Linguistic Portfolios

Volume 9 Article 11

2020

A Tutorial on Acoustic Phonetic Feature Extraction for Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) Applications in African Languages

Ettien Koffi
St. Cloud State University, enkoffi@stcloudstate.edu

Linguistic Portfolios–ISSN 2472-5102 –Volume 9, 2020 |130

A TUTORIAL ON ACOUSTIC PHONETIC FEATURE EXTRACTION FOR AUTOMATIC SPEECH RECOGNITION (ASR) AND TEXT-TO-SPECH (TTS) APPLICATIONS IN AFRICAN LANGUAGES

ETTIEN KOFFI

ABSTRACT

At present, Siri, Dragon Dictate, Google Voice, and Alexa-like functionalities are not available in any indigenous African language. Yet, a 2015 Pew Research found that between 2002 to 2014, mobile phone usage increased tenfold in Africa, from 8% to 83%.[1] The Acoustic Phonetic Approach (APA) discussed in this paper lays the foundation that will make Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) applications possible in African languages. The paper is written as a tutorial so that others can use the information therein to help digitalize many of the continent's indigenous languages.

Keywords: Acoustic Phonetics Feature Extraction, Formant Extraction, Arpabet, Automatic Speech Recognition (ASR), Text-to-Speech (TTS), Acoustic Phonetics of African Languages, Measurement of Speech Signals, Speech Digitalization, Critical Band Theory, JND

1.0 Introduction

This paper is written as a tutorial on acoustic phonetic feature extraction, an indispensable first step in Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) applications. Extracting features is time-consuming but a labor of love that can increase the functionality of indigenous languages and help them survive in the "big eat small" linguistic environment in which indigenous languages find themselves. The processes involved in acoustic phonetic feature extraction are discussed in eight sections. The first highlights the benefits of speech digitalization, the second discusses the three main approaches used in speech synthesis, the third provides a short history of the Arpabet, the fourth surveys basic theoretical issues concerning phonetic invariance, the fifth deals with extracting consonant features, the sixth focuses on vowels features, and the seventh discusses suprasegmental features. The final installment mentions briefly the additional steps not discussed in this paper but that are necessary for building full-fledged speech-enabled artificial intelligent systems for indigenous languages in Africa and beyond.

2.0 Rationale for Extracting Acoustic Phonetics Features

The phonetic and phonological diversity found in African languages makes it a prime candidate for a tutorial on feature extraction. This richness is described by Clements (2000:123) as follows:

> The African continent offers a generous sample of the great variety of phonological systems to be found in the world's languages, as well as some original features of its own. African phonological systems range from the relatively simple to the staggeringly complex. Those on the more complex end of the spectrum contain phonemic contrasts little known elsewhere in the world, rich patterns of morphophonemic alternations, and intricate tonal

[1] http://www.pewglobal.org/2015/04/15/cell-phones-in-africa-communication-lifeline/. Retrieved on November 10, 2017.

> and accentual systems, all offering stimulating grounds for phonetic and phonological study.

In other words, if we can successfully extract the acoustic phonetic features found in African languages, we can apply the acquired skills to myriads of indigenous languages around the globe, so that they too can benefit from the revolution in speech-enabled technologies going on under our very eyes. Rabiner and Schafer (1978:6-8) list some of the benefits that speech digitalization could potentially afford speakers of any language:

1. Speech Synthesis
2. Speech therapy
3. Voiced enabled assistive technologies
4. Digital transmission and storage of speech
5. Speaker verification and identification
6. Enhancement of signal quality

Howley et al. (2020:2) list six additional benefits. Without feature extraction, speakers of indigenous languages cannot use their smart devices in the same ways that speakers of English use Siri, Google Voice, Dragon Dictate, Alexa, etc. The remaining sections of the paper will take us through various steps that make speech digitalization possible.

3.0 Focus on the Acoustic Phonetic Approach

Rabiner and Juang (1993:42-67) offers a good overview of three approaches used in ASR and TTS systems. The common thread among them is that they all depend crucially on feature extraction. The three models in question are:

1) the Acoustic Phonetic Approach (APA)
2) the Pattern Recognition Approach (PRA)
3) the Artificial Intelligence Approach (AIA)

The focus of this paper is APA because it can be readily applied to any indigenous language for which basic phonological descriptions exist. Rabiner and Juang (1993: 42-3) describe it as follows:

> The acoustic-phonetic approach is based on the theory of acoustic phonetics that postulates that there exist finite, distinctive phonetic units in spoken language and that the phonetic units are broadly characterized by a set of properties that are manifest in the speech signal, or its spectrum, over time. Even though the acoustic properties of phonetic units are highly variable, both with speakers and with neighboring phonetic units (the so-called co-articulation of sounds), it is assumed that the rules governing the variability are straightforward and can readily be learned and applied in practical situation.

According to Kent and Read (2000:253), this is the preferred speech synthesis method because it is "the product of many studies in acoustic phonetics, coupled with principles of phonology." Assuming that the language under consideration has an adequate description of allophonic rules and phonotactic constraints, the researcher can move directly to feature extraction. But a few

questions must be answered first. Why is feature extraction necessary? What principles should guide features extraction? How many features are to be extracted? Fant (1998:1249) answers the first question as follows:

> One object of speech analysis is to extract essential parameters of the acoustical structure, which may be regarded as a process of data reduction and an enhancement of information-bearing elements.

Three guiding principles on what features to extract are enunciated by Baken and Orlikoff (2000:3) as follows:

1. The measurements must have a known (or at least a very likely) and specific relationship to recognized aspects of speech system physiology.
2. A measurement must have clear relevance.
3. The sole value of a measurement is in its interpretation.

Fourcin and Abberton (2008:40) add that a distinction is to be made between measurements meant for "analytical precision" and those intended for "matching human perceptual abilities." Since the latter is the goal of ASR and TTS systems, only features that capitalize on intelligibility are worth extracting and measuring. Rabiner and Schafer (1978:45) and Kent and Read (2002:247-9) answer the question of how many features to extract by highlighting F1, F2, F3, and duration correlates. However, since African languages are understudied acoustically, F0, VOT, and Center of Gravity (CoG) should also be included among the extractable features.

3.1 The Overall Architecture of APA

A quick examination of the overall architecture of APA by way of Figure 1 found in Rabiner and Juang (1993: 45) gives a glimpse of the methodological steps one must follow in feature extraction.

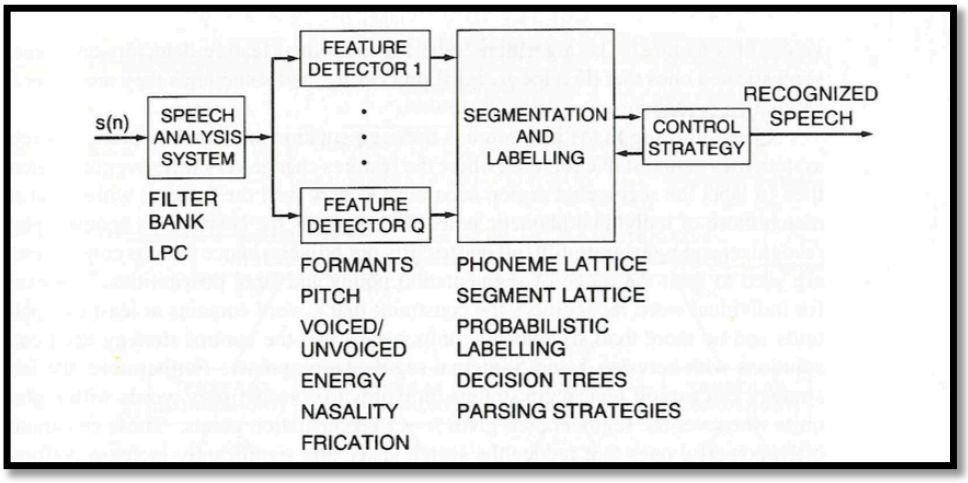


Figure 1: Overall Architecture of APA

The starting point of digitalization begins with the "Speech Analysis System" block in Figure 1. Given the popularity of Praat, it is safe to assume that it is the software that most linguists will use

to extract acoustic phonetic features. Praat is a very powerful software that offers a wide array of possibilities. Furthermore, it is free, which is music to the ears of researchers in the developing world who may not have the financial resources to afford expensive software. The second block is "Feature Detector 1 … Feature Detector Q." This is where most of the extraction activities happen. Questions related to what features to extract are listed below block 2.

The procedure that one can/must follow to extract acoustic phonetic features is illustrated by Figure 2 with the example of the approximant fricative [ɥ] in the word <ahyʊa/ɛhyʊa> (night time stories). It is a rare sound in Anyi, occurring only 10 times in a corpus of more than 5,000 words. A border is drawn around the whole word. The segment [ɥ] occurs only in nouns, as noted on the parts of speech (POS) tier. In this case, it occurred only in the singular, as also noted on the number tier. Thereafter, inner boundaries are drawn around [ɥ] by itself because it the segment from measurements are collected. The measurements appear on individual tiers: F0, F1, F2, F3, intensity, duration, and Center of Gravity (CoG):

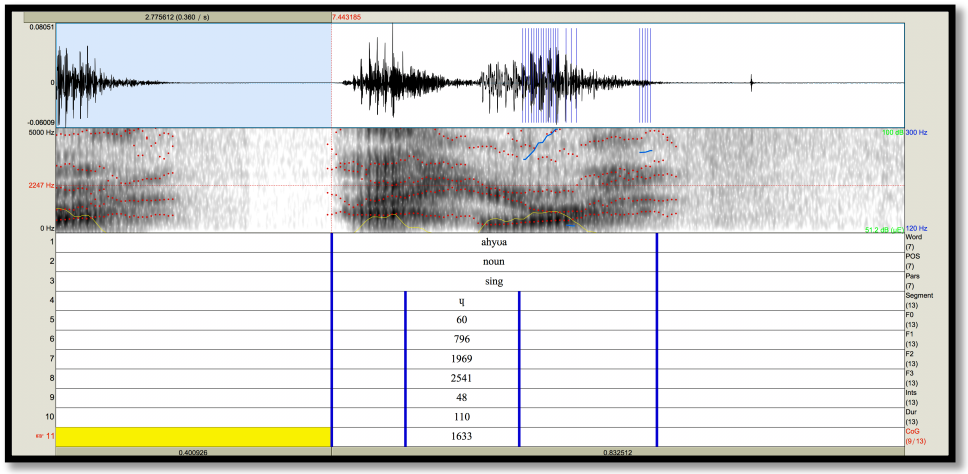


Figure 2: Annotation and Feature Extraction

The totality of these seven correlates represent the features that are extracted for the approximant fricative [ɥ]. This procedure is repeated for the consonants, vowels, and suprasegments of the language that are "essential parameters" and "information-bearing elements," in the language (Fant 1998:1249).

3.2 Methodological Issues

Feature extraction raise several methodological issues. Some relate **speech style** and others to the **number of participants**. In regard to the former, the question is whether or not the extracted features should come from words in **citation form** or in **running speech**. Ladefoged et al. (1976) and Koffi and Krause (2020) have shown that as far as most acoustic correlates of vowels are concerned, speech style has no effect on intelligibility. With regard to the latter, a minimum of **six** speakers are recommended for most acoustic phonetic studies (Ladefoged 2003:67). Jongman et al. (2000:1255) have shown that with as few as 20 participants (10 males and 10 females), one can adequately represent the speech signals of an entire speech community. However, if features are to be extracted from 20 speakers, this would represent a massive amount of data if a minimum of seven acoustic correlates were to be considered. For example, if a language has 9 oral vowels and 23 consonants, and if 20 participants produce them, this yields

4,480 tokens (9 x 7 x 20 + 23 x 7 x 20). This will take a professor who is not on sabbatical and who does not have graduate assistants to help him/her, two to three years to extract all necessary features if the person works on the data on weekends, some holidays, and parts of the summer. Once all the data have been extracted, they will need to be sorted by gender and also by correlates. Assuming that 20 speakers produced [ɥ], their individual production should be tabulated. The data should be tabulated by gender, then "first order statistical analysis," i.e., means and standard deviations should be calculated (Rabiner 1998:1267). This should be done for each segment!

Clearly, if this is the methodology, then there will not be enough human power to do feature extraction and analysis for the 2,000 or so indigenous languages spoken in Africa. Fortunately, there is a simpler solution which is less tedious but yields excellent results. Increasingly, experts are using **a single human exemplar** as the **Artificial Intelligent** (AI) agent for ASR and TTS applications. I have extracted vowel features from one such person whose voice is used in a pronunciation application that is used worldwide. Synthesizing the speech of a single individual saves time and efforts. There are just a few precautions to take to maximize intelligibility. The exemplar's voice should be as accent neutral as possible. In other words, speakers of the language should not be able to easily pinpoint the region where the speaker is from. The recordings should be of excellent quality. If at all possible, the human exemplar should be recorded in a studio. If this requirement is unduly burdensome, the recording should take place in a quiet room with noise cancellation equipment. The price of the latter is no longer exorbitant.

3.3 Critical Band Theory (CBT) and Just Noticeable Differences (JNDs)

Intelligibility is the ultimate goal of communication. For this reason, the interpretations of the measurements should be based on the **Critical Band Theory (CBT)**. CBT originated from the groundbreaking research done at Bells Research Laboratories from the 1920s to the 1960s. In the 1940s, physicist Harvey Fletcher pioneered a psychoacoustic methodology to gauge how the ear transduces acoustic signals into intelligible utterances. Another physicist, von Bekesy, demonstrated clinically that Fletcher's theory of Critical Bands was anchored in anatomical and auditory reality. For this, von Bekesy was awarded the Nobel Prize in Medicine/Physiology in 1961. Fletcher's and Bekesy's approach to intelligibility has revolutionized contemporary understanding of the processes involved in encoding and decoding speech signals. The third-octave response system, such as those listed in Rabiner and Juang (1993:186) and elsewhere, replicate as closely as possible how the human ear perceives speech signals (Everest and Pohlmann 2015:529). Zwicker (1961:248) and Pope (1998:1347) report that they have been endorsed by reputable bodies, such as the American National Standard Institute (ANSI), the International Standardization Organization (ISO), and the International Electrotechnical Commission (IEC)). CBT has uncovered important Just Noticeable Differences (JND) thresholds at which segments become intelligible or not. The main ones are summarized in Table 1:

N0	Acoustic Correlates	JND Thresholds
1.	F0	≤ 1 Hz
2.	F1	≤ 60 Hz
3.	F2	≤ 200
4.	F3	≤ 400
5.	Intensity	≤ 3 dB
6.	Duration	≤ 10 ms

Table 1: Intelligibility Thresholds

The symbol "≤" means that variations between segments of less than the indicated values are not perceptually salient. Stevens (2000:225) notes that for an item to qualify as a valid JND, it must elicit at least 75% of correct responses from a large pool of participants. Extensive discussions of CBT thresholds are available in Stevens (2000:203-241) and Rabiner and Juang (1993).

3.4 Phonetic Invariance

The above-mentioned JNDs and other like them have established beyond the shadow of a doubt that phonetic invariance is real. This does not mean that variability does not exist. Even ardent proponents of phonetic invariance do not deny that intraspeaker and interspeaker variability are a linguistic fact of life. Instead, they argue that phonetic variability is tightly regulated so as to assure intelligibility within the same speech community. Indeed, ASR and TTS systems have exploited the availability of JNDs to build robust systems that "understand" and are "understood by" an ever increasing number of speakers, even speakers of English with foreign accents.[2] Phonetic invariance is the main reason why digitalizing the speech of a single human exemplar is enough for building robust and "smart" ASR and TTS systems. Appendices 3 to 7 display the features extracted from the speech of one speaker of Anyi, an Akan language spoken in eastern Côte d'Ivoire.

4.0 Introducing the Arpabet

Broadly speaking, there are three transcription systems: the conventional orthography, the International Phonetic Alphabet (IPA), and the **Arpabet**. The first is used worldwide by all literate individuals. The second was officially created in 1888 (Pullum and Ladusaw 1986:xix) and is used primarily by linguists. Gambarage (2017:457) refers to the IPA as "the lingua franca for field linguists and phonologists/phoneticians." Since its creation predates modern computers, it is not compatible with them. For this reason, the Arpabet was introduced in the 1970s. It operates on the same principles as the IPA. Its main advantage is that it is fully compatible with ASCII symbols available on all computers. Jurafsky and Martin (2000:94-95) note that the Arpabet was initially designed for English but it is not English-centric. It can be expanded to transcribe African languages, as will be seen in the latter sections of this paper. The English Arpabet is used as the starting point of our discussions before the system is broadened to include African languages.

[2] All the smart systems on my computer or iPhone understand me even though I'm not a native speaker of English. I have noticed a very high level of intelligibility over the past 10 years since I have been testing voice-enabled devices.

N0	Arpabet Phoneme	IPA	Example	Arpabet Transcription
1.	AA	ɑ	odd	AA D
2.	AE	æ	at	AE T
3.	AH	ʌ	hut	hut
4.	AO	ɔ	ought	AO T
5.	AW	ɑʊ	cow	K AW
6.	AY	ɑɪ	hide	HH AY D
7.	B	b	be	B IY
8.	CH	ʧ	cheese	CH IY Z
9.	D	d	dee	D IY
10.	DH	ð	thee	DH IY
11.	EH	ɛ	Ed	EH D
12.	ER	ɚ	hurt	HH ER T
13.	EY	e	ate	EY T
14.	F	f	fee	F IY
15.	G	g	green	G R IY N
16.	HH	h	he	HH IY
17.	IH	ɪ	it	IH T
18.	IY	i	eat	IY T
19.	JH	dʒ	gee	JH IY
20.	K	k	key	K IY
21.	L	l	lee	L IY
22.	M	m	me	M IY
23.	N	n	knee	N IY
24.	NG	ŋ	ping	P IH NG
25.	OW	o	oat	OW T
26.	OY	ɔɪ	toy	T OY
27.	P	p	pee	P IY
28.	R	r	read	R IY D
29.	S	s	sea	S IY
30.	SH	ʃ	she	SH IY
31.	T	t	tea	T IY
32.	TH	θ	theta	TH EY T AH
33.	UH	ʊ	hood	HH UH D
34.	UW	u	two	T UW
35.	V	v	vee	V IY
36.	W	w	we	W IY
37.	Y	j	yield	Y IY L D
38.	Z	z	zee	Z IY
39.	ZH	ʒ	seizure	S IY ZH ER

Table 2: English Arpabet

A few noteworthy observations about Arpabet conventions are in order. First, all vowels are represented by two letters (digraphs). Secondly, the IPA symbols [θ, ð, dʒ, tʃ, ʒ] and aspirated [h] are also represented by diagraphs. Third, Arpabet transcriptions appear either in all in **capital**

(upper case) or lower case letters. However, capitalized transcriptions are more widespread. Fourth, the Arpabet system focused mostly on full-fledged phonemes. However, over the years, limited number of allophones have been included, as shown in Table 3:[3]

N0	Arpabet Phoneme	IPA	Example	Arpabet Transcription
1.	AX	ə	comma	K AA M AX
2.	EL	l̩	bottle	B AH Q EL
3.	EM	m̩	rhythm	R IY DH AX EM
4.	EN	n̩	butto**n**	B AH Q EN
5.	Q	ʔ	bu**tt**on	B AH Q EN

Table 3: Allophones of Arpabet

Fifth, the Arpabet also represents suprasegmentals by using Arabic numbers, as shown in Table 4:

N0	Arpabet Phoneme	IPA
1.	0	unmarked
2.	1	ˊ
3.	2	ˋ

Table 4: Suprasegmentals in Arpabet

The word <rhythm> can be transcribed in the Arpabet system as follows: R IY 1 DH AX EM 2. It is worth noting that stress indices appear at syllable boundaries. In reality though, only primary stress is indicated in most Arpabet systems. So, <rhythm> is transcribed as R IY1 DH AX EM 2. Dialectal variations can also be transcribed. In fact, many AI systems recommend that significant alternative pronunciations be transcribed to increase the "smartness" of the intelligent agent. All in all, the Arpabet transcription system is challenging for English because its orthography is **opaque**, which means that there is no one-to-one correspondence between spelling and pronunciation. Even so, current ASR and TTS systems work surprisingly well. This means that the Arpabet will work well for African languages because their orthographies are **transparent**. They are for the most part based on the **phonemic principle** which calls a straightforward one-to-one correspondence between phoneme and grapheme. Because of this, we are confident that once the relevant features have been extracted, speech synthesis based on the APA model can be implemented successfully in for African languages. Tone marking can be a challenge, but the solution proposed in 6.3 is supposed to work.

5.0 Extractable Consonant Features

Ordinarily, linguists rely on place of articulation, manner of articulation, and voicing to describe consonants exhaustively. However, for some African languages, one might need to add two additional features, namely **voice stream mechanism** and **double closure**. The former addresses issues having to with the pronunciation of implosives and ejectives, while the latter deals with labiovelars (to be explained below). For acoustic phonetic measurements, we will pay closer attention to manner of articulation because, as Reetz and Jongman (2009:199) explain, "It is easier to identify cues to manner of articulation and voicing than place of articulation." This explains why all the extractable features that Rabiner and Schafer (1978:43) display in Table 3.1 are manner

[3] The digraph [nx] is used to represent the allophone of /t/ when it is pronounced as [n] when it follows an [n] as in <twenty>, <winter>, <Hunter>, etc.

features. In other words, the extractable features used in speech synthesis are stops, fricatives, affricates, nasals, liquids, and glides/semi-vowels. Consonants are dealt with first in the next several sections because they are more numerous in any given language than vowels.

5.1 Extracting Stop Features

The most robust acoustic correlate that talkers and hearers rely on to encode and decode stops is voice onset time (VOT). It has to do with the amount of time that elapses when the articulators come together and when they part. Lisker and Abramson's (1964) article investigating the VOT of voiced and voiceless segments in 11 languages is by far the most influential acoustic phonetic study of its kind. Their methodology has been widely used to study VOT in many languages. Kent and Read (1992:120) contend that "VOT has been one of the most frequently measured phenomena in speech research." Ladefoged (2003:98) adds that "When making the description of a language, the VOT of stops consonants should always be given, as it varies considerably from one language to another." It needs to be pointed out that VOT can be positive or negative depending on the speaker and/or the language. In the former, the vocal folds begin vibrating even before the release of closure. In the latter, the vocal folds vibrate only after the release of closure. Katz (2013:252) has highlighted a correlation between VOT and degrees of voicing, namely, "If a language sets a voiced sound to be so negative in VOT, then the voiceless counterpart doesn't have to be strongly voiceless." Our investigation will go well beyond measuring the VOT of the plain voiceless stops [p, t, k] and the plain voiced stops [b, d, g] to include the VOT measurements of implosives, ejectives, labiovelars, and clicks that are found almost exclusively in African languages.

5.2 Focus on Implosives

The stop segments [ɓ, ɗ, ɠ] are called **implosives** because, in producing them, speakers suck air from outside into the oral cavity. Maddieson (1984:111-4) notes that about 10% of world the languages have implosives. Many of them are found in West African languages. We learn from Ladefoged and Maddieson (1996:87) that implosives are not very loud and are mostly voiced. They describe these sounds aerodynamically as follows, "The closed glottis is lowered so that the air pressure in the mouth decreases considerably. When it is about -4 cm H_2O, the vocal folds start vibrating and the oral pressure starts increasing. Shortly afterwards the lips come apart and air flows out of the mouth." VOT measurements of implosives are hard to come by. Ladefoged and Maddieson (1996:82-3) display a couple of spectrographic annotations of implosives.

5.3 Focus on Ejectives

Ejectives are segments that are produced forcefully. Unlike implosives, they are usually voiceless. The three ejectives commonly found in African languages are [p', t', k']. Ladefoged and Maddieson (1996:78) describe them aerodynamically as follows: "The pressure behind the closure in the oral cavity is often increased to about double the normal pulmonic pressure (i.e., about 16 cm H_2O). The oral closure is then released, and, owing to the greater supraglottal pressure, there is a greater amplitude in the burst." They occur in 16.40% of world languages, many of which are in West Africa (Maddieson 1984:101). Hausa is well known for its ejective stops [p', t', k']. We deduce from Ladefoged and Maddieson (1996:80, Figure 3.15), that the VOT of ejectives will be considerably long, ≥ 50 ms. Hausa in an interesting language in that it has plain stops, implosives, and ejectives. Studying its VOT will provide insights into how the speakers encode and decode subtle variations in stops.

5.4 Focus on Labiovelars

The feature **double closure** is used to describe the **labiovelars** [k͡p] and [g͡b]. These segments have two simultaneous closures, "the labial-velar closure does have a similarity to a velar one while its release has similarity to a labial one." (Ladefoged and Maddieson 1996:336). Thirty-three of the 61 (54%) West African languages in Ladefoged's (1968) survey have labiovelars. Various instrumental tools have been used and continue to be used to investigate their articulatory and aerodynamic characteristics. Yet, comprehensive VOT measurements of [k͡p] and [g͡b] that include F0, F1, F2, F3, VOT, and duration are hard to find. We note in passing Connell (1994) provide some measurements on five Nigerian languages, and De Jong (1997) also investigated the F2 of [g͡b] in relation to back vowels.

5.5 Focus on Clicks

Some of the rarest sounds in world languages are **clicks**. They are mostly found in southern African languages. As many as five clicks have been identified in Zulu: the bilabial [ʘ], the dental [ǀ], the alveolar [!], the lateral [ǁ], and the palatal [ǂ] (Ladefoged and Maddieson 1996:258). An important acoustic characteristic of clicks highlighted by Ladefoged and Maddieson (1996:259) is intensity. They note that, in general, their intensity is 6 dB greater than that of surrounding sounds.

5.6 Arpabet Notation of Stops in African Languages

All in all, 13 unique stops are found in African languages in addition to those commonly found in world languages. This calls for an expanded Arpabet system to accommodate: [pʼ, ɓ, tʼ, ɗ, kʼ, ɠ, k͡p, g͡b, ʘ, ǀ, !, ǁ, ǂ]. A two-letter Arpabet is suggested for these unique segments. The grapheme "H" is added to indicate ejectives, "Q" for implosives, and "K" for clicks.

Arpabet	IPA	F0	F1	F2	F3	VOT	Duration
P	p						
B	b						
T	t						
D	d						
K	k						
KP	kp						
G	g						
GB	gb						
PH	pʼ						
TH	tʼ						
KH	kʼ						
BQ	ɓ						
DQ	ɗ						
GQ	ɠ						
BK	ʘ						
FK	ǀ						
TK	!						
LK	ǁ						
ZK	ǂ						

Table 5: Expanded Arpabet for Stops

5.7 Extracting Fricatives and Affricates Features

From Clements (2000:125), Welmers (1973:50-56), and Ladefoged (1968), we know that the following fricatives [ɸ, β, f, v, s, z, ɕ, ʑ, z, ʃ, χ , ɣ h, ɦ] and affricates [tʃ, dʒ, tɕ, dʑ] occur in African languages. An important acoustic correlate of **fricatives** and **affricates** is center of gravity (COG) (Jongman and Wayland's 2000). Extracting this feature makes it possible to know as precisely as possible the articulatory characteristics of these segments. This is the reason why COG is listed in Table 6 as an additional extractable feature. Ladefoged and Maddieson (1996:139) lament the fact that there is a worldwide shortage of acoustic phonetic data on fricatives, "There have been surprisingly few studies of the acoustics of fricatives." Extracting many features helps to document the acoustic behavior of fricatives in African languages, thereby helping to address the shortage of data. In an expanded Arpabet, I suggest that fricatives and affricates that do not have counterparts in English be represented with "H" as the first digraph for glottal fricatives and F as the second diagraph for all other fricatives. "H" should be the first diagraph because H as the second diagraph is already used to represent <i>. The Arpabet annotation of fricatives that do not occur in English appear in the last 10 rows of Table 6.

Arpabet	IPA	F0	F1	F2	F3	Duration	CoG
F	f						
V	v						
S	s						
Z	z						
SH	ʃ						
CH	tʃ						
JH	dʒ						
HY	ç						
HH	h						
HJ	ɥ						
FF	ɸ						
BF	β						
CF	ɕ						
ZF	ʑ						
XF	χ						
GF	ɣ						
XF	ɦ						

Table 6: Expanded Arpabet for Fricatives and Affricates

5.8 Extracting Nasal Features

The segments [m, n, ɲ, ŋ] are the most commonly found nasals in world languages. Ladefoged's (1968:45-63) survey shows that 47 of the 61 West African languages (77.04%) have all four **nasals**. Ladefoged and Maddieson (1996:117) note that "There have been relatively few studies of the acoustic distinctions between nasals in natural languages, and many of those that do exist are limited to **m** and **n**." Extracting five features from nasals will document less studied nasals.

Arpabet	IPA	F0	F1	F2	F3	Duration
M	m					
N	n					
NG	ŋ					
NY	ɲ					

Table 7: Expanded Arpabet for Nasals

Nasals are particularly difficult to synthesize because their pronunciation involves two cavities: the oral and the nasopharyngeal cavities. In producing nasal sounds, a more or less significant portion of air molecules are diverted into the sinuses when the velum lowers. This causes the areas above F2 to be less intense on spectrographs. This area is known as **zeros**. The difficulties in synthesizing nasals and nasalized segments are amply discussed by Rabiner and Schafer (1978:450).

Additional difficulties surface in extracting features from prenasalized segments, that is, nasal sounds that occur immediately before an obstruent. These clusters are pervasive in African languages but relatively uncommon in world languages. Maddieson (1984:67) found that only 19 languages out of the 317 in the UCLA Phonological Segment Inventory Database (UPSID) have prenasalized obstruents. Welmers (1973:69-72) opines that prenasalized segments may have started as prefixes in an earlier stage of Niger-Congo languages. The IPA transcription of prenasalized consonants is controversial. When a prenasalized segment occurs before [f], [v], [k͡p], and [g͡b], some transcribe the sequence as [ŋf], [ŋv] while others write them as [nf] and [nv]. Controversial also is the way in prenasalized [k͡p], and [g͡b] are transcribed. Welmers (1973:65)[4] offers the following opinion, "I have personally preferred /ŋkp, ŋgb/ in the cases I have met, but again no great theoretical issue is at state." Yet others prefer transcribing them as [ŋk͡p] and [ŋg͡b]. Ladefoged and Maddieson (1996:334, 6) indicate that spectrographic evidence from Efik and Logbara are inconclusive as to the preferred method of transcription. Martinez and Rosenbaum (2017) offer acoustic phonetic and aerodynamic descriptions of prenasalized segments in Somali Chizigula which underscore the articulatory complexity of these segments.

The orthographic representation of prenasalized consonants that occur in word-medial positions are no less challenging because such clusters raise theoretical questions about syllable structure. Koffi (2009:92-102, 112-114) devotes 15 pages sorting out various orthographic options. The same issues surface in the Arpabet transcription of word-medial prenasalized segments. For example, in the word [kĩ$_{\sigma}$ndɛ] (to look for), there are two possible options, both of which have theoretical undertones.[5] If the vowel [ɪ] of the first syllable is nasalized, one would have to assume that the underlying phonemic representation is /kɪn$_{\sigma}$ndɛ/. In this case, the nasal tilde on [ĩ] is the result of a tautosyllabic nasalization rule. The other option would posit that the underlying phonemic representation is /kɪ$_{\sigma}$ndɛ/. Therefore, the surface form [kĩndɛ] is the result of a nasalization rule that crosses syllable boundaries. The tendency in the phonological literature is to assume that nasalization rules are tautosyllabic. If this assumption is accepted, the orthographic form of [kĩ$_{\sigma}$ndɛ] should be <kɪnndɛ>, not <kɪndɛ>. Thus, the Arpabet transcription should be K IH N ND EH, not K IH ND EH. In other words, one should expect the digitalization

[4] Transcriptions such as [mŋkp, mŋgb] and [ŋ͡mk͡p] or [ŋ͡mg͡b] that appear in Welmers (1973:65) or Clements (2000:129) are hard to justify acoustically.
[5] Here, the symbol "□" is used as a marker of syllable boundary.

of prenasalized segments to present formidable challenges in speech digitalization, as they are for phonological theory.

5.9 Extracting Liquid Features

Most African languages have the **liquids** [l] and [r] (Ladefoged 1968). However, in most cases, /l/ is the basic phoneme and [r] is an allophone. In Anyi, for example, /l/ is the basic phoneme, and [r] is an allophone that occurs only immediately after coronals. In some languages, [l] and [r] occur in free variation. Regardless of their phonological status, [l] and [r] features need to be extracted separately because they have salient acoustic characteristics that make them perceptually different. F3 is an important feature to extract because, according to O'Connor et al.'s (1976:301, 306), "The distinction between /l/ and /r/ seems to depend primarily on the third-formant transition." A comprehensive study done by Epsy-Wilson (1992) support the view that a liquid is perceived as a **lateral** if its JND is ≥ 2,600 Hz, unless it is trilled. Laterals include clear [l]s, syllabic [l̩]s, or approximant [l̠]. Segments whose F3 are below ≤ 2,200 Hz, are taken to be rhotics such as [r], [ɾ], or the approximant rhotic [ɻ] . Segments whose F3 fall between 2,500 Hz and 2,200 Hz are hard to perceive clearly. Such is the case of the pronunciation of /l/ in /bala/ (woman) in Anyi Morofu which produced in the Anyi Bona dialect of either as [ba l̩a] or [ba ɻa]. In such cases, a vibration calculation may help disentangle the perceptual difficulties. Ladefoged (2003:151) provides the following formula for calculating the degree of **trilling**.

$$\text{Vibration in Hz:} \quad \frac{\text{Absolute Duration in Milliseconds}}{\text{Relative Duration of Segment}}$$

The numerator is the absolute duration in milliseconds. It is always 1000 because 1 second equals 1000 milliseconds. The denominator is the duration of the actual segment under consideration. The JND for trilling is ≥ 22 Hz. Any [r] whose value below this threshold is considered flapped, tapped, or not trilled. If the vibration rate of [r] exceeds this JND, it must be accounted for in the Arpabet system by transcribing it as [RR]. Trilling effectively differentiates between laterals and rhotics because no human language has a trilled lateral. As soon as trilling is heard, whether it is forceful or faint, it should be transcribed in Arpabet as R or RR.

Arpabet	IPA	F0	F1	F2	F3	Vibration
L	l					
R	r					
RR	R					

Table 8: Expanded Arpabet for Liquids

5.10 Extracting Glide Features

Ladefoged (1968) lists [w], [j], and [ɥ] as the **glides** commonly found in West African languages. Espy-Wilson (1997) and others have noted that F2 is the most robust correlate for discriminating between [j] and [w].

Arpabet	IPA	F0	F1	F2	F3	Duration
W	w					
Y	j					
HJ	ɥ					

Table 9: Expanded Arpabet for Glides

The phonological literature on African languages reports pervasive co-articulated when the glides [w] and [j] occur immediately after [p], [b], [t], [d], [k],[g], [f], [v], [s], [z], [tʃ], and [dʒ]. When [j] occurs with [p], phonologists refer to it as **palatalization**. It is transcribed in the IPA system as [pʲ], [bʲ], [tʲ], [dʲ], [kʲ], [gʲ], [fʲ], [sʲ], [zʲ], [tʃʲ], [dʒʲ]. When [w] follows the same segments, it is known as **labialization** and transcribed as [pʷ], [bʷ], [tʷ], [dʷ], [kʷ],[gʷ], [fʷ], [vʷ], [sʷ], [zʷ], [tʃʷ], [dʒʷ]. F3 can help gauge the degree of palatalization and labialization. The same JND of ≥ 2,600 Hz used to discriminate between [l] and [r] also applies here. Palatalized segments call for lip protrusion and therefore have F3 values that exceed ≥ 2,600 Hz. Consequently, its F3 values are expected to be < 2,600 Hz. Labialized segments have F3 values similar to [m], [r], and [w] because lip rounding entails the lowering of the velum. If palatalization and labialization are deemed salient enough to be worth represented in the Arpabet system, then palatalized segments could be represented as **segment + j**, and labialized segments as **segment + W**. Caution should be exercised before implementation so as not to introduce three-segment symbols into the Arpabet transcription system.

6.0 Extracting and Measuring Vowel Features

Vowels are fewer than consonants in all languages. Yet, they play as great a role, and sometimes a greater role in intelligibility than consonants. Prator and Robinett (1985:13) overemphasize their role in teaching English as a second language, arguing that when vowels are poorly pronounced, intelligibility is severely compromised. Rabiner and Juang (1993:21-3) express a similar view regarding the role of vowels in ASR and TTS systems:

> The vowel sounds are perhaps the most interesting class of sounds in English. Their importance to the classification and representation of written text is very low; **however, most practical speech-recognition systems rely heavily on vowel recognition to achieve high performance**.[6] … The vowel sound produced is determined primarily by the position of the tongue, but the position of the jaw, lips, and to a small extent, the velum, also influences the resulting sound.

According to Welmers (1973:20-45), vowel systems in African languages vary in size from five to ten vowels. However, seven /i, u, e, ɛ, o, ɔ, a/ or nine /i, ɪ, u, ʊ, e, ɛ, o, ɔ, a/ vowel systems are the most common. In such systems, /a/ is the only central vowels. There are a few exceptions. Welmers (1973:20) lists some languages as having [ɨ, ə, ɨ̈]. Marchese (1989:128) notes that Bete and Godie, and possibly other Eastern Kru languages have [ʉ] and [ʌ]. I should hasten to add that both Welmers' and Marchese's descriptions of these vowels are based on impressionistic data analysis. If indeed [ɨ, ɨ̈, ʉ, ə, ʌ, a] occur as central vowels, they should be presented in the Arpabet, as shown in Table 10. I make the following suggestions for represented central vowels. Except for [a], [ʌ], and [ə], for which Arpabet symbols exist already, I suggest adding "C" to represent

[6] Highlighting not in the original quote.

the central vowels for which no Arpabet is available. Thus, [ɨ] is transcribed as IC, [ʉ] as UC, and [ɪ̈] as [YC].

Arpabet	IPA	F0	F1	F2	F3	Duration
IY	i̇					
IH	ɪ					
IC	ɨ					
YC	ɪ̈					
EY	e					
EH	ɛ					
AE	æ					
AA	a					
AO	ɔ					
OW	o					
UW	u					
UH	ʊ					
UC	ʉ					
AH	ʌ					
AX	ə					

Table 10: Arpabet for Vowels

6.1 Vowel Length Transcription in the Arpabet System

Very little research has been done on vowel length in African languages. The silence may be due to the fact that vowel length and tonal contour go hand in hand. Ladefoged (1968:33) explains the complexities of correlation between vowel length and tone as follows:

> Discussion of vowel length is always complicated by the interaction of the phonological analysis of length and tone. … In general, it would seem that when, as in many Kwa languages, perceptually long vowels can be on one pitch or involve a change in pitch, and when these vowels can occur in the same phonological structures as sequences of different vowels, then it is preferable to regard them as two vowels.

Let's set tone aside for the moment and focus purely on vowel quality. Here is an example from Anyi for dealing with vowel length by itself. The language makes a three-way contrast between [bó] (to break), [bô] (nose), and [bǒ] (forest). Impressionistically and instrumentally, there is a durational difference between the [o] in the three vowels. The [o] of [bó] is differentiated from [bô] (nose) by being slightly longer. Thus, in an Arpabet system, the two can be transcribed respectively as follows: [B OW] and [B OW OW]. Duration is indicated in Arpabet by doubling the vowel. The [bǒ] (forest) is also slightly longer than [bô] (nose). How then, should it be represented in the Arpabet system? Transcribing it as [B OW OW] is not enough because this transcription does not differentiate it from [bô] (nose), which is also transcribed as [B OW OW]. What can be done in such cases? A solution is proposed in 6.3 that takes into account information about vowel duration and tone contrasts, as discussed in 6.2 below.

6.2 Tone Transcription in the Arpabet System

Before answering the question as to how [bŏ] (forest) and [bô] (nose) are to be differentiated in the Arpabet system, a little detour is necessary in order to discuss auditory illusion and the Critical Band Theory (CBT). Katz (2013:179) notes that auditory illusions are as common as optical illusions. Baken and Orlikoff (2000:1) state that "The ear is too easily fooled." These realities caution us not to rely solely on impressionistic methodologies to make statements about tonal structures in African languages. Koffi (2017) has proposed an acoustic phonetic analysis of tone that is firmly anchored in the Critical Band Theory (CBT). This approach takes the view that the 1/3-octave response system used in audio engineering is "close to the critical bandwidth of the ear" (Pohlmann 2015: 36, 529). *Handbook of the IPA* (1999:14) reports that phoneticians all over the world accept the view that there are only five pitch registers in all human languages: extra low, low, mid, high, and extra high. Furthermore, it is accepted as uncontroversial that the pitch that the human vocal apparatus can produce "extends from about 60 Hz to about 500 Hz" (Fry 1979:68). However, the default settings in Praat go from 75 Hz to 500 Hz because nobody speaks with an F0 lower than 75 Hz. The maximum pitch level is set at 500 Hz because, except for colicky babies, nobody produces F0 ≥ 500 Hz. Putting all these acoustic phonetic facts together, Koffi (2017) has proposed the thresholds in Table 11 in which F0 measurements correlate with pitch registers systematically:

N0	Tone Registers	Lower Limits	Center Frequency	Upper Limits	Arpabet
1.	Extra low	71	80	88	0
2.	Low	89	100	113	1
3.	Mid	114	125	141	2
4.	High	142	160	176	3
5.	Extra high	177	200	225	4

Table 11: Critical Bands for Men and Arpabet Notation

Except for Anyi data that I collected and analyzed myself, I have not come across outside data to verify the postulates in Table 11. F0 measurements on African languages are extremely hard to come by because, as noted previously, most of the statements about tone and tone patterns are based solely on the impressionistic evaluation of the linguist doing the research. I was therefore elated to have run into Professor Bearth in Geneva in December 2019. He informed me that he had done an instrumental study of Toura tones and willingly sent me a copy of the paper he published in 1968, that is, almost 49 years before I published my paper demonstrating that the JNDs in Table 11 correlate F0 measurements with tone registers accurately. His instrumental and statistical analyses and my CBT-based correlations are astoundingly identical. Below are the correlations between F0s and tone registers proposed by Bearth (1968:47) for Toura:

N0	Tone Registers	Center Frequency
1.	Low	110 Hz
2.	Mid	120 Hz
3.	Mid-high	140 Hz
4.	High	160 Hz

Table 12: Center Frequency of Toura Level Tones

Now, when one is dealing with female speech, the measurements in Table 11 (and Table 12, too) should be raised by 50%, or multiplied by 1.5 (Kent and Read 2002:191). Table 13 correlates

female F0 with tone registers:

N0	Tone Registers	Lower Limits	Center Frequency	Upper Limits	Arpabet
1.	Extra low	106	120	132	0
2.	Low	133	150	169	1
3.	Mid	170	185	211	2
4.	High	212	240	264	3
5.	Extra high	265	300	337	4

Table 13: Critical Bands for Women and Arpabet Notation

6.3 Transcribing Vowel Length and Tonal Patters in the Arpabet System

The information in the two previous sections makes it possible to provide an Arpabet transcription that combines both vowel duration and tone contrasts. The tone bearing unit (TBU) of the word [bó] (to break) is high tone. In the suprasegmental annotation in Arpabet, this corresponds to the numerical index of "3." Consequently, [bó] is transcribed as a [B OW3]. The TBU of [bô] (nose) has a high-low contour tone and corresponds in the Arpabet system to "3" and "1." Therefore, [bô] is transcribed as [B OW3 OW1]. Similarly, the TBU of [bǒ] (forest) is low-high and corresponds to "1" and "3." The Arpabet transcription of [bǒ] is [B OW1 OW3]. The Arpabet system can accommodate segmental duration correlated with any combination of contour tone patterns.

7.0 Beyond Acoustic Phonetic Feature Extraction

The ultimate goal of feature extraction is to get to a point where indigenous languages can be fully digitalized and used in speech synthesis for ASR and TTS applications. As this tutorial has shown, the process is long and tedious. Even though various scripts are available to automatize various aspects of the task, to the best of my knowledge, there is not a single script that can extract five or more correlates at once. Furthermore, many phoneticians are leery about the accuracy of scripts. Because of misgivings about scripts, one is better off annotating and extracting all the information oneself, even though doing so is laborious and even if speech synthesis is based off on one prototypical speaker. When all the features have been extracted, the first step is complete. However, the overall goal of ASR and TTS is far from being achieved. Figure 1, which displays the overall architecture, shows that two more blocks remain. The third block is "Segmentation and Labelling." It involves several subsidiary steps, such as phoneme lattice, segment lattice, probabilistic labelling, decision trees, and parsing strategies. The "Control Strategy" block also remains to be implemented. In light of all that remains, the following statement by Kent and Read (2003:243) appears to be too optimistic:

> The remaining of this chapter describes different types of speech synthesis. Most of these are based on acoustic models of the speech signal and are most commonly used today. Because many of these have been implemented on ordinary microcomputers, anyone with a personal computer and some ancillary equipment can experiment with speech synthesis. This fact has accelerated progress in the field.

This optimism should be tampered with a dose of reality. It is uncontroversial that personal computers abound, but what are "some ancillary equipment" that are needed? Almost all the speech synthesizer currently in use call for a large amount of speech data for "deep" learning neural networks (Haley et al. 2020:2-6). The data needed is astronomical, amounting dozens of hours,

and in some cases, hundreds of hours of good quality audio recordings. What we badly need for the indigenous languages in Africa is an acoustic phonetic synthesizer that can produce authentic speech quality based on a comparatively small dataset, i.e., 500 to 1,000 words. Moreover, in my experience and practice, a successful ASR or TTS systems cannot be designed by a single person. It calls for a collaboration with (signal processing) engineers and computer scientists with expertise in artificial intelligence and/or natural language processing. I'm fortunate to work with such a team of experts at my university. We have embarked on a speech synthesis project that obviates the need for "deep" learning.

8.0 Summary

The steps outlined in this tutorial opens up possibilities that were unimaginable just a few years ago. With the ongoing revolution in speech technologies and speech enable-AI systems, developing, endangered, and moribund indigenous languages of Africa and elsewhere have a new lease on life. Whatever status a language finds itself in, feature extraction can help modernize, revitalize, and preserve it for current and future generations. Armed with a vocabulary of only 500 words, an AI agent endowed with an adequate phonological, morphological, and syntactic description and rules can generate novel utterances and understand new ones as it encounters them. There is hope that, not too far from now, moribund and even "dead" languages can be revitalized again if their features are extracted. With the proliferation of mobile devices that are becoming increasingly smarter and smarter, speech-enabled AI agents can be designed for indigenous languages that will teach indigenous peoples their own native tongues. Speech synthesis will pay huge dividends for linguists, language activists, speakers of indigenous languages, and policymakers as they endear to preserve linguistic diversity.

ABOUT THE AUTHOR

Ettien Koffi, Ph.D. linguistics, teaches at Saint Cloud State University, MN. He is the author of four 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 perspectives 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) and Text-to-Speech (TTS), and voice biometrics for speaker verification. He can be reached at enkoffi@stcloudstate.edu.

References

Bearth, Thomas. 1968. Etude Instrumentale des Tons du Toura (Côte d'Ivoire). *Cahiers Ferdinand de Saussure* 24: 45-58.

Bekesy, George V. 1947. The Variation of Phase along the Basilar Membrane with Sinusoidal Vibrations. *Journal of the Acoustical Society of America* 19 (3): 452-460.

Clements, Nick G. 2000. Phonology. In Bernd Heine and Derek Nurse (eds), *African Languages: An Introduction*. New York: Cambridge University Press, 123-160

Connell, Bruce. 1994. The Structure of Labio-velar Stops. *Journal of Phonetics* 22:441-476

de Jong, Kenneth. 1997. Labiovelar Compensation in Back Vowels. *Journal of the Acoustical Society of America* 101 (4): 2221-2232.

Fletcher, Harvey. 1940. Auditory Patterns. *Reviews of Modern Physics* 12: 47-65.

Fry, Dennis B. 1976. *Acoustic Phonetics: A Course of Basic Reading.* New York: Cambridge University Press.

Eckert, Penelope. 2008. Where do Ethnolects Stop? *International Journal of Bilingual*ism 12, Numbers 1 &2: 25-42.

Everest, Alton F. and Ken C. Pohlmann. 2015. *Master Handbook of Acoustics*. Fifth Edition. New York: McGraw Hill Education.

Hawley, Scott H., Vasileios Chatziiannou, and Andrew Morrison. 2020. Synthesis of Musical Instrument Sounds: Physics-Based Modeling or Machine Learning. *Physics Today* 16 (1):20-28.

Heine, Bernd and Derek Nurse. 2000. Introduction. In Bernd Heine and Derek Nurse (eds), *African Languages: An Introduction*. New York: Cambridge University Press, 1-10

Hirsh, Ira J. 1959. Auditory Perception of Temporal Order. *Journal of the Acoustical Society of America* 31 (6): 759-767.

Fant, Gunnar 1998. Acoustical Analysis of Speech. In Malcolm J. Crocker (ed.) *Handbook of Acoustics*. New York: John Wiley & Sons, Inc., 1245-1270.

Gambarage, Joash J. 2017. Unmasking the Bantu Orthographic Vowels: The Challenge for Language Documentation and Description. *Africa's Endangered Languages: Documentary and Theoretical Approaches*, ed. by Jason Kandybowicz and Harold Torrence, pp. 449-484. New York: Oxford University Press.

Goldsmith, John. 1990. *Autosegmental and Metrical Phonology*. Cambridge, MA: Basil Blackwell.

Jongman, Allard and Ratree Wayland. 2009. *Phonetics: Transcription, Production, Acoustics, and Perception*. Malden, MA: Wiley-Blackwell.

Jongman, Allard and Ratree Wayland. 2000. Acoustic Characteristics of English Fricatives. *Journal of the Acoustical Society of America* 108 (3):1252-1263.

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

Koffi, Ettien. 2018. The Acoustic Phonetic Measurements of Three Tonally Contrastive Grammatical Moods in Anyi. Abstract in *The Journal of the Acoustical Society of America* 143 (2): 290. ASA Meeting in Minneapolis, MN, 7-11 May.

Koffi, Ettien. 2018. Differential Analysis of Lexical Pitch in Accent and Tone Languages. *Linguistic Portfolios* 7:110-131.

Koffi, Ettien. 2017. The Acoustic Vowel Space of Anyi in Light of the Cardinal Vowel System and Dispersion Focalization Theory. In Jason Kandybowics, Travis Major, and Harold Torrence (eds), *African Linguistics on the Prairie.* Berlin: Language Science Press, 3-17.

Koffi, Ettien. 2017. The New Paradigm in Tone Analysis: The Contribution of the Critical Band Theory. *Linguistic Portfolios* 6:147-164.

Koffi, Ettien. 2016. The Lowering of Lax Vowels in Central Minnesota English: Does it Happen in Other Dialects? *Linguistic Portfolios* 5:2-14.

Koffi, Ettien. 2009. *The Interface between Phonology, Morpho(phono)logy in the Standardization of Anyi Orthography*. Bloomington, IN: Revised Unpublished Ph.D. Dissertation.

Labov, William, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis. *Languages* 89: 30-65.

Labov, William, Sharon Ash, and Charles Boberg. 2006. *Atlas of North American English: Phonology and Sound Change*. Berlin: Mouton de Gruyter.

Ladefoged, Peter. 2001. *A Course in Phonetics*. Fourth Edition. New York: Harcourt College Publishing.

Ladefoged, Peter. 1999. American English. In *International Phonetic Association* (ed), *Handbook on the International Phonetic Alphabet.* New York: Cambridge University Press, 41-44.

Ladefoged, Peter, Iris Kemeny, and William Brackenridge. 1976. Acoustic Effects of Style of Speech. *Journal of the Acoustical Society of America* 59 (1):228-231.

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

Ladefoged, Peter and Ian Maddieson. 2015. *A Course in Phonetics. 7th edition.* Cengage Learning: Stamford, CT.

Ladefoged, Peter and Ian Maddieson. 1996. *The Sounds of the World Languages*. Malden, MA: Blackwell Publishers Inc.

Lewis, Paul M. 2009. *Ethnologue: Languages of the World.* Sixteenth edition. Dallas: SIL International.

Lisker, Leigh and Arthur S. Abramson. 1964. A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *Word* 20 (3): 383-422.

Maddieson, Ian and Peter Ladefoged. 1989. Multiply Articulated Segments and the Feature Hierarchy. *UCLA Working Papers on Phonetics* 72:116-138.

Marchese, Lynell. 1989. Kru. *The Niger-Congo Languages*. Ed by John Bendor-Samuel, pp. 119-139. New York: University Press of America and SIL.

Martinez, Michal T. and Vanessa Rosenbaum. 2017. Acoustic and Aerodynamic Data on Somali Chizigula Stops. *Africa's Endangered Languages: Documentary and Theoretical Approaches*, ed. by Jason Kandybowicz and Harold Torrence, pp. 427-447. New York: Oxford University Press.

Mermelstein, Paul. 1978. Difference Limens for Formant Frequencies of Steady-state and Consonant-bound Vowels. *Journal of the Acoustical Society of America* 63 (2): 572-580.

Narayan, Chandan R. 2008. The Acoustic-Perceptual Salience of Nasal Place Contrasts. *Journal of Phonetics* 36:191-217.

O'Connor, J.D., L.J. Gerstman, A.M. Liberman, and F.S. Cooper. 1976. Acoustic Cues for the Perception of Initial /w, j, r, l/. In Dennis Fry (ed), *Acoustic Phonetics: A Course of Basic Readings*. New York: Cambridge University Press, 298-314.

Pope, J. 1998. Analyzers. In Malcom J. Crocker (ed), *Handbook of Acoustics*. Wiley-Interscience Publication: New York,1341-1353.

Pullum, Geoffrey K. and William A. Ladusaw.1986. *Phonetic Symbol Guide*. Chicago: The University of Chicago Press.

Rabiner, Lawrence R. 1998. Machine Recognition of Speech. In Malcolm J. Crocker (ed.) *Handbook of Acoustics*. New York: John Wiley & Sons, Inc., 1263-1270.

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

Rabiner, Lawrence and R. W. Schafer. 1978. *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey: Prentice-Hall. Inc.
Speaks, Charles E. 2005. *Introduction to Sound: Acoustics for the Hearing and Speech Sciences.* Third Edition. Clifton Park, NY: Thomson Delmar Learning.
Stevens, Kenneth N. 2000. *Acoustic Phonetics*. Cambridge, MA: The MIT Press.
International Phonetic Association. 1999. *Handbook of the IPA: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press: New York, NY.
Welmers, William E. 1973. African Language Structures. Berkeley, CA: University of California Press.
Zwicker, E. 1961. Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). *The Journal of the Acoustical Society of America* 33 (2): 248.

Appendices

MANNER OF ARTICULTION	PLACE OF ARTICULATION						
	Labial	Labio-dental	Dental-alveolar	Palatal	Velar	Labio-velar	Glottal
Voiceless Stop	p		t	tʃ	k	kp	
Voiced Stops	b		d	dʒ	g	gb	
Voiceless Fricatives		f	s				h
Voiced Fricatives		(v)	(z)				
Nasals	m		n	ɲ	ŋ		
Liquids			l (r)				
Semi-Vowels				J (ɥ)	w		

Appendix 1: Anyi Consonant Phonemes (Koffi 2009)

Note: The segments between parentheses, (v) and (z) are morphophonological variants of /f/ and /s/ respectively. The segment (r) is an allophone of /l/ when it occurs after coronals.

Note on Consonant Measurements

Unless otherwise noted, the measurements of consonants are based on the speech of a single speaker who produced these segments in running speech.

Arpabet	IPA	F0	F1	F2	F3	VOT[7]	Duration
P	p	60[8]	763	1725	3159	15	20
B	b	130	483	1164	2663	-8	47
T	t	60	711	2062	3186	0	23
D	d	133	332	1861	3019	-43	45
K	k	60	1081	1958	3250	7	73
KP	kp	114	523	1235	2738	-24	46
G	g	60	1075	2062	2750	-45	40
GB	gb	139	445	1396	2927	-37	37

Appendix 2: Stop Measurements in Anyi

Arpabet	IPA	F0	F1	F2	F3	CoG	Duration
F	f	60	1012	2015	3036	6968	82
V	v	122	652	1982	3016	1338	57
S	s	60	1180	2552	3242	5660	137
Z	z	127	581	1626	3065	5050	115
CH	tʃ	60	974	2466	3278	6045	44
JH	dʒ	175	535	2313	3419	753	97
HH	h	60	958	1906	3220	1213	98

Appendix 3: Fricative and Affricate Measurements in Anyi

Arpabet	IPA	F0	F1	F2	F3	Duration
M	m	130	972	1834	2975	93
N	n	135	445	1670	3009	136
NG	ŋ	169	509	1627	2586	117
NY	ɲ	131	748	1800	3236	109

Appendix 4: Nasal Measurements in Anyi

Arpabet	IPA	F0	F1	F2	F3	Vibration	Duration
L	l	138	650	1501	2548	16	64
R	r	135	660	1531	2576	26	38

Appendix 5: Liquid Measurements in Anyi

[7] The VOT measurements must be taken with a grain of salt. VOT in running speech may not work well for languages with open syllables such as Anyi because the voicing of the vowel that immediately precedes the stop consonant bleeds into it. This is particularly tricky if the stop itself is a voiced segment. I did not know this when I started!

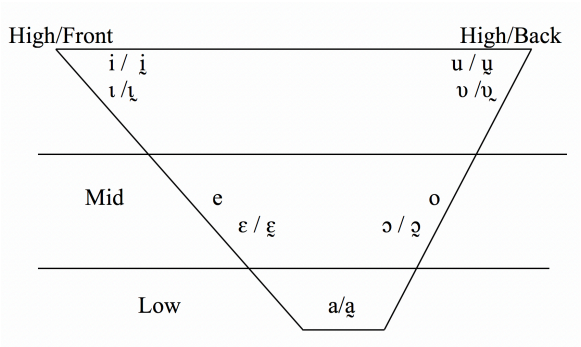
[8] Fry (1978:68) notes that 60 Hz is the lowest possible pitch that the human voice can produce. However, the default pitch setting in Praat is ≥ 75 Hz. Any time Praat F0 measurement as "undefined," I select the default value of 60 Hz for that segment.

Arpabet	IPA	F0	F1	F2	F3	CoG	Duration
W	w	139	485	2075	3455	NA	52
Y	j	124	294	1996	3245	NA	46
HJ	ɥ	60	718	2151	2817	1997	130

Appendix 6: Glide Measurements in Anyi

Note on Vowel Measurements

The measurements in Appendix 7 are based on data published by Koffi (2017). The measurements are based on data collected from 10 speakers. The vowels containing the vowels were produced three times each in citation form.

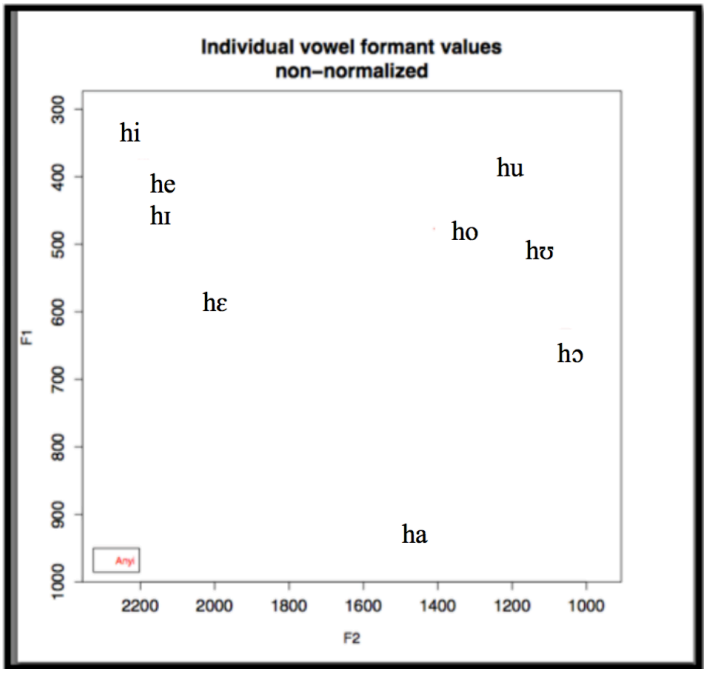


Appendix 7: Anyi Vowel Phonemes (Koffi 2009)

Note: There are some important differences between a prototypical vowel quadrant such as the one above and an actual acoustic such as the one in Appendix 9 based on F1 and F2 measurements obtained by 10 male speakers.

Arpabet	IPA	F0	F1	F2	F3	Duration
IY	i	141	348	2206	3174	118
IH	ɪ	142	399	2174	3055	113
EY	e	146	392	2141	2877	121
EH	ɛ	140	589	2038	2738	114
AA	a	137	925	1486	2506	102
AO	ɔ	139	635	1056	2545	109
OW	o	152	477	1392	2897	121
UW	u	144	388	1249	2832	117
UH	ʊ	147	523	1182	2716	116

Appendix 8: Vowel Measurements in Anyi



Appendix 9: Acoustic Vowel Space of Anyi

Appendix 9

Seventeen Anyi popular proverbs used in various data collections.

1. Sɛ ɛsʊn kʊla wɔ, ɔ di wɔ boó.
2. Sanran b'ɔ si ɛsʊn sʊ, nyanzuo ngan man yi.
3. Kannzɛ ɛsʊn tɪ kpili bɔbɔ ɔ, ɔ nɪn asɪɛ nzɛ man.
4. Anʊnman nvɛ man ɛyaá, bakaá wʊn.
5. Bɛlɛbɛlɛ, yɪɛ anʊnman fá yɔ yɪ suá ɔ.
6. Akɔ ja ngu man yɪ wáa.
7. Sɛ akɔ nyan nɪnka fɪa ɔ, yɪɛ ɔ bisa cɔmaán kosan ɔ.
8. Ekpóo kpa yɪ pieto yuo ɔ, yɪ wʊla nvi man yɪ dua nɪnka.
9. Dodohɔlɔ wan, "Ndɛndɛ tɪ yie, bɛlɛbɛlɛ tɪ yie."
10. Jɪjɪlɪwáa bakaá: bie mɔ lɛ fʊ sʊ ɔ, nɪ́n bie mɔ lɛ jura ba.
11. Bakaá tɔ nzuo nun ɔ, ɔ ngaci man ɛlɛngɛ.
12. Ketebʊɔ nnyʊn kan bo nun ɔ, bɛ ndara man jɪra.
13. Kpɛnzɛ wan, "alɪé atíɪn nnʊn man mmʊ́a.
14. Sanran b'ɔ tin dufalɛ, ɔ ŋjici man yɪ sa nʊan.
15. Wʊntɪnwʊntɪn wan, ɔ́ɔ kɔ Kelegbe, naán kɔ́ nían kɛ, awʊnman a fita yi.
16. Bɛ́ɛ tu nanmuo nun ɔ, bɛ́ɛ nje man nun.
17. Bɛ fulalɪ kʊn naán bɛ finlɪ man yi asɪ ɔ, bɛ nzea kɛ, b'ɔ tɪ ɛlɔ a je.