Principal Components Visualisation of Acoustic-Emotion Profiles in Ibibio

Eno-Abasi & Moses Ekpenyong

Abstract
In this contribution, a principal component analysis (PCA) technique for visualizing the effect of acoustic features on different emotion profiles is proposed. To accomplish this, emotions speech corpus were handcrafted, resulting in seven emotions (anger, fear, joy, normal, pride, sadness, surprise), and recorded under suboptimal conditions. Acoustic features including duration, pitch/F0, intensity and the first four formants (F1-F4) were extracted from the sentence, word and syllable units – for the study – using Praat scripting and component-wise normalisation. The normalised features were then subjected to an unsupervised feature selection and dimension reduction process using PCA. A subsequent visualisation of the principal component dominant features (PCDFs) enabled a proper investigation of acoustic-emotion variability profiles for male and female speakers. The results revealed that speech formants, a direct correlate of tone, constitute the most PCDFs and is important for investigating acoustic-emotion variability profiles in African tone languages (ATLs).Efforts to develop emotion databases for Ibibio emotion recognition systems are ongoing, and a comprehensive statistical evaluation is expected in the future.

Keywords: acoustic-emotion profile; emotion recognition; feature component visualisation; PCA; African tone language.

1. Introduction

In spite of the potential benefits of emotion recognition, the problem still remains open (Taylor, Scherer and Cowie, 2005), given the difficulty of emotion recognition systems to appropriately identify a feature space for classification purposes. Research works on speech and emotion have since shifted from exploratory to the production of (some) substantial evidence – mainly in the field of Human Computer Interaction (HCI) – where progress in this area relies specifically on the development of appropriate emotional databases (Douglas-Cowie, Campbell, Cowie and Roach, 2003). While there exist substantial research evidence on emotion classification, constructing a universally applicable classifier remains unsolved and daunting – largely due to context-dependency and variability of the domain and application (Sintsova, Musat and Pu., 2004). Emotion speech features are mostly lower level features, and as such introduces the difficulty to extract and discriminate them. Until now, there exists no clear cut on which speech features are robust in distinguishing emotions (Zheng, Yu and Zou, 2015) – as these features are easily influenced by speakers, speaking styles, sentences, speaking rates, and

more. Further, research inconsistencies arise when these factors heavily influence the extracted speech features such as pitch and energy contours. Although numerous research works have identified good features of emotion speech signal, no widely acceptable set of speech characteristics has been determined. At the suprasegmental level, emotional conditions are governed by fundamental frequency (F0), intensity and temporal characteristics of speech, as there is the possibility that some segmental features may also be influenced by the speakers' emotional state (Katari, 2000; X. M. Cheng, P. Y. Cheng and L. Zhao, 2009). Previous works have highlighted the importance of syllables during emotions transmission and whereas clinical research methods adopted in prosody concentrate mostly on intonation, technological approaches have however focused on the entire speech signal without recourse to the qualitative variability of the spectral content (Origlia, Cutugno and Galati, 2014). Origlia, Cutugno and Galati (2014) proposed a feature extraction method that explores phonetic interpretation using the concept of syllable, and concentrated on the spectral content of syllabic nuclei, thus reducing the amount of information to be processed. They introduced feature weighting based on syllabic prominence, and evaluated their method on a continuous, three-dimensional model of emotions built on the classical axes of valence, activation and dominance. They found that their method compared favourably with state-of-art. In this paper, we investigate the influence of acoustic features on emotions, and the goal is to discover important acoustic-emotion profiles necessary for aiding emotion recognition systems design for African tone languages (ATLs). The language adopted for this investigation is Ibibio – an under-resourced tone language of the Lower Cross group, from the new Benue Congo language family – spoken by about 4,000,000 speakers in Akwa Ibom State, Nigeria, West Africa.

The remainder of this paper is organised as follows: section 2 reviews related works on the effect of acoustic features on emotions. Section 3 discusses the data collection procedure. Section 4 performs the speech feature extraction. Section 5 deals with the speech feature selection and dimension reduction process using principal component analysis (PCA). Section 6 visualises the acoustic feature-emotion profiles at the sentence, word and syllable levels. Section 7 concludes on the research and offers future research directions.

2. Related Works

Studies of the effects of emotion on acoustic characteristics of speech have shown that the average fundamental frequency (F0) values and ranges differ extensively from emotion to emotion, and which F0 contour spans the entire utterance or corpus. Several reasons have been offered why F0 changes with duration are potent at providing clues about the speaker's emotional state. First, considerable degree in the variations of F0 is expected, since only

certain aspects of the F0 contour carry useful information about the linguistic content of a message. The principal linguistic functions of F0 changes (useful) stress indicators and boundary markers of different types of utterance (word, phrase and sentence). William and Stevens [8] found that anger, fear, and sorrow situations tend to produce characteristic differences in the contour of fundamental frequency, average speech spectrum, temporal characteristics, precision of articulation, and waveform regularity of successive glottal pulses. Further, attributes for a given emotional situation were not always consistent from one speaker to another. Yildirim, Lee, Lee, Bulut, Busso, Kazemzadeh and Narayanan (S. Yildirim, S. Lee, C. M. Lee, M. Bulut, C. Busso, E. Kazemzadeh and S. Narayanan, 2004) analysed changes in temporal and acoustic parameters such as magnitude and variability of segmental duration, fundamental frequency and the first three formant frequencies as a function of emotion. They also explored acoustic differences among four emotions (neutral, sad, angry, happy). Their results showed that anger and happiness emotions were characterised by longer duration; shorter inter-word silence; higher pitch; and root mean squared (rms) energy with wider ranges. Sadness was distinguished from other emotions by lower rms energy and longer inter-word silence; and differences in formant pattern between (happiness/anger) and (neutral/sadness) were better reflected in back vowels than in front vowels. Zhang, Ching and Kong (Zhang, Ching and Kong, 2006) found that vocal expression of the following emotions (anger, fear, joy and sadness) showed specific characteristics as regards pitch (or F0) contour, intensity contour, and timing of utterance. They observed that anger gave the highest F0 and F0 variance, shortest sentence length, and highest short-time amplitude at sentence level. Their outcomes also compared well with the literature. Lin and Fon (Lin and Fon, 2012) investigated pitch and duration cues of emotion speech in Taiwan Mandarin, where a set of acted emotions (anger, joy, sorrow, fear and neutral) were recorded and analysed. Their results showed that F0 height and speech rate were more correlated with arousal dimension, which differentiate emotions of high arousal, such as anger and joy, from low arousal. However, negative emotions such as anger and sorrow, had longer lengthening than positive emotions. In (Guo, Yu, Hu and Y. Ding, 2016), a quantitative analysis of continuous speech emotion of Lhasa Tibetan (a Chinese tone language) was performed. Using pitch, energy and duration features, they investigated four basic emotion patterns (happy, surprise, sad, neutral), and found positive correlation between Lhasa Tibetan emotions and the studied features, and the pitch, energy and duration of negative emotion acoustic parameters appeared higher than positive emotions.

3. Data Collection

Handcrafted text was created to form the corpus for this study. The resultant corpus consists of a group of sentences that portrayed the target emotions.

Participants were made to act/simulate seven types of emotions (anger, fear, joy, normal, pride, sadness and surprise) under suboptimal environment (background effects, device/channel degradation, etc.). Two sentences were constructed for each emotion class. The constructed emotions and their respective gloss are shown in Table I. The recordings were done using a *zoom handy H4n* sound recorder. The choice of this recorder is in its high-quality recording (up to 24bit/96kHz), direct interfacing with a computer system and support for wave (.wav) format (to prevent a loss of fidelity). The recording was done using a sampling frequency rate of 44.1Mhz in stereo mode. Next, Audacity (a software for audio recording and editing) was used to convert the speech signal to mono mode. The reason for this conversion was to make it amenable to use in Matrix Laboratory (MATLAB) – the programming tool environment for this study. The recording sessions spanned a couple of days, as participants were given ample opportunity to rehearse the script before the recording sessions. Participants were made to repeat each sentence (at least) two times, and were given the freedom to act out the emotion (where necessary). Most participants preferred to introduce additional word(s) to enable them read the sentences successfully, and elicit the emotions properly. For instance, a female participant desired to include the word ḿbọ́k 'please', to cue the emotion utterance for anger: ɓákkɑ́ ké úsʌ́ʌ̄ ψák ḿbóíψɔ̀ 'leave the way for me to pass' – a frequently used word to cue such emotions – or to stimulate anger, while a male speaker desired to add the overtone sound 'o-o' to the end of the utterances for sadness – in order to stimulate grief. Another male speaker preferred to emphasize the object of reference while simulating the emotions. For instance, in the sentence: sɑ̀ɪ́! ìnɔ́ ɔ́dɔ̀ ɑ́tɔ́ɔ́nɔ́ ɑ̀dɪ́ ψɑ̀ɑ̀nɑ́ ɑ̀kɔ̀m úfɔ̀k '(exclamation) ... that thief has started removing the roofing sheets', ɑ̀kɔ̀m úfɔ̀k 'roofing sheets' was emphasized.

Table 1. Recorded emotions and their respective gloss

S/No.	Emotion type	Emotion sentence	English gloss
1.	Anger	(i) ʌ̀sínɑ́m ψɑ̀k ɑ̀sʊ́én̄ ɛ̀kɑ̀ m̀mì? (ii) ɓákkɑ́ ké úsʌ́ʌ̄ ψák ḿbóíψɔ̀	(i) Why have you disgraced my mother? (ii) Make way (leave the way) for me to pass
2.	Fear	(i) kɑ̀sé úrʌ̀kíkɔ̀t ɑ́ɓʌ́k úfɔ̀k m̀mì (ii) sɑ̀ɪ́! ìnɔ́ ɔ́ɓɔ̀ ɑ́tɔ́ɔ́nɔ̄ ɑ̀dɪ́ ψɑ̀ɑ̀nɑ́ ɑ̀-kɔ̀m úfɔ̀k	(i) Watch out! (or any other exclamatory start) ... A snake has entered my house (ii) (exclamation) ... that thief has started removing the roofing sheets
3.	Joy	(i) ɑnwɑ́ɑ́n m̀mì ɑ́mɑ̀n é-ψín ɑ́wɔ̀ɓɛ̀ɛ̀n (ii) éψén m̀mì ɑ́ψɑ́ ɑ́ɓɑ́kɑ́ɑ́ɓìɔ́ ḿbɑ̀kɑ́ rɑ́ ʋ̀kpɔ́nō	(i) My wife has given birth to a baby boy (ii) My child will leave for overseas tomorrow
4.	Normal	(i) ʋ̀ψɑ́ kɑ̀ɑ́ úfɔ̀k ʋ̀wɛ̀ɓ mɸín (ii) éψín ékɑ̀ m̀mì ɑ̀mésì-ɛ̀rɛ́ nɜ̀ɓɛ̀	(i) I will go to school today (ii) My brother (or sister), good morning to you too

S/No.	Emotion type	Emotion sentence	English gloss
5.	Pride	(i) ṅsɛ̀ ṅdìɑ̀ ṅkpɔ́ dáṅ̀ɑ́ mmámá (ii) ɛ̀tɛ́ mmì ánlɛ́ éfɑ́ɑ́k ámì	(i) I do eat as freely as I want to (ii) My father owns this street
6.	Sadness	(i) mmm ɛ̀tɛ́ mmì ámɑ̀ ákpɑ́ ṅkpɔ́ṅ̄ (ii) mmm ìdɛ́m ɛ̀kɑ̀ mmì ísɔ́ṅ̄ṅɔ́	(i) mmm ... my father just died yesterday (ii) mmm ... my mother is not feeling well
7.	Surprise	(i) íyɑ́! ákɛ́ dámmɑ́ ìdáɳɑ́ákɛ̀? (ii) ùwɔ́! ɑ̀fɔ̀ kɛ́ ékétɔ́n kɛ́ íkɑ́ṅ̄ ɑ́dɔ́?	(i) (exclamation) ... when did s/he turn mad? (ii) (exclamation) ... were you the victim of that gun-shot?

4. Speech Feature Extraction

First, *Praat* (version 4.1.43) – a speech processing and analysis software, was used to annotate the recorded emotion speech, and the recorded files were saved in wave (.wav) format. From the sound files, the TextGrid (a product of the annotation) was produced. The TextGrid files were then used in extracting the speech features. We focused on three tiers – sentence, word and syllable tiers, to obtain the respective units for these tiers. The annotations allowed for a structured and easily accessible speech corpus and are useful for future speech processing research. The following features were extracted using Praat scripting: duration, pitch/F0, intensity and formants.

Pitch/F0: Pitch represents the perceptual correlate of fundamental frequency (F0), which measures the rate of vibration of the vocal folds (in speech). It is the relative highness or lowness of a tone as perceived by the ear, and depends on the number of vibrations per second produced by the vocal cords. Pitch is the main acoustic correlate of tone and intonation.

Intensity: The intensity of a sound wave represents the power and loudness of the wave. Intensity correlates with the relative mean square (RMS) amplitude of the wave or how high above (compression) or below (rarefaction) the baseline the wave reaches in each cycle.

Formants: A formant is the concentration of acoustic energy around a particular frequency in a speech wave. It can be seen in a wideband spectrogram as dark bands. The first formant (F1) is inversely related to vowel height, i.e., the higher the (F1) formant frequency, the lower the vowel height (and vice versa). The second formant (F2) in vowels is somewhat related to degree of backness, i.e., the more front the vowel, the higher the second formant (but affected by lip-rounding). The lower of the (F2) formant frequency, the rounder shape of the lip (associated back vowel). Syrdal and Gopal (Syrdal and Gopal, 1986) found that the acoustic cues to vowel recognition are F0, F1, F2, F3 and F4. They normalized the vowels to a fully abstract, speaker independent representation, and observed that within each vowel, F0 and F4 represent speaking qualities, while F1, F2 and F3, are related to vowel identity. Further, F3 has been found to be related to lip spreading, while F4 is more related to lip protrusion (Isei-Jaakkola, Naka and

Hirose, 2010). A classic study in Peterson and Barney (1952) for instance, showed that the first two formant frequencies (F1 and F2), have significant variations among different speakers enunciating same vowel. In addition to demonstrating the overlap between the different vowel classes, the F1-F2 plane has been established as the most descriptive, two-dimensional representation of the phonetic quality of spoken vowel sounds.

In order to obtain a normalised speech signal sequence, a simple Praat script was used to scan the long-term F0, intensity and formant features (for sentence, word and syllable units), and the averages collated to yield each instance of the sentence, word or syllable. Tables 2, 3 and 4, show the normalised duration, F0, intensity, F1-F4 extractions for anger emotion of the first female and male speakers, at sentence, word and syllable levels, respectively.

Table 2. Normalised duration, F0, intensity, F1-F4 of male and female anger emotions for sentence unit

Sound Name	Interval Name	Duration	F0	Intensity	F1	F2	F3	F4
fl_anger	Nsinam yak asuenne eka mmi; dakka ke usVN yak mboyo	4.05	54.88	74.41	551.56	1664.51	2660.16	3825.23
ml_anger	nsinam yak asuenne eka mmi; dakka kusVN yak mboiyo	2.92	156.08	58.85	598.27	1659.12	2744.32	4040.73

Table 3. Normalised duration, F0, intensity, F1-F4 of male and female anger emotions for word unit

Sound Name	Interval Name	Duration	F0	Intensity	F1	F2	F3	F4
fl_anger	Nsinam	0.56	362.78	81.88	499.06	1678.99	2521.33	3564.02
	yak	0.18	253.08	85.23	690.99	1472.05	2375.54	3460.19
	asuenne	0.38	229.95	78.06	588.07	1778.77	2782.60	4063.38
	eka	0.22	173.64	74.52	617.12	2063.97	2755.62	4057.35
	mmi	0.39	152.21	68.38	425.23	1635.55	2627.59	3762.61
	Dakka	0.43	278.38	77.69	785.36	1717.14	2605.53	3964.12
	kusVN	0.44	339.94	79.22	493.87	1216.01	2682.29	3900.62
	yak	0.24	211.33	80.31	666.89	1518.31	2509.39	3537.95
	mboyo	0.64	201.50	76.83	398.92	1756.82	2736.78	3823.50
ml_anger	nsinam	0.43	210.79	59.61	713.40	1977.09	2961.17	4178.92
	yak	0.14	175.38	64.81	650.59	1489.51	2490.41	3710.70
	asuenne	0.36	133.28	60.20	540.44	1626.55	2743.84	3943.28
	eka	0.29	132.42	54.47	544.00	1460.88	2147.81	3830.30
	mmi	0.28	109.40	47.91	635.32	1842.19	3010.46	4331.63
	dakka	0.29	159.87	61.29	747.98	1265.57	2574.48	3756.40
	kusVN	0.29	212.96	62.09	591.12	1924.89	3223.33	4106.38
	yak	0.26	152.25	65.22	602.15	1355.93	2347.64	3812.90
	mboiyo	0.52	132.73	59.14	429.14	1659.80	2794.29	4271.31

Table 4. Normalised duration, F0, intensity, F1-F4 of male and female anger emotions for syllable unit

Sound Name	Interval Name	Duration	F0	Intensity	F1	F2	F3	F4
f1_anger	N	0.10	226.03	77.37	398.08	1684.00	2771.38	4226.46
	si	0.24	336.39	79.67	427.27	1760.05	2717.86	3587.20
	nam	0.22	440.27	85.80	614.57	1589.00	2214.89	3293.66
	yak	0.18	253.08	85.23	690.99	1472.05	2375.54	3460.19
	a	0.09	207.89	78.24	732.42	1558.01	2549.41	3562.79
	suen	0.18	250.32	77.34	598.67	1744.59	2792.08	4170.98
	ne	0.11	226.66	79.06	455.17	2010.70	2954.82	4293.73
	e	0.13	180.91	74.53	507.06	2289.76	2851.36	4190.88
	ka	0.09	166.04	74.51	781.60	1726.52	2612.53	3857.78
	m	0.14	159.13	69.77	584.32	1608.67	2640.42	3989.08
	mi	0.25	147.81	67.59	335.77	1650.67	2620.38	3635.26
	dak	0.19	220.11	76.68	692.18	1893.62	2767.38	3954.78
	ka	0.24	323.57	78.48	857.65	1580.23	2479.98	3971.34
	ku	0.12	343.73	82.36	446.76	1194.55	2250.52	3734.76
	sVN	0.32	338.29	78.01	511.93	1224.23	2847.74	3964.17
	yak	0.24	211.28	80.29	666.15	1517.81	2509.20	3534.94
	m	0.07	213.15	76.82	600.16	1687.68	2782.97	4020.44
	bo	0.16	223.61	83.32	430.44	1423.70	2652.76	3680.07
	yo	0.41	187.91	74.11	349.67	1907.88	2761.86	3846.22
m1_anger	nsi	0.20	198.30	55.02	826.94	2522.43	3786.88	4638.15
	nam	0.24	214.54	62.86	629.30	1573.14	2349.55	3887.48
	yak	0.14	175.38	64.81	650.59	1489.51	2490.41	3710.70
	a	0.13	71.59	60.03	663.56	1424.16	3050.41	4057.29
	suen	0.13	144.22	59.26	522.29	1614.15	2787.16	3938.14
	ne	0.11	149.83	61.54	417.37	1880.13	2329.93	3815.14
	e	0.09	146.14	57.17	394.23	1848.68	2213.86	3875.11
	ka	0.20	124.96	53.22	613.12	1281.90	2117.33	3808.22
	m	0.17	109.03	50.93	574.37	1723.37	2877.00	4346.20
	mi	0.10	110.41	42.77	739.18	2044.61	3237.84	4306.81
	dak	0.14	143.91	59.45	708.00	1323.45	2733.33	4063.07
	ka	0.15	174.02	62.94	783.96	1213.48	2431.51	3480.40
	ku	0.10	211.22	62.35	593.57	1567.90	3008.23	3839.89
	sVN	0.19	213.84	61.96	589.88	2106.88	3332.98	4229.00
	yak	0.26	152.25	65.22	602.15	1355.93	2347.64	3812.90
	m	0.08	150.58	63.69	455.57	1197.24	2507.53	3937.47
	boi	0.17	149.86	62.73	414.46	1656.46	2694.71	4308.99
	yo	0.27	112.41	55.39	430.44	1801.30	2943.58	4348.13

5. Feature Selection and Dimension Reduction

Our choice of an unsupervised technique to feature selection is that in many applications, the class labels are unknown (Dash and Liu, 1997). A PCA-based unsupervised selection algorithm (Luo, Xiong and Wang, 2008) was adopted in this work to select the dominant principal component features, and eliminate redundant feature frames that hitherto would have contributed to poor selection. PCA is used in this study because, (i.) it is a powerful tool to visualise high dimensional data, (ii.) it shows quantified difference among observations, (iii.) it is useful for assessing data quality and for the discovery of relationship/variability between data points.

Given an input space $\mathbb{R}^D$ and target space $\mathbb{R}^d$; d << D, let $X \in \mathbb{R}^{N\times D}$ be an input dataset of N samples and D features, and $X \in \mathbb{R}^{N\times d}$ its low-dimensional representation. A dimension reduction technique is a mapping $\phi: \mathbb{R}^D \to \mathbb{R}^d$ that optimises a cost function $\in$: R d $\to \mathbb{R}$ on the target space. This problem can often be reduced to an eigenvalue problem whose eigenvectors defines the embedding Y. Assuming a training set with N samples $\{x_i\}_{i=1}^N$, each sample represented by an n-dimensional vector $x_i = [x_{1i}, x_{2i}, \dots, x_{ni}]^T$, PCA can be considered as a linear transform that maps data from the original measurement space to a new space populated by a set of new variables. Suppose the linear transform is denoted by matrix L, then pattern x in the new space is represented as,

$$y = L^T x \tag{1}$$

where $y = [y_1, y_2, \dots, x_d]^T$, $L = [q_1, q_2, \dots, q_d]^T$ and,

$$q_j^T = [q_{1j}, x_{2j}, \dots, x_{nj}],\ j = 1,2,\dots,d \tag{2}$$

where $d \le n$, but most often $d \ll n$. The new variables y_j, $j = 1,2,\dots,d$, are called principal components (PCs). Now, consider the projection of x_i on the k principal axis, then,

$$y_{ki} = q_k^T x_i = \sum_{j=1}^{n} q_{jk} x_{ik} \tag{3}$$

As seen in equation (3), the projection of a sample on the principal axis is a linear combination of all variables. However, some of the variables might be redundant, irrelevant or insignificant, which indicates that feature selection can only be achieved through the identification of subset of variables whose roles are critical in determining data projections on the principal axes. We observe here that the significance of a single variable x_j can be evaluated based on the value of the corresponding coefficient q_{jk}. An approximate method (Dash and Liu, 1997) is therefore introduced for feature selection in two inter-related steps: (i) select a subset of relevant features; and (ii) select critical features from the relevant features. Further, a recurrence definition of principal component dominant feature (PCDF) about y_k can be defined as follows:

(i) for a specific principal component y_k, a variance with the largest coefficient in the component is a PCDF;

(ii) for the remaining features, if x_j is a relevant feature about y_k, i.e., $\rho(x_j, y_k) > \omega$, and there exists no PCDF x_p – which is subject to $\frac{|\rho(x_j,x_p)|}{|\rho(x_j,x_p)|} \ge \theta\ (0 < \theta \le 1)$, then features x_j is a PCDF about y_k.

6. Acoustic-Emotion Profiles Visualisation

Sentence Unit Visualisation

Table 5 shows an extraction of the first three principal components at sentence level from the speech features of the recorded emotions. Observe

that F1 and F3 features of all the emotions within the first principal component (PC1) for male and female speakers were most dominant (i.e., with eigen values above 1 or |-1|), and captures the most variance, i.e., 99% for male speakers, and 100% for female speakers (see Fig. 1.). From the visualisation plots, it appears that all the speech features (F1, F2, F3) exhibited major differences between the emotion profiles, and are useful for modelling emotion speech variability in Ibibio.

Table 5. Extraction of the first three components for sentence unit

Speech feature	Principal component					
	Male speaker			Female speaker		
	1	2	3	1	2	3
F1_Anger	**-1.2018**	-0.2073	-0.6118	**-1.2081**	-0.4215	-0.2414
F1_Fear	**-1.1782**	-0.0877	-0.1810	**-1.1809**	-0.0157	-0.3120
F1_Joy	**-1.2263**	-0.9512	-0.3669	**-1.1970**	-0.4856	-0.9618
F1_Normal	**-1.2565**	0.3280	-0.3360	**-1.2528**	-0.0744	-0.2215
F1_Pride	**-1.2208**	**-1.0176**	**-1.1988**	**-1.2156**	-0.1714	-0.6754
F1_Sadness	**-1.1734**	**1.1862**	-0.2056	**-1.2477**	**1.8440**	1.0251
F1_Surprise	**-1.1833**	0.3446	-0.0769	**-1.1676**	0.1450	-1.4817
F2_Anger	0.0173	-0.6416	0.3749	0.0412	0.0884	0.4074
F2_Fear	-0.0226	0.8707	**2.4728**	-0.0478	**-1.8127**	2.0674
F2_Joy	0.0057	-0.7485	**1.3864**	0.0300	0.7597	-0.5447
F2_Normal	0.0389	0.1431	0.4308	0.1011	-0.7209	1.3030
F2_Pride	0.0157	-0.7693	-0.1723	0.0670	0.3593	0.5323
F2_Sadness	0.1256	**1.4282**	0.7760	0.0282	**1.3290**	1.7087
F2_Surprise	-0.0044	**1.0833**	0.8873	0.0058	**-1.6350**	0.5096
F3_Anger	**1.1760**	0.6026	**-1.0610**	**1.1624**	0.0612	-1.3066
F3_Fear	**1.2530**	0.2265	0.8210	**1.2087**	-0.6336	-0.3584
F3_Joy	**1.0888**	**-1.9481**	0.4195	**1.1060**	1.3133	-0.6855
F3_Normal	**1.1252**	0.6103	**-1.8592**	**1.1582**	-0.6759	-0.2771
F3_Pride	**1.1105**	**-2.1764**	0.2933	**1.1544**	0.5465	-1.2153
F3_Sadness	**1.3249**	0.9485	-0.2059	**1.2668**	**1.5606**	1.1210
F3_Surprise	**1.1856**	0.7757	**-1.5865**	**1.1879**	**-1.3602**	-0.3931

Word Unit Visualisation

In Table 6, extraction of the first three principal components at word level from speech features of the respective emotions is presented. As can be seen in the table, F1 and F3 features exhibited major differences in emotion

profiles for male speakers, while F0 and F4 features exhibited major differences in emotion profiles in female speakers.

Table 6: Extraction of the first three components for word unit

Speech feature	Principal component					
	Male speaker			Speech feature	Female speaker		
	1	2	3		1	2	3
F1_Anger	**-1.2396**	-0.1176	-0.3342	F0_Anger	**-1.3336**	0.0546	-0.0281
F1_Fear	**-1.1121**	0.7317	-0.1651	F0_Fear	**-1.2769**	0.3279	-0.0001
F1_Joy	**-1.1542**	0.4120	0.2515	F0_Joy	**-1.2981**	0.1829	0.0939
F1_Normal	**-1.2502**	-0.0166	0.0333	F0_Normal	**-1.3477**	0.0913	0.0659
F1_Pride	**-1.2390**	-0.0410	0.0156	F0_Pride	**-1.3364**	0.0781	0.0499
F1_Sadness	**-1.1925**	0.1432	0.3574	F0_Sadness	**-1.3415**	0.1395	0.1536
F1_Surprise	**-1.1648**	0.1589	0.3460	F0_Surprise	**-1.2630**	0.1146	0.1591
F2_Anger	-0.2170	-0.8749	**-1.1545**	F3_Anger	0.0858	**-1.0614**	**-1.2826**
F2_Fear	0.1103	**1.6211**	-0.8802	F3_Fear	0.5032	**1.9123**	**-1.3034**
F2_Joy	0.1810	0.7299	0.4043	F3_Joy	0.4093	0.8019	0.6199
F2_Normal	-0.0802	-0.6937	-0.1090	F3_Normal	0.2390	-0.8086	-0.2011
F2_Pride	-0.0186	-0.8008	-0.3120	F3_Pride	0.2409	-0.8925	-0.4508
F2_Sadness	0.1568	-0.1359	**1.0106**	F3_Sadness	0.4040	-0.0693	**1.2930**
F2_Surprise	0.0709	-0.0007	**1.0830**	F3_Surprise	0.3809	-0.0689	**1.2426**
F3_Anger	0.7935	**-1.6175**	-2.0090	F4_Anger	0.6957	**-1.5373**	**-1.8189**
F3_Fear	**1.5395**	**2.5823**	**-1.8975**	F4_Fear	**1.1923**	**2.5275**	**-1.6789**
F3_Joy	**1.3044**	**1.0531**	0.8233	F4_Joy	**1.1278**	**1.0302**	0.8833
F3_Normal	0.9657	**-1.3030**	-0.3755	F4_Normal	0.9111	**-1.2025**	-0.3772
F3_Pride	0.9827	**-1.4081**	-0.5981	F4_Pride	0.8509	**-1.2652**	-0.6616
F3_Sadness	**1.3310**	-0.2510	**1.7478**	F4_Sadness	**1.1004**	-0.1634	**1.6752**
F3_Surprise	**1.2322**	-0.1714	**1.7625**	F4_Surprise	**1.0559**	-0.1917	**1.5664**

The visualisation plots in Fig. 2 indicates that feature patterns of F1 (for male speakers – Fig. 2 (a)) and F1 (for female speakers – Fig. 2 (b)) tend to cluster together, meaning that speech features in this class maintained similar emotion profiles. Hence, F2 and F3 features in male speakers showed high variability with PC1 (73%) capturing the most variance, while PC2 (14%) and PC3 (5%) captured the least variances. In female speakers: F3 and F4 features showed high variability, with PC1 (79%) capturing the most variance, while PC2 (12%) and PC3 (5%) captured the least variances. We can deduce here that F2, F3 and F4 features exhibit high variability in the emotion profiles, and are useful for investigating emotion feature variability in Ibibio.

Syllable Unit Visualisation

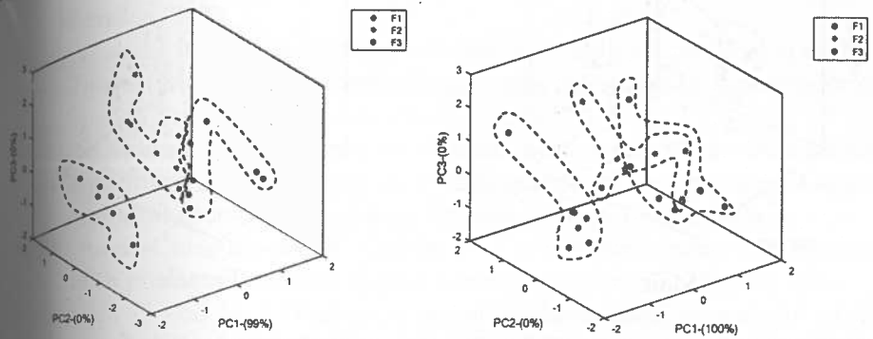
In Table 7, the extraction of the first three principal components at syllable level for obtained speech features of the recorded emotions is presented. As seen in the table, PC1 of the emotion profiles for both speakers does not show significance for F2 and F3 features (for male speakers), and F1 and F4 features (for female speakers).

Table 7: Extraction of the first three components for syllable unit

Speech feature	Principal component						
	Male speaker				Female speaker		
	1	2	3	Speech feature	1	2	3
F2_Anger	**-1.3316**	-0.0416	**-1.0427**	F1_Anger	**-1.3357**	0.0016	-0.2396
F2_Fear	-0.9521	**1.1708**	0.0284	F1_Fear	**-1.2507**	0.4748	-0.0627
F2_Joy	-0.8108	**1.1629**	-0.2291	F1_Joy	**-1.2218**	0.4258	-0.1029
F2_Normal	**-1.2004**	0.0012	0.3795	F1_Normal	**-1.3479**	0.0366	0.0715
F2_Pride	**-1.1719**	-0.0120	0.3235	F1_Pride	**-1.3278**	0.0377	0.1261
F2_Sadness	**-1.1450**	-0.0459	0.1130	F1_Sadness	**-1.3494**	0.0554	0.1842
F2_Surprise	**-1.0580**	0.3108	0.9626	F1_Surprise	**-1.2544**	0.1512	0.3717
F3_Anger	-0.4310	-0.6811	**-1.8223**	F3_Anger	0.0579	-0.8558	-1.5995
F3_Fear	0.4597	**1.4244**	-0.3767	F3_Fear	0.5014	**1.5222**	-0.5718
F3_Joy	0.2978	**1.3791**	-0.2858	F3_Joy	0.4377	**1.3996**	-0.3625
F3_Normal	-0.2833	-0.5797	0.2876	F3_Normal	0.1706	-0.7887	-0.0768
F3_Pride	-0.2653	-0.5762	0.3173	F3_Pride	0.2055	-0.7534	0.1686
F3_Sadness	-0.1624	-0.6289	0.3214	F3_Sadness	0.2512	-0.7454	0.3720
F3_Surprise	-0.0294	-0.0537	**1.5948**	F3_Surprise	0.3319	-0.0061	2.0097
F4_Anger	0.5887	**-1.3983**	**-2.5719**	F4_Anger	0.7537	**-1.2771**	-2.2892
F4_Fear	**1.8921**	**1.5552**	-0.3047	F4_Fear	**1.2656**	**1.9577**	-0.6281
F4_Joy	**1.7243**	**1.4241**	-0.5394	F4_Joy	**1.2669**	**1.8717**	-0.3199
F4_Normal	0.8765	**-1.3062**	0.2383	F4_Normal	0.9035	**-1.1886**	-0.2251
F4_Pride	0.8189	**-1.2561**	0.2956	F4_Pride	0.8866	**-1.0982**	0.1778
F4_Sadness	0.9636	**-1.3247**	0.2936	F4_Sadness	0.9785	**-1.1003**	0.4498
F4_Surprise	**1.2195**	-0.5240	**2.0171**	F4_Surprise	**1.0765**	-0.1207	2.5466

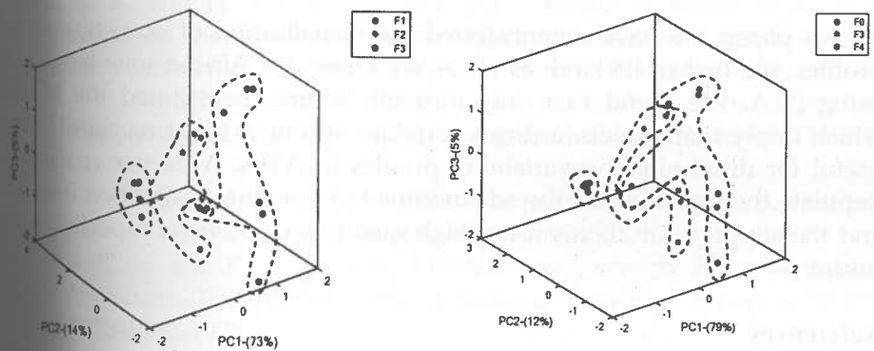
The visualisation plots in Fig. 3 indicate that all the selected features separate into independent clusters. Hence, in male speakers, F3 and F4 features show high features variability with the most variance captured by PC1 (57%), while PC2 (29%) and PC3 (5%) captured the least variances. Similar observations go for female speakers with the most variance captured by PC1 (74%), while PC2 (19%) and PC3 (4%) captured the least variances. F2 and F1 features appear to show close similarities in the emotion profiles for

male and female speakers, respectively, but the bonding appear stronger in female speakers than male speakers. Hence F3 and F4 features constitute high variability emotion profiles, and are useful for investigating emotion feature variability in Ibibio.



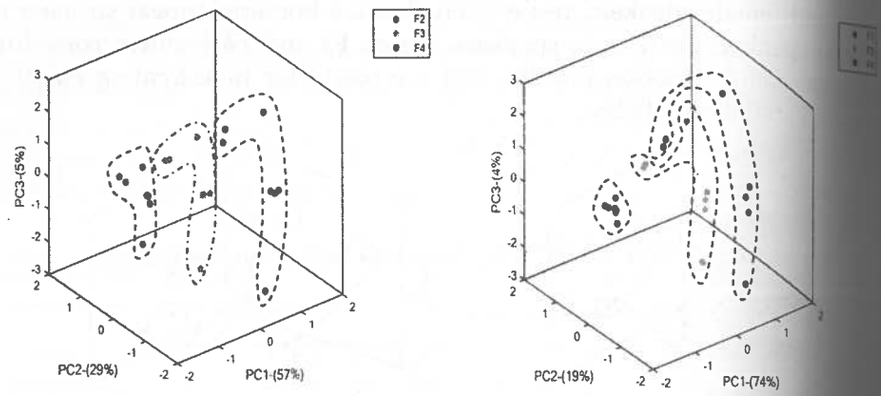
Male Female

Fig. 1. Visualisation of sentence level principal components for male and female speakers



Male Female

Fig. 2: Visualisation of word level principal components for male and female speakers



Male Female

Fig. 3: Visualisation of syllable level principal components for male and female

7. Conclusion and Future Research Direction

In this paper, we have demonstrated the visualisation of acoustic-emotion profiles, the first of its kind, as far as we know, for African tone languages, using PCA. We found that only formant features constituted the PCDFs, which imply that vowels, a direct correlate of tone is most responsible and useful for discriminating variability profiles in ATLs. A deeper research to populate the literature for the advancement of emotion recognition in ATLS, and the support our claims using high quality recordings is expected in the future.

References

Cheng, X. M., P. Y. Cheng and L. Zhao. A Study of Emotional Feature Analysis and Recognition in Speech Signal. In Proceedings of IEEE International Conference on Mastering Technology and Mechatronics Automation, IEEE Publishers: 418-420, 2009.

Dash, M. and H. Liu, H. Dimensionality reduction for unsupervised data. In Proceedings of 9th IEEE International Congress on Tools with AI, Newport Beach, CA, 522-539, 1997.

Douglas-Cowie, E., N. Campbell, R. Cowie and P. Roach. Emotional Speech: Towards a New Generation of Databases. *Speech Communication*, 40, 33-60, 2003.

Guo, H., Yu, A. Hu and Y. Ding. Statistical analysis of acoustic characteristics of Tibetan Lhasa dialect speech emotion. In Proceedings of SHS Web of Conferences 25, EDP Sciences, 1-5, 2016.

Isei-Jaakkola, T., T. Naka and K. Hirose. Comparison of the formant frequencies F3 and F4 on a three-dimensional vowel chart. The Journal of the Acoustical Society of America, 127(3), 2019-2019, 2010.

Katari. S., *Handbook of Neural Network for Speech Processing*, Artecch House, London, 2000.

Lin., H. Y. and J. Fon. Prosodic and acoustic features of emotional speech in Taiwan Mandarin. In Proceedings of 6th International Conference on Speech Prosody, Shanghai, China, 1-4, 2012.

Luo, Y., S. Xiong and S. Wang. A PCA based unsupervised feature selection algorithm. In Proceedings of 2nd International Conference on Genetic and Evolutionary Computing, 299-302. Hubei, China, 2008.

Origlia, A., F. Cutugno and V. Galatà, V. Continuous emotion recognition with phonetic syllables. Speech Communication, 57, 155-169, 2014.

Peterson, G. E. and H. L. Barney. Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2), 175-184, 1952.

Sintsova, V., C. Musat and P. Pu. Semi-Supervised Method for Multi-Category Emotion Recognition in

Syrdal, A. K. and H. S. Gopal. A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086 1100, 1986.

Taylor, J. C., K. Scherer and R. Cowie. Emotion and Brain: Understanding Emotions and Modelling their Recognition. *Neural Networks*, 18,313-31, 2005.

Tweets. In Proceedings of 2014 IEEE International Conference on Data Mining Workshop, IEEE Publishers, 393-402, 2004.

Williams, C. E. and K. N. Stevens. Emotions and speech: Some acoustical correlates. The Journal of the Acoustical Society of America, 52(4B), 1238-1250, 1972.

Yildirim, S., S. Lee, C. M. Lee, M. Bulut, C. Busso, E. Kazemzadeh and S. Narayanan. Study of acoustic correlates associate with emotional speech. The Journal of the Acoustical Society of America, 116(4), 2481-2481, 2004.

Zhang, S., P. C. Ching and F. Kong. Acoustic analysis of emotional speech in Mandarin Chinese. In Proceedings of ISCSLP, Kent Ridge, Singapore, 57-66, 2006.

Zheng, W. Q., J. S. Yu and Y. X. Zou. An Experimental Study of Speech Emotion Recognition Based on Deep Convolutional Neural Networks. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII '15), Xi'an, China, 827–831, 2015.