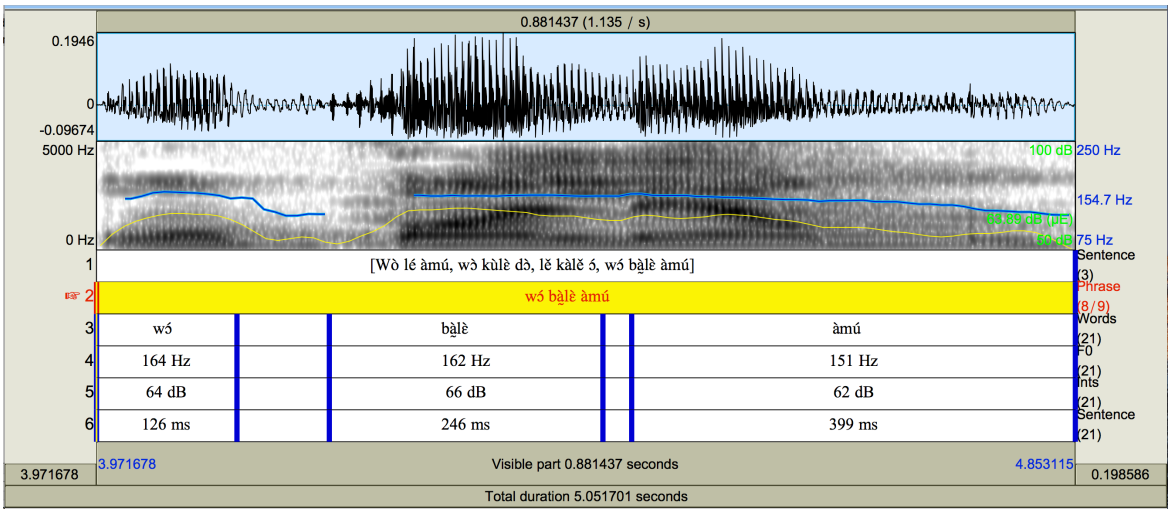


Figure 9: Intonation of Sentence 11- Phrase4

Speech is a three-dimensional physical entity that consists of frequency, intensity, and duration. Therefore, a study of intonation patterns must call on information from these three prosodic domains. They are examined separately in the paragraphs below, but they work in tandem to highlight intonation patterns.

5.1 Pitch Profile in Natural Speech and in Speech Synthesis

The key at which Louise produced Sentence 11 corresponds to the mid-pitch register on the octave scale in Table 7. However, she concludes her utterance in low pitch register scale. A micro-level analysis reveals that the utterance has three **pitch plateaux**. A pitch plateau occurs when the acoustic distance of pitch level between two successive words is less than 5 Hz. The first plateau occurs between [wò] (192 Hz) and [lé] (192 Hz), the second between [àmú] (177 Hz) and [wɔ̀] (178 Hz), and the last between [wɔ́] (164 Hz) and [bà̱lè] (162 Hz). It is worth noting that the last plateau occurs before the terminal fall. In other words, a pitch declination begins from [wɔ́] (184 Hz) and ends with [àmú] (151 Hz).

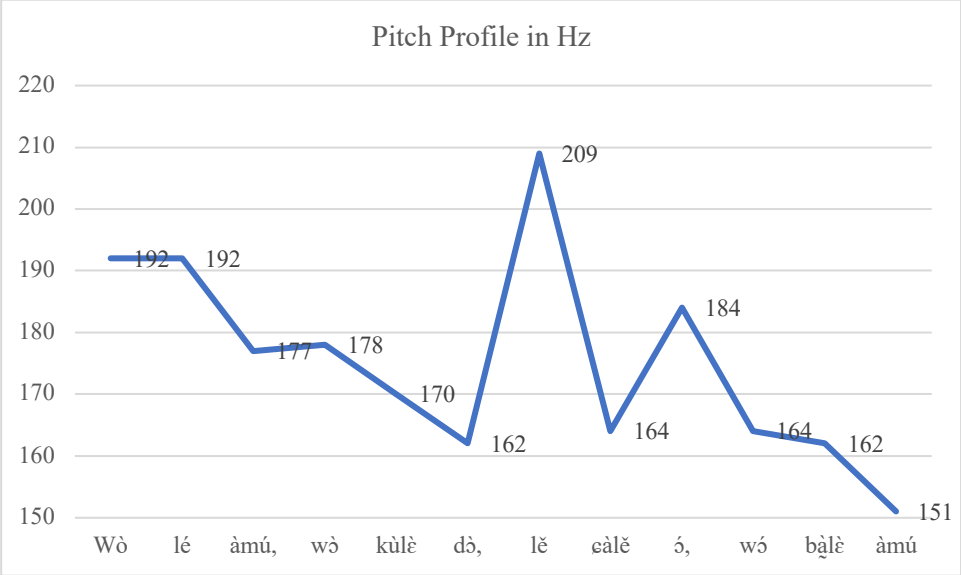


Figure 10: F0/Pitch Profile of Sentence 11

Figure 10 also shows us that there are six **pitch falls**: between [lé] (192 Hz) and [àmú] (177 Hz), between [wɔ̀] (178 Hz) and [kṵ̀lɛ̀] (170 Hz), between [kṵ̀lɛ̀] (170 Hz) and [dɔ̀] (162 Hz), between [lé] (209 Hz) and [càlě] (164 Hz), between [wɔ́] (184 Hz) and [bàl̰ɛ̀] (162 Hz), and between [bàl̰ɛ̀] (162 Hz) and [àmú] (151 Hz). There are two **pitch rises**: the one between [dɔ̀] (162 Hz) and [lě] (209 Hz) and between [ɔ́] (164 Hz) and [wɔ́] (184 Hz). The combination of pitch plateaus, falls, and rises make up the **pitch melody** components of the utterance. All in all, there are 11 pitch melodies. Of these, there are three pitch plateaux that are not auditorily salient to the naked ear, 6 pitch falls, and 2 pitch rises. Consequently, the relative functional load (RFL) of pitch for the utterance is **72.72%**, that is, 8 auditorily salient pitch levels divided by 11 possible movements.

Moreover, a cursory look at Figure 10 shows clearly that **[lě]** is the **intonation nucleus** of the utterance. There is a steep rise of 47 Hz from [dɔ̀] (162 Hz) to [lě] (209 Hz) and a steep fall of 45 Hz from there to [càlě] (164 Hz). A brief comment must be made about what the intonation nucleus means for the overall auditory perception of utterances. When pitch movements are displayed graphically as in Figure 10, the uninformed reader may be misled into thinking that the visual display correlates with auditory reality. It does not! Auditory attention centers around the intonation nucleus at the exclusion of most of the pitch movements in the utterance. In Fry's experiments, he noticed that the size of the step increase in frequency was inconsequential for the perception of pitch. He explains why:

> Change in fundamental frequency differs from change of duration and intensity in that it tends to produce an all-or-none effect, that is to say the magnitude of the frequency change seems to be relatively unimportant while the fact that a frequency change has taken place is all-important (Fry 1958:151).

This quote is extremely important in understanding the concept of intonation nucleus. It means that when we listen to somebody speak, we disregard (subconsciously, of course) slight changes in pitch. Only the word with the highest F0 arrests our auditory attention. So, in listening to Sentence 11, the hearer will identify **[lě]** as the intonation nucleus because it has the highest F0. Even though there are other frequency modulations, the ear ignores them and gravitates around **[lě]** because it is the intonation nucleus of the utterance. In general, the

hearers perceive intonation as robotic or monotonous when an utterance contains more pitch plateaux than pitch movements (rises and falls).

5.2 Sonority Profile in Natural Speech and in Speech Synthesis

Intensity is also perceived on a logarithmic scale, meaning that the naked ear cannot perceive sonority differences of less than 3 dB between speech signals (see Koffi 2020:2-27 for a detailed explanation of what intensity measurements mean in linguistic analysis). Table 8 offers a principled way of making sense of decibel measurements rendered by speech analysis software:

N0	**Speech Levels**	**Intensity Levels in dB(A)**
1.	Quiet whisper	40
2.	Moderately quiet speech	50
3.	Normal conversation	60-64
4.	Elevated/classroom speech	65-75
5.	Loud speech	76-86
6.	Shout	87-97

Table 8: Speech Intensity Levels

When we examine Sentence 11 in light of Table 8, we see that Louise spoke at an elevated speech level. Furthermore, we see that the utterance is dominated by six sonority plateaux, that is, instances when the sonority difference between two consecutive words is less than 3 dB. They occur between [wò] (69 dB) and [lé] (69 dB), between [lé] (69 dB) and [àmú] (67 dB), between [àmú] (67 dB) and [wɔ̀] (68 dB), between [kṵ̀lɛ̀] (64 dB) and [dɔ̀] (66 dB), between [wɔ́] (64 dB) and [bà̰lɛ̀] (66 dB). There is only one sonority rise: the one between [dɔ̀] (64 dB) and [lě] (71 dB). There are four auditorily sonority falls: between [wɔ̀] (68 dB) and [kṵ̀lɛ̀] (64 dB), between [ɔ́] (68 dB) and [wɔ́] (64 dB), [lě] (71 dB) and [ɛàlě] (67 dB), and between [bà̰lɛ̀] (66 dB) and [àmú] (62 dB).

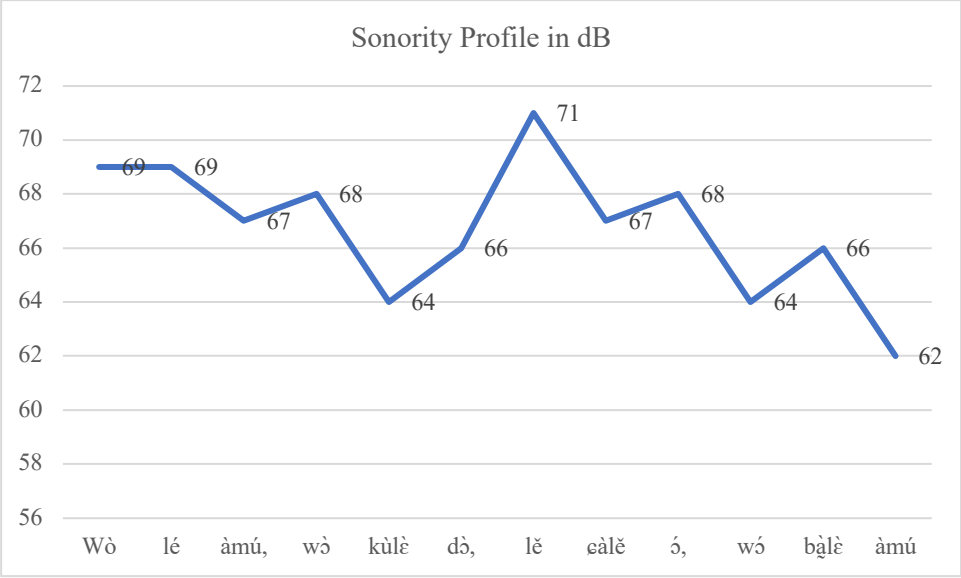


Figure 11: Sonority Profile Sentence 11

The RFL of sonority for the utterance is 45.45%, that is, five auditorily salient sonority levels divided by 11 possible sonority movements. Here, as in the pitch analysis, the word **[lě]** is the sonority nucleus of the utterance. Sonority rises steeply by 5 dB until it peaks at **[lě]** and drops

precipitously by 4 dB on the next word. We see again that **[lě]** is the **sonority nucleus** of the utterance. Many researchers have noted that sonority plays a rather insignificant role in intonation.

5.3 Duration in Natural Speech and in Speech Synthesis

The duration correlate of intonation has not been sufficiently studied. Consequently, there are not many thresholds on which we can count. Our main source comes from Klatt (1976:1210-1211). He gives us an important clue regarding word duration in naturally occurring utterances. He notes that, typically, words last from 150 to 250 msec. Furthermore, he notes that words that are emphasized are 10 to 20% longer than those that are not. In analyzing word duration, we should not lose sight of the fact that multisyllabic words are longer than disyllabic words, which are in turn longer than monosyllabic ones. This said, Klatt (1976:1211) also notes that syllables at the end of sentences and pauses are longer than those that occur elsewhere in the sentence. They are lengthened anywhere from 60 to 200 msec compared to occurrences elsewhere in an utterance. For languages that have open syllables, that is, languages where most syllables end with a vowel, the vowel before the pause can be very long. Since Betine is an open-syllable language, we can expect each word with a final vowel in an intonation unit to be somewhat elongated. The JNDs for perceiving one speech signal as being longer than another is 10 msec for signals lasting less 200 msec, 20 msec for those lasting less than 300 msec, and 30 msec for signals less than 400 msec (see Koffi 2021:7-8). Temporal distances less than these JNDs are not perceived by the naked ear.

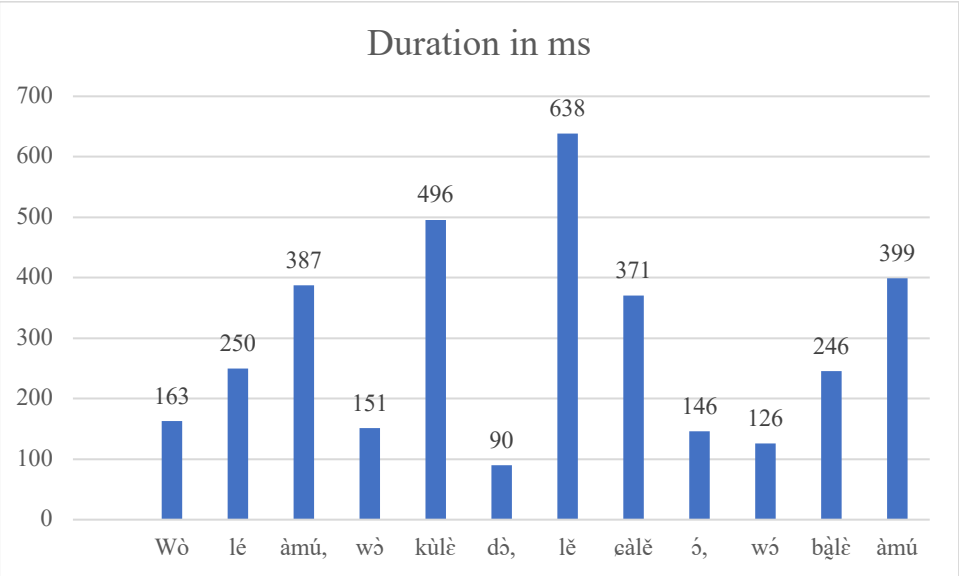


Figure 12: Duration Profile of Sentence 11

When the aforementioned insights are applied to Sentence 11, we see clearly that **rhythmicity** plays an important role in the overall prosody of the utterance. The durational distances between consecutive words are longer than the JNDs listed above. This means that the tempo of the utterance is normal, not too hurried, not too slow. In building a speaker model for speech synthesis, one should ensure that the tempo of the speaker is as natural as possible.

Three of the four words that occur immediately before a phrasal boundary in each rhythmic group are long or elongated. The only exception is the word [dɔ̀] but there is a reason. If we examine the tree diagram in Figure 5, we see that the full word is [ɛ̀dɔ̀]. However, in Betine a deletion rule applies when two homophonic vowels occur right next to each other. The word [kù̱lɛ̀] ends with the vowel [ɛ̀] and the word [ɛ̀dɔ̀] begins with the same vowel [ɛ̀].

For this reason, the [ɛ̀] of [ɛ̀dɔ̀] gets deleted. This explains why [dɔ̀] is the shortest word in the utterance. Another important prosodic characteristic of Betine is the euphonic particle [ɔ́] that occurs at the end of the rhythmic group [lɛ̌ càlɛ̌ ɔ́]. This particle occurs very often when an intensifier/degree adverb is used. Klatt's observation that words that are emphasized in an utterance are up to 20% longer is verified in Sentence 11. In this case, word [lɛ̌] is 28% longer than [kʊ̀lɛ̀] even though the latter has two syllables, and the former has only one syllable. In fact, it is longer than any other word in the utterance. Here again, we see that [lɛ̌] caries prosodic prominence in the utterance. Pitch, sonority, and duration converge to show that **[lɛ̌]** is the intonation nucleus of the utterance. The prosodic behavior of words such as **[lɛ̌**] must be accounted for because they contribute to a naturally sounding synthesized speech.

6.0 Summary

This tutorial has covered relevant issues that are the necessary first steps before a TTS system can be developed for a critically endangered or under-documented language. The main topics have to do with gathering sufficient data for building a robust language model and carefully choosing a speaker model. Additional topics include text normalization and the optimal sampling frequency, i.e.,10000 Hz. The software model was discussed but no firm recommendation is made because we are yet to experiment with Python and C+++.[12] Since MATLAB is a proprietary software and requires subscription, it is unlikely to be the first choice of many language planners. Praat is the ideal software for acoustic feature extraction because it is free and widely accessible. The last installment of the tutorial devoted considerable attention to intonation because not much has been written on it. There is every reason to believe that the psychoacoustic model discussed in this paper can yield naturally sounding synthesized intonation patterns. There is still much to do. Future tutorials will address other aspects of TTS synthesis such as information storage and retrieval.

ABOUT THE AUTHORS

Ettien Koffi, Ph.D. linguistics (Indiana University, Bloomington, IN) teaches at Saint Cloud State University, MN. He is the author of five books and author/co-author of several dozen articles on acoustic phonetics, phonology, language planning and policy, emergent orthographies, syntax, and translation. His acoustic phonetic research is synergetic, encompassing L2 acoustic phonetics of English (Speech Intelligibility from the perspective of the Critical Band Theory), sociophonetics of Central Minnesota English, general acoustic phonetics of Anyi (a West African language), acoustic phonetic feature extraction for application in Automatic Speech Recognition (ASR), Text-to-Speech (TTS), and voice biometrics for speaker verification. Since 2012, his high impact acoustic phonetic publications have been downloaded **28,515** times (analytics provided by Digital Commons) and read **13,680** times (analytics provided by Researchgate.net) as of February 2022. He can be reached at enkoffi@stcloudstate.edu.

Mark C. Petzold, Ph.D. Electrical Engineering (University of Colorado, Colorado Springs) is a professor of Computer Science at St. Cloud State University, where he teaches in the Computer Science and Cybersecurity programs. He is a co-principle investigator of a National Science Foundation S-STEM grant, and has published papers and presented on measures of student belonging. Previously, he worked at Varian Medial Systems as a high voltage and microwave engineer, and for Pericle Communications Company, where he consulted for

[12] Author 2 is translating the Fortran codes that Klatt (1980) used to build his speech synthesis into Python and other coding languages.

several major international airports on communication technology as well as performing patent reviews in the cellular phone industry. He can be reached at mcpetzold@stcloudstate.edu.

Note: The course referred to in this paper was the subject of an article entitled "Students Helping Preserve African Indigenous Language," that appeared in *St. Cloud State Magazine*, Fall 2021/Winter 2022, pp.10-11.

Appendix

RECIT D'ORIGINE DES EOTILE

A Story about the origin of the Eotile People

1. jè ɣi̱ kɛ̄ kɔ̀dìvwár mlɔ̱ nísà bì kɛ̄é sɥ̱ gānà nɛ̄ jɛ́ bètībè ɔ́ ɲą̱ sɥ́ gāná
/Nous/savoir+PRES./que/Côte d'Ivoire/dans/Hommes/beaucoup/quitter+PRES./
Ghana/mais/nous/Betibe/nous+ne/venir+PRES./Ghana/
« Nous savons que la majorité des habitants de la Côte d'Ivoire sont originaires du Ghana. »
We know that most of the people who live in Côte d'Ivoire come from Ghana

2. jɛ́ nɛ̀mjɛ̄ mɥ̱ ɔ́ tòtòlè jɛ́ kɛ̄ bètībè sɥ̱ gbógbó mlɔ̱
/Nos/ancêtre/PL./PART./dire+ACC./nous/que/betibe/quitter+ACC./lagune/dans/
« Nos ancêtres nous ont dit que les Bétibé sont sortis de l'eau. »
Our ancestors have told us that the Betibe come from water

3. wò sɥ̱ gbógbó mlɔ̱ cɛ́ wɔ̀ tàpílɛ̆ wò gbòlè ŋ̀gbą́ fɔ́
/Ils/quitter+PRES./lagune/dans/que/ils/sortir+ACC./ils/demeurer+ACC./sur/
« C'est de l'eau qu'ils sont sortis pour s'établir sur la terre ferme. »
They came from the water and estblished themselves on firm ground
4. ǹsá mɔ̱ lé àmū wɔ̀ tàpílɛ̆ wò gbòlè ŋgbą́ fɔ̀ wò dí ŋbàƗí ɔ̀hūmą̀
/Homme/qui/avec/eux/ils/sortir+ACC./ils/demeurer+ACC./terre/sur/ils/appeler+ACC./N'gban-dji Ohouman/
« L'homme qui les a conduit hors de l'eau pour s'établir sur la terre ferme s'appelle N'gbandji Ohouman. »
The person who led them from the water to live on firm ground is called Ngbandji Ohouman

5. m̀mlɛ̄ mɔ̱ wɔ̀ sɥ̱ gbógbó mlɔ̀ wɔ̀ tàpílɛ̆ ɔ́ wò pèŋé èplí mɔ̱ wɔ̀ dɛ̄ wɔ̀ mɔ̀nɔ̀bàká
/Moment/où/ils/quitter+PRES./lagune/dans/ils/venir/PART./ils/rester+PRES./lieu/que/il/appeler+PRES./il/Monobaha/
« Quand ils sont sortis de l'eau, ils ont demeuré dans le lieu que l'on appelle Monobaka »
After they sprang up from the water, they lived in a place called Monobaka

6. mɔ̀nɔ̀báká wàlɛ̀ mą́cą́ báká kò mɔ̱ wò ɔ̀gbɔ̄ báká fɔ̀
/Monobaka/était/village/grand/un/qui/ils/être/île/grand/sur/
« Monobaka était un grand village qui était situé sur une grande île. »
Monobaka was large village on a large island.

7. mą́cą́ cɔ̌ nísà ŋɥ́ cɛ́ jɛ́ sà
/Village/là/homme/deux/qui/nous/diriger+PRES./
« Ce village-là était administré par deux personnes. »
Two people ruled that village.

8. àmú mn̰ wɔ̀ sà mn̰cn̰ ókó dī wɔ̀pú sɨ̰ŋɛ̰ ɔ̀kɔ̀ àbà dī wɔ̀pú nìgbēní
/Ceux/qui/ils/diriger+PRES./village/un/s'appeler+PRES./WopouSiguin/un/aussi/s'appeler+PRES./Wopou Nigbeni/
« Ceux qui administrent ce village, un se nomme Wopou Siguin et un autre Wopou Nigbeni. »
One administrator of the village was called Wopou Siguin, and the other Wopou Nigbeni.

9. wɔ̀pú sɨ̰ŋɛ̰ lé wɔ̀pú nìgbēní wɔ̀ sàlɛ̀ mn̰cn̰ lě m̀mlɛ̄kɔ̀ mlɔ̰ ɛ̀ŋípú wɔ̀ sʊ̰ gānà fɔ̃ wɔ̀ bă
/WopouSiguin/avec/WopouNigbeni/ils/diriger+ACC./village/jusque/moment/un/dans/Agni/ils/quitter+PRES./Ghana/dans/ils/venir+PRES./
« Wopou Siguin et Wopou Nigbéni administrèrent le village jusqu'au temps où les Agni sont arrivés du Ghana. »
Wopou Siguin and Wopou Nigbeni ruled the village until the time when the Agni arrived from Ghana.

10. àmú lě ɔ̀dɔ̀àkʊ̰ wɔ̀ bàlɛ̀ wɔ̀ dɔ̌ndɔ̀lɛ̀ àmú
/Ils/tenir+PRES./guerre/ils/venir+ACC./ils/trouver+ACC./eux/
« Ils (Agni) sont venus les (Eotilé) trouver avec la guerre »
They (the Agni) came and found the Eotile and waged war against them.
wɔ̀ bàlɛ̀ wɔ̀ dɔ̌ndɔ̀lɛ̀ àmú.

11. wɔ̀ lé àmú wɔ̀ kʊ̰lɛ̀ dɔ̀ lě càlɛ̌ ɔ́ wɔ́ bn̰lɛ̀ àmú
/Ils/avec/eux/ils/battre+ACC./guerre/finir/PART./ils/chasser+ACC./eux/
« Ils firent la guerre jusqu'à ce qu'ils(les Eotilé) soient vaincus et chassés à la fin. »
They waged war against the Eotile until they defeated them. Finally, they chased them out.

12. wɔ̀ bn̰lɛ̀ àmú kɩ̰Ɓɛkɛ̀ wɔ̀pú sɨ̰ŋɛ̰ lé wɔ̀pú nìgbēní mɔ̰ wàlɛ̀ èhīmlɔ̰ ànú ɔ́ àmú ɔ̰tó ǹsn̰ tòlɛ̀ ɛ̀ɓlā fɔ̃
/Ils/chasser+ACC./eux/parceque/WopouSiguin/et/WopouNigbeni/qui/être+ACC/frère/deux/PART./leur/bouche/entendre+NEG./cause/femme/sur/
« Ils ont été chassés parce que Wopou Siguin et Wopou Nigbeni qui sont deux frères étaient en désaccord au sujet d'une femme. »
They were chased out of their village because the two brothers, Wopou Siguin and Wopou Nigbeni, no longer saw eye to eye because of a woman.

13. éhìmlɔ̰ mù ànú cɔ̌ ɔ́ ɔ̀kɔ̀ jɔ̀ klû wɔ̀ jímlɔ̰mjɛ̰ jɔ̀
/Frère/PL./deux/là/PART./un/aimer+PRES./ il/frère/femme/
« Ces deux frères là, un courtise la femme de l'autre »
The two brothers fell in love with the same woman.

14. tòlɛ̀ cɔ̌ ɔ́ fɔ̃ m̀mlê mɔ̰ ɛ̀ŋípú ɔ́ lé wɔ̀pú sɨ̰ŋɛ̰pú ɔ́ wɔ̀fɔ̀ wō kʊ̰ ɔ́
/Cause/cela/sur/moment/qui/Agni/PART./avec/WopouSiguin+sien/PART./PRON.réfléchi/ils/battre+PRES./PART./
« A cause de cette raison, au moment où les Agni et les partisans de Wopou Siguin se guerroyaient, »
For this reason, when the Agni were waging war against Wopou Siguin's troops

15. wɔ̀ jìmlɔ̰ mjɛ̰ a̰ būká wɔ̀ ɛ̀ŋípú ɔ́ bn̰lɛ̀ wɔ̀pú sɨ̰ŋɛ́
/Il/frère/ACC.+NEG/aider/lui/Agni+PL./ vaincre+ACC./WopouSiguin/
« Son frère ne lui a pas prêté main forte permettant ainsi aux Agni de vaincre

Wopou Siguin/ »
His brother (Wopou Nigbeni) did not come to his rescue. So, the Agni defeated Wopou Siguin.

16. wò sṳ̀ ɔ̀gbɔ̄ fɔ̀ wò sòlè wò dèdèlè mlɔ̰̀
/Ils/venir+PRES./île/sur/ils/lever+ACC/ils/disperser+ACC./dans/
« Ils quittèrent l'île et se dispersèrent partout »
They (Wopou Siguin and his warriors) fled and were scattered about

17. ǹcô kɔ̀lè gàná ǹcô kɔ̀lè àƗɛ́kɛ́ ǹcɔ́ àbà bàlɛ̀ frɔ̰̀sí
/Certains/partir+ACC/Ghana/certains/partir+ACC./Adiaké/certains/aussi/venir+ACC
./France/
« Certains partirent au Ghana, certains partirent à Adiaké, certains aussi sont vénus
au quartier France. »
Some went to Ghana, some settled in the town of Adiake, some went to live in a suburd called France.

18. cɔ̌ cɛ́jè mɔ̰̀ jè wò vìtré
/Ceux-là/c'est/nous/qui/nous/être/Vitré/
« Ceux-là, c'est nous qui sommes à Vitré. »
The people who went to live there, that is us, who live presently in Vitré.

References

Butcher, Andrew. 2013. Research Methods in Phonetic Fieldwork. In. M. J. Jones and R. Knight (Eds.), *The Bloomsbury Companion to Phonetics*, pp. 57-78. New York: Bloomsbury.

Cheveigné, Alain de. 1999. Formant Bandwidth Affects the Identification of Competing Vowels. *International Congress Phonetics Science 14*: 2093-2096.

Clements, Nick G. 2000. Phonology. In H. Bernd and D. Nurse (Eds.), *African Languages: An Introduction,* pp.123-160. New York: Cambridge University Press.

Crystal, David. 2000. *Language Death.* New York: Cambridge University Press.

Dunn, H. K. 1961. Methods of Measuring Vowel Formant Bandwidths. *Journal of the Acoustical Society of America* 33: 1737-1745.

Dusosky, Scarlet. 2022. Speech Digitalization, Coding, and Nasali(ized) Vowel Synthesis: Demonstration with Beti, a Critically Endangered Language. *Linguistic Portfolios* 11: 81-90.

Eberhard, D.M., Simons, G.F., and Fenning, C. D. (Eds,.). 2019. *Ethnologue: Languages of Africa and Europe*. Twenty-second Edition. SIL, Dallas.

Foba, Kakou A. 2009. Syntaxe de l'Éotile: Language Kwa de Côte d'Ivoire. Parler de Vitre. Doctorat Unique. Universite Felix Houphouet Boigny. Abidjan: Côte d'Ivoire, W. Africa.

Fry, Dennis. B. 1958. Experiments in the Perception of Stress. *Language and Speech* 1 (2): 126-152.

Fry, Dennis. B. 1955. *The Physics of Speech.* New York: Cambridge University Press.

Gropou, Djaki C. 2006. Morphophonologie de l'Eotile (Betinene). In Morphophonologie des Langues Kwa de Côte d'ivoire, pp. 261-273, ed by. Firmin Ahoua and William R. Leben. Cologne, Germany: Rudiger Koppe Verlag.

Hale, Krauss. 1992. Language Endangerment and the Human Value of Linguistic Diversity. *Language* 68 (1) 35-42.

Jurasfsky, Daniel and James Martin H. 2000. *Speech and Language Processing: An Introduction to Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

Hérault, G. 1983. L'Éotile. In *Atlas des Langues Kwa de Côte d'Ivoire*, pp. 403-424. Abidjan, Côte d'Ivoire: Institute de Linguistique Appliquée, Université d'Abidjan.

Klatt, Dennis H. 1990. Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence. *JASA 59 (5):1207-1211.*

Klatt, Dennis H. 1980. Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence. *JASA 59 (5):1207-1211.*

Klatt, Dennis H. 1976. Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence. *JASA 59 (5):1207-1211.*

Krauss, Michael. 1992. The World's Language in Crisis. *Language* 68 (1):4-10.

Koffi, Ettien. 2021. Language Endangerment Threatens Phonetic Diversity. Acoustics Today 17 (2) 23-31.

Koffi, Ettien. 2020. A Tutorial on Acoustic Phonetic Feature Extraction for Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) Applications in African Languages. *Linguistic Portfolios* 10: 130-153.

Koffi, Ettien. 2012. *Paradigm Shift in Language Planning and Policy: Game-Theoretic Solutions*. De Gruyter Mouton, New York.

Ladd, Robert D. 2008. Intonational Phonology. Second Edition. New York: Cambridge University Press.

MATLAB. 2020. Version 9.8.0.1396136 (R2020a). Natick, Massachusetts: The MathWorks Inc.

Oppenheim, Alan V. and Schafer, Ronald W. 2010. Discrete-Time Signal Processing. New York: Pearson.

Rabiner, Lawrence. R. 1998. "Machine Recognition of Speech. In *Handbook of Acoustic Phonetics*. In M. J. Crocker. John Wiley & Sons, Inc.1263-1270. New York: Wiley-Interscience Publication.

Rabiner, Lawrence R. and Schafer Ronald. W. 2011. *Digital Speech Processing: Theory and Applications.* New York: Pearson.

Rabiner, Lawrence R. and Schafer Ronald. W. 1978. *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Rabiner, Lawrence. R. and Juang, Biing-Hwang. 1993. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Silverman, Kim, Mary Beckman, John Pitrelli, Mary Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. TOBI: A Standard for Labeling English Prosody. Proceedings of the *2nd International Conference of Language Processing (ICSLP 92)*, October 12-16, 1992. Banff, Alberta, Canada.

Snell, Roy C. and Milinazzo, Fausto. 1993. Formant Location from LPC Analysis Data. IEEE Transactions on Speech and Audio Processing. April 1993.

t' Hart, Johan. 1981. Differential Sensitivity to Pitch Distance, Particularly in Speech. *Journal of the Acoustical Society of America*, 69 (3): 811-821.

Welmers, Wm. E. 1973. *African Language Structure*. Los Angeles: University of California Press.