Linguistic Portfolios

Volume 13 Article 8

2024

SPEECH SYNTHESIS BY SYLLABLE A CONCATENATION: EXPERIMENTATION WITH BETINE

Ettien Koffi
St. Cloud State University

DANIEL FABRES

SWATHI R. PINGILI

GOPIKRISHNA K. CHAVA

Linguistic Portfolios – ISSN 2472-5102 –Volume 13, 2024 | 126

SPEECH SYNTHESIS BY SYLLABLE A CONCATENATION: EXPERIMENTATION WITH BETINE

ETTIEN KOFFI WITH DANIEL FABRES, SWATHI R. PINGILI, GOPIKRISHNA K. CHAVA[1]

ABSTRACT

The endangerment of minority languages has reached pandemic proportions. No country or continent is spared. The editors of Ethnologue (2019:14-15) report that 2,923 of the world's 7,111 languages are critically endangered. McWorther (2003:257-8) adds that, statistically speaking, "A language dies roughly every two weeks." UNESCO (2010) projects that 90% of the world's indigenous languages will be dead by the end of 2100. In Africa alone, Kandybowicz and Torrence (2017:3) note that 201 languages of the estimated 2000 languages have already died, and 308 others are on the brink of extinction. Minority languages worldwide are in such deplorable conditions that the United Nations (UN) has sounded the alarm about this impending linguistic catastrophe of apocalyptic proportions. It has declared 2022-2032 the International Decade of Indigenous Languages (IDIL 22-32). A simpler and nimbler speech synthesis that is based on syllable concatenation is described in this paper with the hope that it can be duplicated to revitalize critically endangered languages in Africa and elsewhere.

Keywords: Beti Language, Beti Speech Synthesis, Beti Text-to-Speech Synthesis, Syllable Concatenation, Syllable Synthesis, Text-to-Speech Synthesis

1.0 Introduction

Over the past several years, my colleague Dr. Mark Petzold, an engineering and computer science professor and I have joined forces and are teaching a computational linguistics course that attracts engineering, computer science, and linguistics students. The goal of the course is to put our assets together and come up with a simpler and nimbler speech synthesis model that can help revitalize critically endangered languages, in keeping with Crystal's (2000:141), Postulate 6, which states that "An endangered language will progress if its speakers can make use of electronic technology." The language that we have chosen for our experiments is Betine, also known as Eotile, represented by the International Standardization Organization (ISO) number of (639-3-eot). Most of our past and continuing efforts have focused on formant-based synthesis. Yet, we are open to pursuing other methods such as syllable or phone concatenation. In this paper, we explore syllable concatenation as a possible avenue for text-to-speech (TTS) synthesis. It has been successfully tried elsewhere. We are experimenting with it because it has been argued that it requires less data but yields excellent results. Ordinarily, Rabiner and Schafer (2011:929) note that, TTS synthesis systems require no less than 1,000 sentences. Yet, the data we are experimenting with has only 16 sentences and 221 words. These words were collected by Author 1 who secured a Proposal Enhancement Grant (PEG) from St. Cloud State University, his home

[1]**Authorship responsibilities:** The preposition "with" is used instead of the coordinating conjunction "and" because Daniel Fabres, Swathi Pingili, and Gopikrishna Chava did not contribute to the writing of this paper, except for the codes and the syllabification algorithm. Gratitude is expressed to Dr. Mark Petzold for overseeing the coding aspect of this project. Author 1 bears full responsibility for any analytical or interpretive errors in this publication, except for the codes and algorithms.

Linguistic Portfolios – ISSN 2472-5102 –Volume 13, 2024 | 127

institution.[2] Even with this limited amount of data, the preliminary results reported here are encouraging. The syllable concatenation approach is described here with the hope of encouraging others to try it on minority languages. The paper has five main installments. The first provides a succinct overview of the Betine language. The second discusses the results of a sociolinguistic survey that was carried out to see if speech synthesis is worth attempting. The third and four installments are devoted theoretical methodological issues. The final section addresses the thorny issue of prosody in speech synthesis.

1.1 The Beti Story of Origins

A window into how an indigenous group sees itself is often reflected in the story of their origins. This is also true for the Beti people who tell the following story:

> *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 Anyi 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 Anyi 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 Anyi were waging war against Wopou Siguin's troops, his brother did not come to his rescue. So, the Anyi defeated Wopou Siguin. Wopou Siguin and his warriors fled and were scattered about. Some went to Ghana, some settled in the town of Adiaké, some went to live in a suburb now nicknamed "France." The people who went to live there, are presently in the villages of Vitré 1 and Vitré 2.*

A terminological clarification is needed on the uses of the terms "Eotile," "Beti," "Betine," and "Betibe." "Eotile" is the term used in official documents in Côte d'Ivoire, West Africa. However, because of its derogatory connotations, we have opted for the terms that the people of Beti descent use to refer to themselves and to their language. The suffix <-ne> refers to the language or the speakers of the language, while the suffix <-be> designates people of Beti ethnicity.

This story of Beti origins has been passed on from one generation to the next. Koffi and Petzold (2022:28) note that, except for their subterranean origins that cannot be independently confirmed, the history of wars and occupations encapsulated in the story are true. They are amply documented by Perrot (2008). Her 260-page book recounts that the war that the Anyi waged against them took place in 1725. This led to the subsequent resettlement of the fugitives in the villages of Vitré 1 and Vitré 2. Those who remained in Adiaké, the largest town in Betiland were assimilated to the Anyi and have lost their language. The escapees settled on an island 55

[2] The grant in the amount of $6,940 allowed Author 1 to travel to the Betine-speaking areas of Côte d'Ivoire to collect sociolinguistic data and to pay the graduate students.

kilometers (some 35 miles) away from Adiaké. Later, French colonial officials resettled them into the adjoining villages of Vitré 1 and Vitré 2 which are separated by only two miles. They managed for a while to keep their native language, but the combined pressure from Anyi and French has taken its toll. Given the significance of the Beti Story of Origins to the community, we used it to experiment with our speech synthesis model.

1.2 An Overview of the Language

Betine is a moribund language with an 8a status. Eberland et al. (2019:14) describes a language with this status as follows: "The only remaining active users of the language are members of the grandparent generation and older." Twenty-three years ago, the language had only 200 speakers left. The last true monolingual speaker died in 1993. Most of the elderly speakers who can still use the language are bilinguals with Anyi and/or French as their dominant language(s). To make matters worse, Betine is not well documented at all. To date, we have been able to find only five published sources on the language. Two additional sources are referenced in publications, but we have not been able to locate them. Betine is genetically related to the lagoon languages spoken in south-eastern Côte d'Ivoire. It belongs to the Akan sub-family, and to the bigger Kwa family. The available sources indicate it has 23 consonants, as shown in Table 1.

	Bilabial	Labiodental	Alveolar	Palatal	Velar	Labiovelar	Glottal
Stop	p/b		t/d	c/ɟ	k/g*	k͡p/g͡b	
Fricative		f/v*	s/z*				h
Nasal	m		n	ɲ	ŋ		
Liquid			l/r*				
Glide				j		w	

Table 1: Consonant Chart

The consonants with an asterisk are allophones of the phonemes to their left. When pairs of phonemes are listed, the one to the left is voiceless. Betine has nine oral vowels, five of which are nasal or nasalizable before the nasal segments /m, n, ɲ, ŋ/. Hérault (1983:407) opines that the vowels /ɪ, e, ʊ, o/ are not nasalizable. Suprasegmentally, the vowels in Table 2 can bear two types of phonemic tones, a high and a low.

Front	Central	Back
i		u
ɪ		ʊ
e		o
ɛ	a	ɔ

Table 2: Vowel Chart

1.3 Syllable Structure and Syllable Count

Since syllable concatenation is the main framework used in this paper, we must provide a basic definition of what constitutes a syllable. The classic definition is that a consonant and vowel combine to form syllables. More will be said about this in 4.2 when we discuss syllable algorithms. Suffice it to say for now that the canonical syllable structure of Betine is prototypical of many sub-Saharan African languages (Clements 2000:140-9). Betine has five main types of monosyllables:

1. V
2. CV
3. C^1VC^2, where C^2 is a nasal consonant, usually /m, n, ɲ, ŋ/
4. C^1C^2V where C^2 is either /l/ or /r/
5. NCV where /N/ can surface as /m, n, ɲ, ŋ/ in accordance with the place of articulation of the first consonant in the root.[3]

When a vowel constitutes a syllable all by itself, as in the first example, that vowel must be /e, ɛ, a, ɔ, o/. The other vowels, /i, ɪ, u, ʊ/ cannot be syllabic. When syllable types are computed, Betine ends up with **2,318** syllables in its inventory.[4] This relatively small number of syllables makes Betine a prime candidate for syllable-based synthesis. In fact, the same can be said for many sub-Saharan African languages. Maddieson (1983:22) gives examples of syllable inventories of nine prototypical world languages, including Yoruba and Gã, two West African languages. Yoruba has only 582 syllables, while Gã has 2,331 syllables. The latter has 13 more syllables than Betine.

2.0 The Sociolinguistic Survey

It is recommended that before attempting to revitalize a moribund language, a sociolinguistic survey must be carried out to gauge the speakers' interest. This is what Author 1 did in the summer of 2021. He secured a grant in 2019 and had hoped to carry out the sociolinguistic survey in the summer of 2020 but the Covid-19 pandemic forced a postponement of plans. After a delay of one year, he was finally able to travel to Betineland to survey the speakers and collect their attitude and sentiments about the current state of their language, and to gauge whether or not a TTS synthesis is worth building for them. The survey was administered in Adiaké and Vitré 1/ Vitré 2 villages, the two main areas mentioned in the Betine Story of Origins. Two hundred participants took part in the survey. The respondents fell into the following demographic categories:

1. Five age groups (18-30, 31-45, 46-60, 61-75, and older)
2. Six educational levels (no schooling, elementary school, middle school, high school, BA, MA, doctorate)
3. Eight occupations (farmers, college students, teachers, healthcare professionals, civil servants, private sector, others).

2.1 Survey Results and Analyses

The survey had 22 questionnaires from which The Statistical Consulting and Research Center at St. Cloud State University generated a 234-page document. Six questionnaires are singled out to help probe deeper into the respondents' sociolinguistic attitudes and sentiments.

[3] This is called "homorganic assimilation", and it is widespread in Akan languages (Fromkin et al. 2017:220-2). In most NCV instances, /N/ is a syllabic nasal.

[4] The count goes as follows: 23 consonants x 9 oral vowels = 207 syllables. Four syllabic nasals x 207 vowels = 828 syllables. Five syllabic vowels x 207 CV syllables = 1,035 syllables. 23 consonants x 5 nasal vowels = 115 syllables. Two consonant clusters (C^1C^2), in which C^1 must not be a nasal consonant, and C^2 is either /l/ or /r/. There are 18 syllables that can occur in C^1 position. The total possible syllable combination is 2,318.

Q 5: What language(s) do you use in your everyday life? The responses are as follows: 27% use French, 45% use Anyi, 3.5% use Betine, while 24% use other languages. It is worth noting that there are two respondents from Adiaké and five from Vitré 1 and Vitré 2 who reported using Betine every day. Unsurprisingly, they are in the 61-75 or older age brackets. Our survey confirms the status of 8a given to Betine in *Ethnologue* (2019:125), meaning that the language is used only by people of the "*grandparent generation and older*."

Q 11: Are you aware that Betine is an endangered language? elicited the following responses: the vast majority (90.5%) responded "yes," 8% said "no," 1.5% were "unsure." These responses show that the Betibe are keenly aware that their ancestral language is critically endangered.

Q 12: If Betine is no longer spoken, I will be: ... Here, the participants were given four choices: 72% responded with "very sad," 17% chose "sad," 10% checked "it doesn't matter," and one person did not offer an opinion. All in all, 89% of the respondents are saddened by the current state of their language.

Q 15: Are you proud to be Betibe? This question was meant to evaluate any putative correlation between ethnicity and language. The responses were as follows: 93% were "proud," 5.5% were "somewhat proud," and 1.5% said that "ethnicity did not matter" to them. We note in passing that it is hard for outsiders to know who an ethnic Betibe is or is not. For this reason, census results vary wildly. However, Perrot (2008:7-8) estimates Betibes to be between 2,000 and 3,000. She disputes the 1998 census that put their number at 20,000. She contends that the methodology of the census was flawed because the government lumped all the inhabitants of Adiaké together as though they were all ethnic Betibes. The Anyi and many other ethnic groups live in Adiaké.

Q 16: If I can learn Betine, I will jump on this opportunity. This question is important to our research team because how it was answered would determine whether or not a TTS synthesis will be worth pursuing. The vast majority, 82.5% said they would be "very interested," 10.5% said they would be "interested," 4% would be "somewhat interested," and 3% responded that it would be "a complete waste of their time." Taken together, Questions 15 and 16 show that ethnolinguistic loyalty ranks very high (93%) in both communities. In other words, language revitalization efforts are very likely to succeed because, as noted by Grinevald (2007:67), strong ethnic identity correlates well with success in language revival. In fact, Perrot (2008:15) notes that in the 1960s, after the government cracked down brutally on an attempt by the Anyi to secede from CIV, there was enthusiasm among the Betibe to return to their native tongue. This effort was led by a charismatic leader by the name of Aiko Etyua Émile (Perrot 2008:19-20). But this attempt was short-lived. We interpret the high ethnolinguistic score to mean that, if the speakers have the chance to use their language, they will.

Q 17: In your opinion, which language threatens Betine the most? This question was meant to gauge the respondents' knowledge as to which language(s) has caused/is causing the death of Betine. Overwhelmingly, the respondents from Adiaké (95%) blamed Anyi as the "killer" language, but only 13% from Vitré 1/Vitré 2 saw Anyi the same way, even though 42% of them responded that they use Anyi in their daily life. Even so, 84% of people from Vitré 1/Vitré 2 see

French as the "killer" language. Both groups are right, and their responses underscore two different sociolinguistic realities about language endangerment. In some instances, indigenous African languages cause a lot of harm to other minority languages (see Brenzinger 2007:196-8, Bokamba 2008:100). In other instances, it is the language of the former colonial power that is the "killer" language. The respondents from Vitré 1/Vitré 2 are correct in identifying French as the "killer" language because Abidjan, a megacity of several million people, has sprawled to the doorstep of their villages. The completion of a four-lane freeway in 2020 has sped up an unprecedented urbanization process so much so that Bassam which used to be 20 miles away from Abidjan is now a suburb. Prior to the freeway, Vitré 1/Vitré 2 were secluded and almost inaccessible. Now, they are less than two miles away from the freeway. During the fieldwork, the chief and his notables did not mince words describing the ruthless tactics that developers are employing, enticing villagers to sell their lands. Since urbanization is synonymous with modernization, and since French is associated with both urbanization and modernization, the respondents from Vitré 1/Vitré 2 have every reason to believe that French is more menacing than Anyi. Blench (2007:152), Krauss (2007:2), and Mesthrie (2008:329) discuss the role of urbanization and modernization in language death across sub-Saharan Africa and elsewhere.

3.0 Idealistic Language Documentation

Rabiner contends that a full-blown TTS system would require a minimum of 1,000 sentences. Time is of the essence for critically endangered languages. With a corpus of 1,000 sentences, one can embark on a full morphological, syntactic, and lexical analysis of the language. From such a large corpus, one can have illustrative samples of all the 2,318 syllables that occur in the language. However, for this experiment, we have only the Betine Story of Origins read by a 30-year-old female who earned a Ph.D. in linguistics. Our consultant can read but cannot speak the language, though she is ethnically a Betibe. She is the same speaker model used in Koffi and Petzold (2022). More is said about speaker model in 6.0. The two most basic requirements for concatenated syllable synthesis are the International Phonetic Alphabet (IPA) and the Arpabet transcription of the data. Both are illustrated below.

3.1 Exemplification of IPA Transcription

Gambarage (2017:457) refers to IPA transcription as a technical code that linguists use to describe the sounds of a language. It is an efficient system of documentation that has been in use since its inception in 1888. O'Grady et al. (2017:67) note that there are five distinct levels of transcriptions, with varying shades of narrow and broad transcriptions. The one used to transcribe the Betine Story of Origins corresponds to IPA transcription Level 5.

> [jè ɣi̱ kɛ̄ kɔ̀dìvwár mlɔ̰ nísà bì kɛ̄é sṵ gānà nɛ̄ jé bètībè ɔ́ ɲą́ sṵ́ gāná// jé nèmjɛ̄ mṵ ɔ́ tòtòlè jé kɛ̄ bètībè sṵ gbógbó mlɔ̰// wò sṵ gbógbó mlɔ̰ cé wɔ̀ tàpílɛ̌ wò gbòlè ŋ̀gbą́ fɔ̀// ǹsá mɔ̰ lé àmū wɔ̀ tàpílɛ̌ wò gbòlè ŋgbą́ fɔ̀ wò dí ŋbàJí ɔ̀hūmą̀// m̀mlɛ̄ mɔ̰ wɔ̀ sṵ gbógbó mlɔ̀ wɔ̀ tàpílɛ̌ ɔ́ wò pèŋé èplí mɔ̰ wɔ̀ dɛ̄ wɔ̀ mɔ̀nɔ̀bàká// mɔ̀nɔ̀báká wàlɛ̀ mą́cą́ báká kò mɔ̰ wò ɔ̀gbɔ̄ báká fɔ̀// mą́cą́ cɔ̌ nísà ɲṵ́ cé jé sà// àmú mɔ̰ wɔ̀ sà mą́cą̂ ókó dī wɔ̀pú sị̀ŋɛ́ ɔ̀kò àbà dī wɔ̀pú nìgbēní// wɔ̀pú sìŋɛ́ lé wɔ̀pú nìgbēní wɔ̀ sàlɛ̀ mą́cą́ lě m̀mlɛ̄kò mlɔ̰ èŋípú wɔ̀ sṵ gānà fɔ̀ wɔ̀ bă// àmú lě ɔ̀dɔ̀akṵ́ wɔ̀ bàlɛ̀ wò dŏndòlè àmú// wò lé àmú wɔ̀ kṳ̀lè dɔ̀ lě càlɛ̌ ɔ́ wɔ́ bą̀lè àmú//. wɔ̀ bą̀lè àmú kị̀Ɉɛ̌kɛ̀ wɔ̀pú sị̀ŋɛ́ lé wɔ̀pú nìgbēní mɔ̰ wàlè èhīmlɔ̰ ą̀nú ɔ́ àmú ɔ̰tó ǹsą́ tòlɛ̀ èɓlā fɔ̀// éhìmlɔ̰ mù ą̀nú cɔ̌ ɔ́ òkò jò klû wò jímlɔ̰mjɛ̰́ jɔ̀// tòlè cɔ̌ ɔ́ fɔ̃ m̀mlê mɔ̰ èŋípú

ɔ́ lé wɔ̀pú sɩ̰ŋɛ́pú ɔ́ wɔ̀fɔ̀ wō kṵ ɔ́// wò jìmlɔ̰ mjɛ̰ ą̰ būká wɔ̀ ɛ̀ɲípú ɔ́ bą̰lɛ̀ wɔ̀pú sɩ̰ŋɛ́// wò sṵ ɔ̀gbɔ̄ fɔ̀ wò sòlɛ̀ wò dèdèlè mlɔ̰// ǹcô kɔ̀lɛ̀ gànáǹcô kɔ̀lɛ̀ àƗɛ́kɛ́ ǹcɔ́ àbà bàlɛ̀ frą̰sí// cɔ̌ cɛ́jɛ̀ mɔ̰ jè wò vìtré//]

A transcription such as this provides relevant insights about the phonological rules that undergird pronunciation. In transcribing Betine, diacritics are used to indicate nasal vowels, as in [ḭ] of the word [ɣḭ]. The tilde is written under the vowel, in order to leave space over the vowel for tone diacritics, as in [à] and [ú] of [àmú]. Vowel length is indicated by the circumflex diacritic, as is the case of [ɛ̌] in [tàpílɛ̌]. The vowel [ɛ̌] has a low-high contour tone in which the first vowel has a low tone and the second has a high tone. Both tones are superimposed on the phone [ɛ].

3.2 Exemplification of Arpabet Transcription

The Arpabet transcription method was introduced in the 1970s because many of the symbols used by the IPA were (and still are) incompatible with the ASCII characters on standard computer keyboards. The Arpabet is in every regard like the IPA with the added advantage of being compatible with computers and smart devices. It is the preferred transcription method in TTS synthesis. In his 2020 paper, Koffi expanded the Arpabet system to African languages and beyond. Because the Arpabet transcription takes a lot of space, only the first sentence of the Betine Story of Origins is used to exemplify this transcription method:

Jè	ɣḭ	kɛ	kɔ	dìvwár		mlɔ̰	nísà		bì	kēé
YEY1	GHIN	KEH	KOA2	DIYVWAA1R		MLOAN2	NIY1SAA2		BIY2	KEY0EY1
sṵ	gānà		nɛ̄	jɛ́	bètībè		ɔ́	ɲą̰	sṵ	
SUWN2	GAANAA2		NEH	YEH1	BEY2TIYBEY2		OA2	NYAAN2	SUWN1	
gāná										
GAANAA1										

By convention, the Arpabet transcription uses capitalization to set it apart from normal spelling. Consonants often are represented by a single grapheme, unless two graphemes are necessary, as is the case of <GH> that stands for [ɣ], <NY> for [ɲ], and <NG> for [ŋ]. Vowels are always represented by two graphemes. The expanded Arpabet proposed by Koffi (2020) uses numbers to symbolize tones: **1** written after a vowel indicates a high tone, **2** is used for a low tone, and **0** stands for a mid-tone. With these additions, the Arpabet can be used to transcribe any language. No wonder that in a relatively short period of time this paper has been downloaded more than 2,000 times from all over the world. Needless to say, the IPA and Arpabet transcriptions are time-consuming and onerous processes. But they are indispensable parts of TTS synthesis (Rabiner and Schafer 2011:86-87).

4.0 Rationale for Synthesis by Syllable Concatenation

The syllable appears to be the optimal unit of speech synthesis for as many as six reasons. The usefulness of the syllable in speech synthesis was recognized early on by Klatt (1987:758). He considered it for English but quickly changed his mind because English has a very large inventory of syllables, more than 10,000 in fact. For languages such as English syllable-based synthesis is not a good option. Languages with large inventories of syllables include Vietnamese and Thai which according to Maddieson (1984:22) have respectively 14,430 and 23,638 syllables. Yet, for languages with a smaller syllable inventory such as Betine, and this includes many sub-

Saharan African languages, syllable-based synthesis is optimal. It has yielded excellent results for other languages with similar syllable sizes. This includes Mandarin (Ouh-Young et al. 1986), Malay (Tiun et al. 2011:68-74), Indonesian (Samsundin et al. 2001) and (Soedirdjo et al. 2011), Tamil (Rama et al. 2002), and Punjabi (Singh and Lehal 2011). This explains why this approach is being tried on Betine.

4.1 Exemplification of Syllable Inventory Documentation

There are numerous steps that one must follow when using this approach. First and foremost, an audio recording of the language must be available. The Betine data was recorded and sampled (in Praat) at 44100 Hz, 16 bits per sample. Now that computers have incredible storage capabilities, there is no reason to downsample any file. Secondly, an IPA transcription, as illustrated in 3.1, must be done. It is what one uses for extracting individual syllables. Individual syllables are extracted and saved as "mono" instead of "stereo." The .wav file format is ideal because, according to Singh and Lehal (2011:3), it reduces storage capacity by 90%. Sentence 3 is used to illustrate the several steps in the process.

σ Count	Spelling	IPA	Arpabet	File Name
1.	wo	wò	WOW2	WOW2.wav
2.	sun	sṵ̀	SUWN2	SUWN2.wav
3.	gbó	gbó	GBOW1	GBOW1.wav
4.	gbó	gbó	GBOW1	GBOW1.wav
5.	mlɔn	mlɔ̰	MLOAN2	MLOAN2.wav
6.	cɛ́	tʃɛ́	CHEH1	CHEH1.wav
7.	wɔ	wɔ̀	WOA2	WOA2.wav
8.	ta	tà	TAA2	TAA2.wav
9.	pí	pí	PIY1	PIY1.wav
10.	lɛ́	lɛ́	LEH1	LEH1.wav
11.			WOW2	WOW2.wav
12.	gbo	gbò	GBOW2	GBOW2.wav
13.	le	lè	LEY2	LEY2.wav
14.	mgbán	ŋ̀gbá̰	MGBAAN1	MGBAAN1.wav
15.	fɔ	fɔ̀	FOA2	FOA2.wav
16.	nsá	ǹsá	NSAA2	NSAA2.wav

Table 3: Exemplification of Syllabification Rules and Patterns

This sentence contains 16 individual syllables. Even if the same syllable occurs numerous times in a text, one and only one prototype is selected. Only the best syllable is kept in the syllable bank. The first column contains the word in normal orthography (spelling). Since the 1970s, the Côte d'Ivoire government has decreed how Ivorian languages are to be spelt. The second column lists the word in IPA transcription, the third reflects Arpabet transcription of the syllable. The fourth column contains the file name of individual syllables. It is important that the Arpabet transcription and the file name be identical so as to facilitate search and retrieval operations.

The Betine Story of Origins yielded 350 syllables, of which 107 occurred two or more times. In other words, the story contains only 243 unique syllables out of the **2,318** in the Betine

syllable inventory. This represents only 10.48% of all possible syllables. The syllable make-up of the words in the Story of Origins is as follows:

1. Three of the 221 words have four syllables.
2. Twenty-two of the 221 words have three syllables.
3. Seventy-two of the 221 words have two syllables.
4. One hundred twenty-four of the 221 words are monosyllabic.

Statistically, monosyllabic words represent 35.42% of the corpus, and disyllabic words are 20.57%. Both types of syllables account for 56% of all the words in the story.

4.2 Syllabification Algorithm

After the preliminary work was done, the research team built an algorithm that helps to parse the words found in the Betine Story of Origins accurately. The parser can be diagrammed as follows:

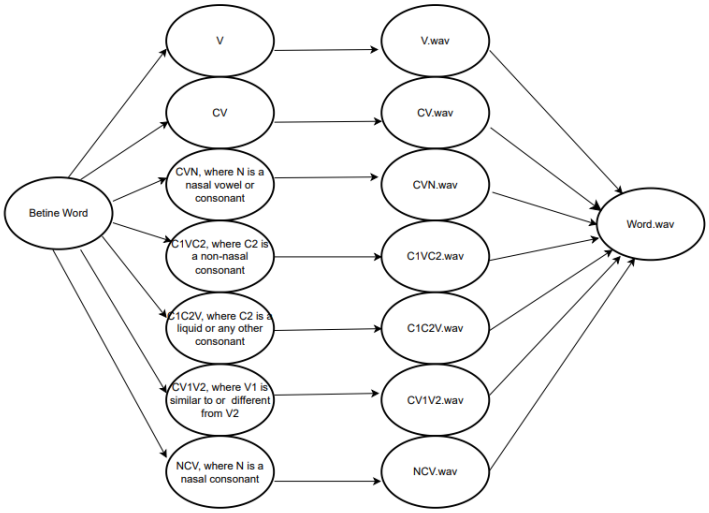

Figure 1: Syllabification Algorithm

The algorithm is based on the classic definition of the syllable that takes the nucleus (usually the vowel) as the main rule for syllabification. In any given word, the number of vowels corresponds to the number of syllables. For the sake of simplicity, we ignore syllabic nasals in words such as [m̀mlɛ̄kò] and [m̀mlɛ̂] where [m̀] is a syllabic nasal. So, for any given word, the parser looks for a vowel. When it finds one, it marks the unit as containing one syllable. Thereafter, it looks back to identify and categorize the preceding consonant. The algorithm looks back because Betine is a codaless language, which means that every syllable ends with a vowel. The algorithm works very well with the Betine Story of Origins. However, it must be borne in mind that this story contains only 10.48% of the available syllables in the language. The remaining 89.52% do not occur in this text. As noted previously, for a robust TTS system, a minimum of 1,000 sentences is required. By contrast, the Story of Origins has only 16 sentences.

4.3 Glitches and Solutions

The research team has made great progress, but only through trials and errors because we had no blueprints for a syllabification algorithm to follow. The following are examples of problems that we encountered and solved. There is no doubt that many more will surface when more sentences become available for analysis.

1. **Distortion:** The very first experiment with the word [ta.pi.lɛ] was an abject failure. When the three syllables were merged and played back, there was a lag between each syllable, which created a huge distortion. However, this problem was quickly fixed when a code was written for an audio joiner in Python 3.9. Thereafter, one could not tell any difference between the words produced by the model speaker and the words that were concatenated via synthesis.

2. **Syllabification difficulties with nativized loanwords:** Samsundin et al. (2001:3) encountered similar difficulties in their speech synthesis for Malay. They explained the issues as follows: "Although structures are already available, the synthesizer might still not be able to segment all words correctly. This is because there are a lot of loanwords in Malay vocabulary. The origins of those loanwords are also varied such as English, Arabic, Sanskrit, Indian and Javanese." In the Beti Story of Origins, "Côte d'Ivoire" is pronounced as [kɔ̀dìvwár]. This nativized pronunciation does not conform to Betine syllabification rules. If it is parsed as [kɔ̀.dìv.wár], the syllables [dìv] and wár] are problematic because Betine is an open-syllable language. The only consonant that can occur in the coda is a nasal. Yet, if the syllable is parsed as [dì], we run into another problem because the last syllable becomes [vwár], which also violates syllable onset rules because no Betine word has a complex onset unless the first consonant is a nasal or the second segment is a liquid. Given that language contact between French and Betine began as far back as 1692 (Perrot 2008:37), we expect many loanwords from French. Consequently, it is anticipated that the algorithm in Figure 1 will have to be revised many times to account for loanwords. For speech synthesis to work optimally, one must account for loanwords.

3. **Indexing errors:** The algorithm parsed the words correctly, but trisyllabic words were read incorrectly. So, instead of [ta.pi.lɛ], it first read <ta.pi> and dropped <.lɛ>. This perplexed us for about two weeks until it was discovered that it was an indexing problem that needed fixing. A code was written to fix it.

With these fixes, the algorithm yields very good results. The concatenated words are read as naturally as the ones produced by the model speaker. We cannot claim victory yet, because our sample size is very small, accounting for only 10.48% of possible Betine syllables. The remaining 89.52% of the 2,318 syllables are still unaccounted for. Nevertheless, great progress was made with the available corpus. What remains is a matter of collecting more data and increasing the syllable count until we reach **2,318**.

4.4 An Onerous Process

It took Author 1 a considerable amount of time to manually syllabify all the words in the Story of Origin, an example of which was displayed earlier in Table 3. With languages dying at such a high rate, a human-supervised syllabification is a time-consuming and an inefficient

process. For concatenative syllable synthesis to be used on a large amount of data, i.e., 1,000 sentences, an algorithm such as Figure 1 should be built to extract syllable data automatically or semi-automatically. The robustness of the syllabification algorithm can be reliably tested by comparing its results with a human-supervised extraction. It is quite likely that the algorithm will have to be tweaked gradually until the automatic syllabification matches human-supervised syllable extraction.

In order to test the robustness of our algorithm, we will need at the very least 1,000 sentences. Even though our goal is for a fully automated syllable extraction method, we may have to settle for a semi-automated system. We anticipate that sonorants, a class of speech sounds that comprises the approximants [j] and [w], the liquids [l] and [r], and the nasals [m, n. ɲ, ŋ], will pose serious challenges to automatic syllabification because their spectral characteristics are often hard to distinguish from those of the vowels that immediately follow them. Even so, we expect automatic syllabification to be straightforward when the consonant in the onset is a stop or a fricative. Even if we manage to syllabify some words semi-automatically, this will be a giant leap forward because it will drastically reduce the amount of time that human-supervised syllabification requires.

5.0 Rationale for Prosodic Documentation

The goal for all TTS systems is to arrive at a naturally sounding speech synthesis. Therefore, sooner or later the issue of prosody must be addressed. Klatt (1987) laments the fact that most TTS systems lack naturalness. Great strides have been made in English and other elite languages because huge amounts of data exist. However, for a critically endangered language such as Betine, naturalness can be achieved only if a large databank is available. This means recording a wide variety of utterances and eliciting different types of grammatical constructions.

5.1 Theories of Prosody

To date, theories of prosody are not easily conducive to speech synthesis. Experimentations are being done with the ToBI (Tone and Break Indices). ToBI was originally designed for English and has to be adapted to each language (Silverman et al. 1992). Koffi (2023) has proposed a psychoacoustic model based on the Critical Band Theory (CBT). It applies more broadly to all human languages because it is based on how the human auditory-perceptual system works. Since all human beings have similar auditory and neurolinguistic systems, CBT can be used to study the prosodic patterns of any language. At its core is the view that the three acoustic correlates of speech, namely F0/pitch, intensity/sonority, and duration/rhythmicity are all important in determining prominence because the auditory-perceptual system processes prominence in three separate phases (Yost 2007:236, 246). The discrimination phase is the one during which acoustic signals are simply perceived by the ear. In the integration phase, the acoustic signals are ferried to the Central Auditory Nervous System for processing. Finally, during the resolution phase, the word with the most prominent correlate or combination of correlates wins out and is perceived as the most prominent in the utterance. The prosodic pattern of Sentence 11 can be analyzed as follows:

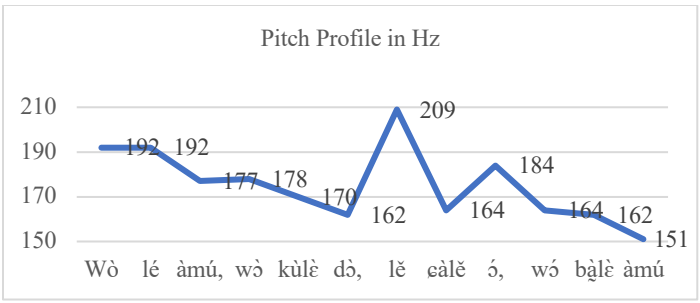

Figure 2: F0/Pitch Measurements

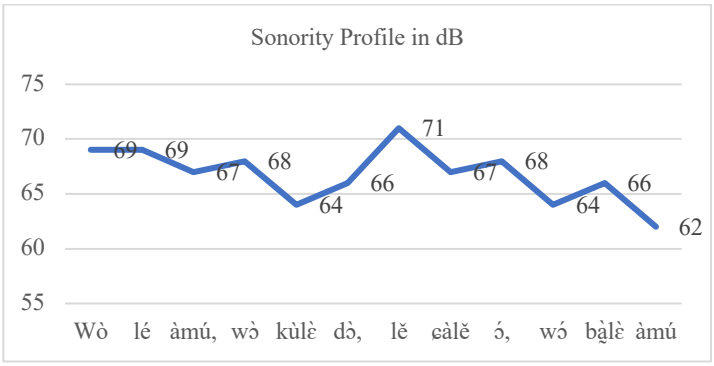

Figure 3: Sonority Measurements

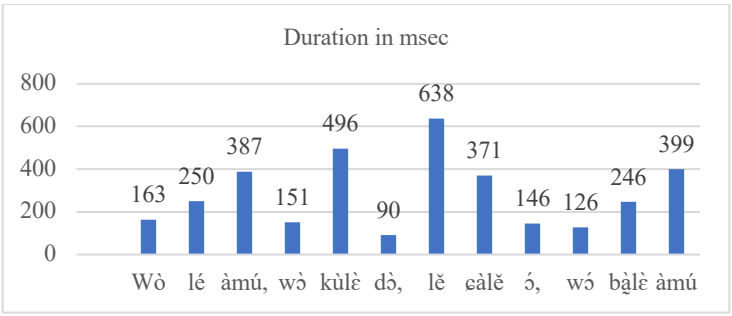

Figure 4: Duration Measurements

The word [lɛ̌] carries prosodic prominence in this utterance because it has the highest pitch, the greatest intensity, and the longest duration, even though it is a monosyllabic word. But, why? This is so because [lɛ̌] is a quasi-ideophone that underscores the fact that the combats were fierce, and the war went on for a long while. In African action narratives, words such as [lɛ̌] are often interjected to enhance the liveliness of the story. In building a TTS synthesis that replicates natural intonation, we must account for the acoustic characteristics of such words. A prosodic analysis that takes into account F0, intensity, and rhythmicity increases the naturalness of synthesized utterances.

6.0 The Speaker Model

The speaker model can be one person or a group of speakers. If it is one speaker, he or she reads all 1,000 sentences or more. If there are multiple speakers, they all read the same 1,000 sentences. There is a cost associated with having multiple voices. For one, increasing the number of speakers also increases the workload. This may in turn delay the speech synthesis project by many years, if not decades. For critically endangered languages, time is of the essence. For this

reason, we recommend having a single speaker model per language. Ideally, the selected person would be a person with a very good voice quality. There is a precedent for using a single voice. Klatt synthesized his own voice for the synthesizer that he developed. For nearly a decade or so, Siri, the voice assistant in Apple's products, was based on a single speaker's voice. Once the speaker model has been found and recorded, one extracts all the syllables needed for concatenation. In every language, some syllables occur more frequently than others. Regardless, a syllable should be represented only once in the databank. The exemplar should be the best possible pronunciation of that syllable. For Betine, it means that the syllable databank would contain **2,318** unique syllables. Needless to say, the recordings should take place in a professional recording studio.

7.0 Summary

Speech synthesis by syllable concatenation helps generate "Big Data" that can be put to a wide variety of theoretical and applied linguistic uses. Even if one were to collect no more than 1,000 sentences, the data will provide valuable insights on many aspects of Betine. From these sentences, one can learn a lot about its inflectional and derivational morphology, its word formation processes, and its syntax. All the 1,000 or so sentences should be glossed following the model in Sentence 11 to maximize using the data for other types of linguistic research:

Wò lé àmú, wɔ̀ kǹ̩lɛ̀ dɔ̀, lɛ̌ càlɛ̌ ɔ́, wɔ́ bǹ̩lɛ̀ 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.

Another benefit for critically endangered languages is that if one is able to collect 1,000 sentences, one can anticipate building a lexicon of some 4,000 words. To conclude, since smart phones are ubiquitous everywhere, speech synthesis by syllable concatenation should be give serious consideration because it can breathe new life into critically endangered languages. Heritage speakers of these languages can learn new vocabulary items and their pronunciation of words through apps and other derivative tools.

Codes[5]

These are the codes written for the experimentation phase of the project. The first is the Arpabet code and the second is the syllabification code.

Arpabet Code

```
#! python
from typing import Counter
from playsound import playsound

syllablesAndAudiosMap = {"mc": "MAO","nc": "NAO","ba":"BAA1","ka":
"KAA1","wa":"WAA","le":"LEH",
"wc":"WAO","ban":"BAN","le":"LEH","a":"AA","mu":"MUW","kin":"KIN","jee":"JHEHEH",
"ke":"KEH","wc":"WAO","pu":"PUW",
```

[5] We had hoped for a bigger grant to continue working on this project. Unfortunately, our application was not funded. We make our preliminary data available so that others can improve on them. We hope that our efforts will lead to a simpler and nimbler speech synthesis for endangered and less commonly spoken languages.

```
"ma":"MAAN","ca":"CHAAN","ko":"KOW","mɔ":"MAON","wo":"WOW","c":"AO","gbcn":"
GBAON","fc":"FAO",
"sin":"SIN","nge":"NGEH","ni":"NIY","gben":"GBEYN","mcn":"MAON","e":"EY","hin":"HI
YN"
}
def speak(word, syllablesAndAudiosMap):
    startIndex = 0
    endIndex = 1
    while(1):
        if(endIndex > len(word) or startIndex > len(word)):
            break
        subString = word[startIndex:endIndex]
        syllable = search(syllablesAndAudiosMap, subString)
        if(syllable is None):
            endIndex+=1
        else:
            playFile(syllable)
            startIndex = endIndex
def search(values, searchFor):
    if searchFor in values.keys():
        return values[searchFor]
    return None
```

Syllabification Codes

```
from typing import Counter
def splitting_into_syllables(input_word):
    count = 0
    word1 = input_word.lower()
    vowels = set("aeiouɔɪɛʊ")
    syll = list()
    temp = 0
    for letter in word1:
        if letter in vowels:
         count += 1
    if count == 1:
        print(count)
        return word1
    for index in range(1,len(word1)):
        if word1[index] in vowels and word1[index - 1] not in vowels:
            w = word1[temp: index+1]
            print(w)
            if len(w) != -1:
                syll.append(w)
                temp = index+1
    return syll
user_input = input()
```

```
print(splitting_into_syllables(user_input))
```

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), voice biometrics for speaker verification, and infant cry bioacoustics. Since 2012, his high impact acoustic phonetic publications have been downloaded **68,201** times (**47,355** as per Digital Commons analytics), (**20,846** as per Researchgate.net analytics), and several thousand downloads from Academia.edu, as of **February 2024**. He can be reached at enkoffi@stcloudstate.edu.

Daniel Fabres is a student pursuing a master's degree in computer science at St. Cloud State University. He has a one-year experience as a software developer and has experience interning with NASA at the Goddard Space Flight Center. His interests lie in technology and games of all sorts. He plans to use his knowledge and abilities for improving lives through technological development and is focused on AI development. He can be reached at daniel.fabres@go.stcloudstate.edu or at dafabres@gmail.com.

Pingli Swathi is a student pursuing a master's degree in computer science at St. Cloud State University. She has six years of experience as an Application Programmer. Her interests extend beyond technology and includes traveling. She is committed to using her knowledge and skills to help preserve critically endangered languages. She can be reached at swathi.pingili@go.stcloudstate.edu or at swathipingili28@gmail.com.

Gopikrishna Chava is currently pursuing a master's in computer science at St. Cloud State University. He has 3 years of experience as a Data Analyst. He is an enthusiastic open-source contributor, actively participating in projects that promote collaboration and innovation in the tech industry. His passion for open-source development serves both as a hobby and a platform for continuous learning and skill enhancement. He is interested in cultural explorations, particularly in the preservation of the linguistic heritage of less commonly taught languages. He can be reached at gopi.chava@go.stcloudstate.edu
or at chavagopikrishna98@gmail.com

Works Cited

Atal, B. S. 1972. Automatic Speaker Recognition Based on Pitch Contours. *Journal of the Acoustical Society of America*, 52 (6):1687-1697.

Blench, Roger. 2007. Endangered Languages in West Africa. In Matthias Brenzinger (Ed.), *Language Diversity Endangered,* pp.140-178. Berlin, Germany: Mouton de Gruyter.

Boersma, Paul, and David Weenink. 2022. *Praat*. Praat: doing Phonetics by Computer, Version 6.2.13. https://www.fon.hum.uva.nl/praat/.

Bokamba, Eyamba G. 2008. The Lives of Local and Regional Congolese Languages in the Globalized Linguistic Markets. In C. B. Vigouroux and S.S. Mufwene (Eds.), *Globalization and Language Vitality: Perspectives from Africa,* pp.97-125. New York, NY: Continuum International Publishing Group.

Brenzinger, Matthias. 2007. Language Endangerment in Southern and Eastern Africa. In Matthias Brenzinger (Ed.), *Language Diversity Endangered,* pp.179-204. Berlin, Germany: Mouton de Gruyter.

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.

Dusosky, Scarlet. 2022. Speech Digitalization, Coding, and Nasal(ized) Vowel Synthesis: Demonstration with Beti, A Critically Endangered Language. *Linguistic Portfolios* 11, 82-92.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (Eds.). 2019. *Ethnologue: Languages of Africa and Europe*, 22nd ed. Dallas, TX: SIL International.

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.

Fromkin, Victoria, Robert Rodman, and Nina Hyams. 2017. *An Introduction to Language*. 11th Edition. Boston, MA: Cengage.

Gambarage, Joash, J. Unmasking the Batu Orthographic Vowels: The Challenge for Language Documentation and Description. In J. Kandybowicz and H. Torrence (Eds.). *Africa's Endangered Languages: Documentary and Theoretical Approaches*, pp. 449-484. New York, NY: Oxford University Press.

Grinevald, Colette. 2007. Endangered Languages in Mexico and Central America. In Matthias Brenzinger (Ed.), *Language Diversity Endangered,* pp.59-86. Berlin, Germany: Mouton de Gruyter.

Hansen, Mitchell. 2022. Students Helping Preserve African Indigenous Language. *The St. Cloud State Magazine, Fall 2021/Winter* 2022, pp. 10-11.

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.

Himmelmann, Nikolaus P. and Robert D. Ladd. 2008. Prosodic Description: An Introduction for Fieldworkers. *Language Documentation and Conservation* 2: 244-274.

Kandybowicz, Jason and Harold Torrence. 2017. Africa's Endangered Languages: An Overview. In J. Kandybowicz and H. Torrence (Eds.). *Africa's Endangered Languages: Documentary and Theoretical Approaches*, pp. 1-10. New York, NY: Oxford University Press.

Klatt, Dennis H. and Laura C. Klatt. 1990. Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers. *Journal of the Acoustical Society of America*, Volume 87 (2):820-857.

Klatt, Dennis H. 1987. Review of Text-to-Speech Synthesis Conversion for English. *Journal of the Acoustical Society of America*, Volume 67 (3):971-995.

Klatt, Dennis H. 1980. Software for a Cascade/Parallel Formant Synthesizer. *Journal of the Acoustical Society of America*, Volume 82 (3):737-793.

Koffi, Ettien and Mark Petzold. 2022. A Tutorial on Formant-Based Speech Synthesis for the Documentation of Critically Endangered Languages. *Linguistic Portfolios* 11, 26-55.

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.

Krauss, Michael. 2007. Classification and Terminology for Degrees of Language Endangerment. In Matthias Brenzinger (Ed.), *Language Diversity Endangered,* pp.1-8. Berlin, Germany: Mouton de Gruyter.

Ladefoged, Peter. 2003. *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques.* Malden, MA: Blackwell Publishing.

Ladefoged, Peter. 1968. *A Phonetic Study of West African Languages: An Auditory-Instrumental Survey.* Cambridge, UK: Cambridge University Press.

Maddieson, Ian. 1984. *Patterns of Sounds: Cambridge Studies in Speech Science and Communication*. New York, NY: Cambridge University Press.

McWhorter, John. 2003. *The Power of Babel: A Natural History of Language*. New York, NY: Perennial.

Mesthrie, Rajend. 2008. Trajectories of Language Endangerment in South Africa. In C. B. Vigouroux and S.S. Mufwene (Eds.), *Globalization and Language Vitality: Perspectives from Africa,* pp.32-50. New York, NY: Continuum International Publishing Group.

O'Grady, William, John Archibald, Mark Aronoff, and Janie Rees-Miller. 2017. *Contemporary Linguistics: An Introduction*, 7th edition. New York: Bedford/St. Martins.

Ouh-Young, Ming, Chin-Jiang Shie, Chiu-Yu Tseng, and Lin-Shan Lee. 1986. A Chinese Text-to-Speech System Based Upon a Syllable Concatenation Model. *ICASSP 86*: 2439-2442.

Perrot, Claude Hélène. 2008. *Les Éotilé de Côte d'Ivoire aux XVII^e^ et XIX^e^ Siècles: Pouvoir Lignager et Religion*. Paris, France: Publications de la Sorbonne.

Pew Research Center. 2015. Cell Phones in Africa: Communication Lifeline. https://pewrsr.ch/3RDuVRZ.

Python Software Foundation. 2022. Version 3.9, https://www.python.org/.

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

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

Rama, G.L. Jayavardhana, A. G. Ramakhrisha, R. Muralishankar, and P. Prathibha. 2002. A Complete Text-to-Speech Synthesis in Tamil. *Proceedings of 2002 IEEE Workshop*, Santa Monica, CA, USA.

Samsudin, Nur-Hana, Sabrina Tiun, and Enya Kong Tiun. 2001. A Simple Malay Speech Synthesizer Using Syllable Concatenation Approach. *National Conference on Research and Development in Computer Science*.

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.

Singh, Parminder and Gurpreet Singh Lehal. 2011. Text-to-Speech Synthesis System for Punjabi Language. DOI:10.1007/978-3-642-19403-0_54

Soedirdjo, Subaryani D.H., Hasballah Zakaria, and Richard Mengko. 2011. Indonesian Text-to-Speech Using Syllable Concatenation for PC-based Low Vision Aid. *Proceedings of 2011 International Conference on Electrical Engineering and Informatics*, July 17-19, 2011. Bandug, Indonesia.

Tiun, Sabrina, Rosni Abdullah, and Enyakong Tang. 2011. Subword Unit Concatenation for Malay Speech Synthesis. *IJCSI International Journal of Computer Science* 8 (5) 68-74.

UNESCO. 2010. *Atlas of the World Languages in Danger*, 3rd ed. UNESCO, Paris, France.

Yost, William A. 2007. *The Fundamentals of Hearing: An Introduction.* 5th Edition. New York, NY: Elsevier.