Speaker Variability for Emotions Classification in African Tone Languages

Moses Ekpenyong[1](✉), Udoinyang Inyang[1], Nnamso Umoh[1], Temitope Fakiyesi[1], Okokon Akpan[1], and Nseobong Uto[2]

[1] University of Uyo, P.M.B. 1017, Uyo 520003, Nigeria
{mosesekpenyong,udoinyanginyang}@uniuyo.edu.ng,
nnamso_obong@yahoo.com, okokonakpan@uniuyyo.edu.ng
[2] University of Saint Andrews, St Andrews KY 169SS, UK
npu@st-andrews.ac.uk

Abstract. In this paper, we investigate the effect of speaker variability on emotions and languages, and propose a classification system. To achieve these, speech features such as the fundamental frequency (F0) and intensity of two languages (Ibibio, New Benue Congo and Yoruba, Niger Congo) were exploited. A total of 20 speakers (10 males and 10 females) were recorded and speech features extracted for analysis. A methodological framework consisting of 4 main components: speech recording, knowledge base, preprocessing and analytic. ANOVA was used to test the intra- and inter-variability among speakers of various languages on emotions and languages, while the predictive analysis was carried out using support vector machine (SVM). We observed that language and speech features are dependent on speakers' characteristics. Furthermore, there exists a highly significant variability in the effect of emotions and languages on speech features. Results of SVM classification yielded 66.04% accuracy for emotions classification and 79.40% accuracy for language classification. Hence, classification performance favored the language classifier compared to emotion classifier, as the former produced low root mean squared error (RMSE) when compared with the later.

1 Introduction

Speech is a natural way of communication between humans. It represents a reliable footprint that embeds phonetic and emotional characteristics of any speaker. A speech signal therefore provides cues for expressing emotions – as it represents time-varying indicator that conveys multiple layers of information (words, syllables, languages, etc.). This phenomenon does not only convey linguistic content of a message but also the expression of attitude and speakers' emotion (Mozziconacci and Hermes 1999). The role of speech features on classification performance is vital in speaker and emotion recognition, as most of the existing recognition systems are executed under certain acoustic conditions. Features commonly utilized in speech based emotion classification systems should capture both emotion-specific information and speaker-specific information (Sethu et al. 2013), and the absolute fidelity of content that defines the classification performance

Z. Vetulani et al. (Eds.): LTC 2017, LNAI 12598, pp. 174–185, 2020.
https://doi.org/10.1007/978-3-030-66527-2_13

rests on how the speech is produced by the speaker (‘acted’ vs. ‘spontaneous’) and/or the environment in which the speech is produced (‘optimal’ vs. ‘suboptimal’). Batliner and Huber (2007) addressed these interrelated issues of speaker characteristics (personalization) and suboptimal performance of emotion classification, with the argument that:

- inherent multi-functionality and speaker-dependency of speakers makes its use as a feature in emotion classification less promising, and
- constraints on time and budget often prevent the implementation of an optimal emotion recognition module.

Ideally, the only source of variability in extracted speech features arise from differences in the emotions being expressed. Speech features variability may also come from other reasons as well, including linguistic content (differences between what is being said) and speaker identity (differences between who said it), and these additional sources of variability are known to degrade classification performance (Cao et al. 2015; Chakraborty et al. 2017). While previous studies discriminate speakers using static fundamental frequency (F0) parameters, recent works focus on the dynamic and linguistically structured aspects of F0 – owing to the dynamic nature of lexical tones. Chan (2016) explored the speaker-discriminatory power of individual lexical tones and of the height relationship of level tone pairs in Cantonese, and the effects of voice level and linguistic condition on their realization. Results showed that F0 height and F0 dynamics are separate dimensions of a tone and are affected by voice level and linguistic condition in different ways. Moreover, discriminant analyses reveal that the contours of individual tones and the height differences of level tone pairs are useful parameters for characterizing speakers.

It is assumed that the emotion system is governed by the central nervous system and it is fast to react, able to switch quickly from one state to another, and produces only one emotion instance at a time. However, the intensity of emotion is a non-monotonic function of deterrence to the goal of emotion. Several experiments using supporting data as well as selected theoretical problems have been carried out to support these assumptions (c.f. Brehm 1999). Each emotion induces physiological changes which directly affect speech (Kassam and Mendes 2013). Physiological changes include affects in measures of speech features such as pitch, intensity and speech rate or duration (Kim 2007). High arousal emotions trigger higher values of speech features. For instance, anger and joy emotions have same arousal state, but differ in affect (positive and negative valence), and consequently raising serious concerns on how to accurately discriminate emotions that are at the same level of arousal, and those that have lower values, and in same arousal space. Sethu et al. (2008) investigated the effect of speaker- and phoneme-specific information on speech-based automatic emotion classification. They compared the performance of the classification system using established acoustic and prosodic features (pitch, energy, zero crossing rate and energy slope) for different phonemes, in both speaker-dependent and speaker-independent modes, using the linguistic data consortium (LDC) emotional prosody speech corpus comprising of speech from professional actors trying to express emotions while reading short phrases consisting of dates and numbers in order to ensure no semantic or contextual information is available. Their results indicate that speaker variability is more significant than phonetic variations; and features commonly used in emotion classification systems do not completely disassociate emotion-specific information from speaker-specific information (Batliner and Huber 2007).

This paper investigates the intra- and inter-variability of speech features on speakers, emotions and languages. The speech features considered include the fundamental frequency: F0 (the acoustic correlate of speech) and intensity (a measure of loudness). A support vector machine (SVM) classification system is then developed to predict emotions and languages. The remainder of this paper is outlined as follows: Sect. 2 discusses the methods and includes the proposed system framework and their respective components. Section 3 presents the results, discussing the intra- and inter-variability analysis and the classification results. Section 4 concludes on study and offers future research perspective.

2 Methods

2.1 Proposed System Framework

The framework defining our methodology is presented in Fig. 1, and consists of four main components: speech recording, knowledge base, preprocessing and analytic modeling phases. The speech recording component captures the various speech emotions from speakers of various languages. In this paper, speech recordings were obtained from native speakers of Ibibio (New Benue Congo, Nigeria) and Yoruba (Niger Congo, Nigeria). The speech sources might emanate from multiple locations/sources/speakers—homogenous or heterogeneous, therefore may be having significantly varying and inconsistent data formats and types, ambiguous, poor quality and may pose some challenges during analysis. Pre-processing is an essential task adopted to adapt heterogeneous speech corpora into a homogenous corpus. This work reduces pre-processing into five stages as follows 1) data cleaning 2) transformation 3) Integration 4) feature extraction and 5) feature selection. Data cleaning and transformation detect and remove outliers, insert missing entries as well as other operations required to standardize the data-points into a format that is computationally less expensive to model. Both stages produce a reconciled version of the hitherto heterogeneous speech corpora, and make it suitable for automatic speech corpora integration. During integration, the speech corpora was fused into a uniform and consistent version through schema integration approaches, object matching and redundancy removal, and pushes them into the speech feature database. Syllable units of two speech features, F0 and intensity were extracted using a Praat script. These features are selected in addition to each speaker identity and emotion for the analytic phase.

The analytic engine performs two main tasks: speech feature variability analysis, and emotion and language predictive analytics. The speech feature variability component performs intra-language and inter-language variability assessments with analysis of variance (ANOVA) test, while the predictive analytics component builds and executes the support vector machine (SVM) model. The results are finally produced to a decision support engine for appropriate evaluation.

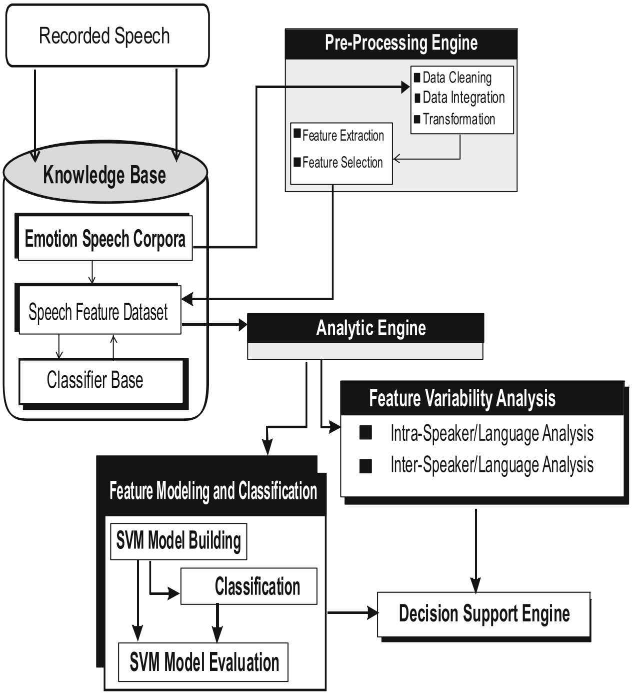

Fig. 1. Proposed system framework.

Emotion Speech Corpora. A total of 20 native speakers, 10 from each language (Ibibio and Yoruba), were selected for this study. The speakers were presented with two speech corpora that embed two negative emotions (anger and sadness), and told to act naturally, the respective emotions. Each speaker was recorded twice and the best speaking style selected. Table 1 documents the emotion corpora (column 4) use in this study with translations into Ibibio and Yoruba (column 3). Figure 2 shows a Praat speech analysis window showing the waveforms, spectrogram, point and syllable TextGrid annotations. In this paper, we are interested in the overall effect of the syllable units rather than specific units. Hence the blind labeling of the syllable units with the repeating label 'syl'. A future study is expected to address the effect of specific units with the right labels and compare their performance with other units such as phonemes and words.

Table 1. Recorded emotions speech corpora

Language	Emotion	Recorded speech	English gloss
Ibibio	Anger	sọp idem nyaak afọn. nsinam anduok etab? nsinam asueen̄ eka mmi? eka mmi ado n̄ka mfo? yak akuppọ utebe inua mfo asọn̄ anye anyen. usukponoke owo? akpe maana asio uyo; nya ubeek edet ado	Come on! leave my shirt/dress. Why spit on me? Why insult my mother? Is my mother your mate? That you open your dirty mouth to insult her? Don't you have respect? If you speak again; I'll destroy your dentition
Yoruba	Anger	jowo fi aso mi sile. kilode ti ofi tuto simi lara? kilode ti ofi dojuti iya mi? se egbere ni iya mi je? ti ofi ya enu buruku re lati soro si won. se iwo koni aponle ni? to ba tun soro; mo ma ba eyin reje	
Ibibio	Sadness	hmmm! mbre mbre, udọn̄ọ ami aya awod owo ama. nso ke adodo? afịd awo edifefeen̄e korona, awo ikọọmọ owo ubọk aba. abasi mmi! ubọk mfo-o! hmmm! n̄kitan̄a abiọọn. idaha ami, ndiweek in̄week ke anyen. abiọọn aya awot awo ama	Hmm! Bit by bit, this sickness will kill everyone. What is it? Everyone is afraid of coronavirus, no one greets with hand again! My God! I your hands I rest. Hmmm! Not to mention hunger. Now I breathe through the eyes. Hunger will kill us all
Yoruba	Sadness	hmm! die die, aisan yi yoo pa gbogbo eniyan. kini gan? gbogbo eniyan lohun beru korona, kosi eniti ofe bo eniyan lowo mo. oluwa, saanu! Hmm, ka ma tiso tebi. nisinyi, oju ni mofi hun mi. ebi yoo paniyan	

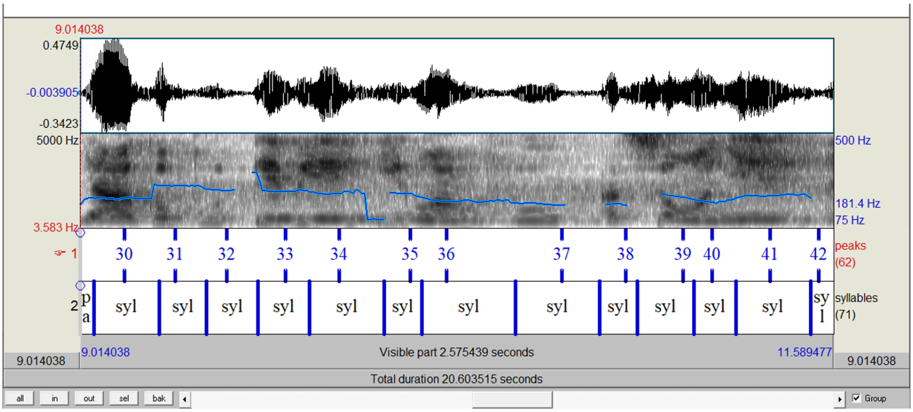

Fig. 2. Praat speech analysis window of a recorded speech corpus

Feature Extraction and Selection. A Praat script was then written to extract the syllable units of the pitch (F0) and intensity features. Syllable units were used because they are the closest and most stable features for detecting speech fluency and pronunciation, and reveals clearly, the syllable nucleus (most often a vowel) and an optional initial and final margins (typically, consonants). The extracted features were then labeled to form the speech feature datasets to which the classifier base connects with. A snippet of the labeled datasets is given in Table 2, with the parameters coded to reflect the source of

Table 2. Labeled datasets

Speaker	Emotion	Language	F0	Intensity
IM8	AN	IB	96.89	68.78
IM8	AN	IB	121.31	72.73
IM8	AN	IB	122.25	57.32
IM8	AN	IB	123.15	66.83
IM8	AN	IB	125.43	78.21
IM8	AN	IB	128.66	74.75
IM8	AN	IB	131.01	76.41
IM8	AN	IB	132.97	77.80
IM8	AN	IB	133.52	69.38
IM8	AN	IB	136.55	70.39
IM8	AN	IB	138.40	77.51
IM1	SAD	IB	112.49	72.61
IM1	SAD	IB	114.23	70.84
IM1	SAD	IB	116.95	64.87
IM1	SAD	IB	117.22	71.86
IM1	SAD	IB	118.38	74.74
IM1	SAD	IB	120.36	65.38
IM1	SAD	IB	120.56	74.50
IM1	SAD	IB	120.74	67.36
IM1	SAD	IB	123.00	71.33
IM1	SAD	IB	126.30	74.10
IM1	SAD	IB	126.67	72.81
IM1	SAD	IB	127.98	70.85
IM1	SAD	IB	128.06	77.61
YM1	SAD	YU	197.99	41.29
YM1	SAD	YU	199.13	53.44
YM1	SAD	YU	199.54	57.60
YM1	SAD	YU	200.03	62.04
YM1	SAD	YU	200.18	60.20
YM1	SAD	YU	201.50	60.70
YM6	AN	YU	188.93	61.01
YM6	AN	YU	191.18	58.54

(*continued*)

Table 2. (*continued*)

Speaker	Emotion	Language	F0	Intensity
YM6	AN	YU	191.84	57.56
YM6	AN	YU	192.02	59.06
YM6	AN	YU	194.95	57.04
YM2	AN	YU	195.32	59.04
YM2	AN	YU	195.80	47.12
YM2	AN	YU	199.13	50.13
YM2	AN	YU	202.12	53.76
YM2	AN	YU	205.21	63.15
YM2	AN	YU	206.54	61.68
YM2	AN	YU	207.57	57.26
YM2	AN	YU	207.83	66.04

the input, e.g. IM8 codes the eighth speaker of the Ibibio language, AN codes the anger emotion, and IB codes the Ibibio speaker.

3 Results

3.1 Intra-variability Analysis

A two-way analysis of variance (ANOVA) was performed on 10 speakers each of Ibibio and Yoruba languages, to examine the effect of speaker and emotion on F0 and intensity, respectively (see Table 3). The results in Table 4 show that there was significant difference in the mean values of F0 across speakers ($F = 2.38$, $p = 0.011$) and emotions ($F = 597.8$, $p = 0.00$) of Ibibio language. A similar result is also observed for mean values of

Table 3. ANOVA test for intra-variability analysis of emotions and languages on speech features

Response	Factors	Ibibio		Yoruba	
		F	p-value	F	p-value
F0	Emotion	116.55	0.00	597.84	0.00
	Speaker	2.38	0.011	74.36	0.00
	Interaction	8.81	0.00	8.48	0.00
Intensity	Emotion	17.68	0.00	34.97	0.00
	Speaker	6.43	0.00	13.28	0.00
	Interaction	17.17	0.00	11.07	0.00

intensity; speakers ($F = 6.43, p = 0.00$) and emotions ($F = 17.68, p = 0.00$). The mean differences of speakers and emotions are statistically significant in Yoruba language (p < 0.01, p = 0.00). However, the impact on F0 by emotion (F = 597.84) is the highest. Also noticed is the significant interaction between the intra-language effects of speakers and emotion on F0 ($F = 8.81, p = 0.00$; $F = 8.48$; $p = 0.00$) as well as intensity ($F = 11.07, p = .000$; $F = 8.48$; $p = .000$) for Ibibio and Yoruba languages respectively at 95% confidence level. This implies that F0 and intensity of speech in a given language are significantly dependent on the speaker as well as the speaker's emotion.

Table 4. Mean values of intra-lingual variability

Factors	Level	Ibibio		Yoruba	
		F0	Intensity	F0	Intensity
Speaker	S1	163.59	67.46	196.945	56.08
	S2	172.52	67.59	187.77	57.25
	S3	151.14	68.04	231.361	58.60
	S4	172.39	71.71	169.85	60.07
	S5	159.35	67.28	171.20	60.12
	S6	164.72	68.14	162.14	58.47
	S7	170.63	67.22	144.96	60.40
	S8	155.69	68.15	190.67	55.38
	S9	157.46	69.96	197.29	59.02
	S10	157.41	67.85	225.34	62.31
Emotion	Anger	179.10	69.08	211.998	59.85
	Sadness	145.88	67.60	163.507	57.69

In Table 4, a Turkey multi-comparison test shows that the mean F0 and mean intensity values of male speakers of Ibibio vary significantly for anger emotions (mean F0 = 179.10, mean intensity = 69.08) than sadness emotion (mean F0 = 145.88, mean intensity = 67.60). The same inference is drawn for Yoruba language.

3.2 Inter-variability Analysis

Table 5 is the results of a 2-way ANOVA for inter-variability analysis. For F0 response, we found that among speakers of various languages, there exist a highly significant variability in the effect of emotions on F0 (p < 0.01, p = 0.00). Similarly, for speakers expressing various emotions, there is significant variability in the effect of spoken languages on F0 (p < 0.01, p = 0.00). Moreover, there exists a highly significant variability in the effect of interaction between emotions and languages on F0. This implies that the variability in the effect of emotions on F0 changes for different languages. Also, variability in the effect of languages changes for different emotions (p < 0.00, p = 0.00).

Moreover, there exists a highly significant variability in the effect of interaction between emotions and languages on F0, this implies that the variability in the effect of emotions on F0 changes for various languages. Also, variability in the effect of languages changes for various emotions ($p < 0.01$, $p = 0.00$). For intensity response, we found that among speakers of various languages, there exist a highly significant variability in the effect of emotions on intensity ($p < 0.01$, $p = 0.00$). Also, for speakers expressing various emotions, there exist a highly significant variability in the effect of languages on intensity ($p < 0.01$, $p = 0.00$). However, there exist no significant variability in the effect of interaction between emotions and languages. This implies that variability in the effect of emotions on intensity does not change across various languages. Similarly, variability in the effect of languages on intensity does not change significantly for various emotions ($p > 0.05$, $p = 0.214$).

Table 5. ANOVA test for inter-variability analysis of emotions and languages on speech features

Response	Factor	F	p-value
F0	Emotion	417.10	0.00
	Language	159.51	0.00
	Interaction	14.58	0.00
Intensity	Emotion	45.08	0.00
	Language	1241.98	0.00
	Interaction	1.54	0.214

The results from Tukey's simultaneous tests (see Table 6) indicate that the mean level for intensity in Ibibio language (mean = 68.34) is significantly higher than that of Yoruba language (mean = 58.78) while the mean level of F0 for Yoruba language (mean = 187.75) is significantly higher than that of Ibibio language (mean = 162.49). However, for the various languages, anger emotions produce higher mean values of F0 (mean = 195.55) and intensity (mean = 64.47), compared to sadness emotions, which had mean F0 and intensity values of 154.60 and 62.64, respectively.

Table 6. Mean values of inter-lingual variability

Factors	Level	F0	Intensity
Emotion	Anger	195.55	64.47
	Sadness	154.60	62.64
Language	Yoruba	187.75	58.78
	Ibibio	162.49	68.34

3.3 SVM Classification

The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points, where N is the number of features. A binary SVM classifier was used in the classification and prediction of emotions and languages. Binary sequential minimal optimization (SMO) was adopted for the training of the SVM with linear kernel: k(x, y) = <x, y> using F0 and intensity as input features. Results of emotion and language classification are discussed in this section. The emotion SVM model used 0.04 s for model building and execution with 62,600 kernel evaluations (60.695% cached), while the language SVM model evaluated and utilized a total of 48,475 kernels out of which 60.197% were cached in 0.07 s. The model coefficients of the input parameter as shown in Table 7 indicate that for emotions model, intensity has a higher coefficient value than F0, compared to the language model, which yielded a higher coefficient value for F0 than intensity.

Table 7. SVM model coefficients

Input parameter	Emotion SVM model	Language SVM model
F0	−6.1328	3.8187
Intensity	−1.7383	−7.7709
Error term	2.7082	3.7399

The classification performances are evaluated with derivatives of confusion matrix and receivers operating characteristics (ROC) curve including, true positive rate (TPR), Recall, precision and area under the curve (AUC) in addition to kappa statistics, root mean squared error (RMSE) and coverage. The classification confusion matrix of emotion and language is given in Table 8 and Table 9, respectively. The emotion classifier has an overall accuracy of 66.04% (1,849 instances), while the language classifier has an overall accuracy of 79.4% (2,224 instances). In terms of classification errors, the language SVM classifier performed better than emotion classifier (i.e., language RMSE = 0.4536; emotion RMSE = 0.58), despite a higher relative absolute error of 0.68.

Table 8. Confusion matrix for emotion classification

		Predicted	
		Anger	Sadness
Actual emotion	Anger	837	563
	Sadness	388	1012

The model performance evaluation for emotions and languages is documented in Table 10. As shown in Table 10, sensitivities (TPRs) of 60% and 72% are recorded for

Table 9. Confusion matrix for language classification

		Predicted	
		Ibibio	Yoruba
Actual emotion	Ibibio	1121	279
	Yoruba	297	1103

anger emotions and Sadness emotions, respectively, while false positive rates (FPRs) of 28% and 40% are produced by the classifier. The language SVM classifier outperforms the emotion SVM classifier in terms of sensitivity to instances in the respective language classes. Also, 78.80% of instances belonging to the Yoruba class were correctly predicted while the true positive rate (TPR) of instances in Ibibio class is 80.10%, i.e., the highest performance of the two classifiers. Furthermore, the overall weighted AUC for language and emotion prediction performance are 79.40% and 66.00%, respectively.

Table 10. SVM model performance evaluation for emotions and languages

Class label	Sensitivity	FPR	F-Measure	AUC
Anger	0.6000	0.2800	0.6380	0.6600
Sadness	0.7200	0.4000	0.6800	0.6600
Ibibio	0.8010	0.2120	0.7960	0.7330
Yoruba	0.7880	0.1990	0.7930	0.7940
Weighted average (emotion)	0.6600	0.3400	0.6590	0.6600
Weighted average (language)	0.7940	0.2060	0.7940	0.7940

4 Conclusion and Future Works

Classifying emotions using speech features is a relatively new area of research, and has many potential applications. But there exists considerable uncertainty as regards the best algorithm for classifying emotions. This uncertainty however deepens for tone languages – as there are no sufficient resources to empower rigorous research in this area. This paper proposed a generic framework using a state-of-the-art classify – the SVM, to further emotion research for African languages by exploiting basic speech features such as F0 and intensity, and concentrated on the overall effect of syllable units. Two negative emotions (anger and sadness) were used to investigate the intra- and inter-variability of the speech features on emotions and languages, using Ibibio and Yoruba as case study. Results obtained show valid implications useful for advancing speech processing of tone languages. Future directions of this paper include: (i) a study of other emotional types, as well as the creation of large corpus datasets –to enable efficient learning of speech features and prediction; (ii) adoption of a hybrid learning methodology – to improve the robustness of the classifier; (iii) elimination of insignificant feature(s) contribution, hence, reducing computational time; and, (iv) inclusion of more speakers and languages – to improve diversity of the classifier.

References

Batliner, A., Huber, R.: Speaker characteristics and emotion classification. In: Müller, C. (ed.) Speaker Classification I. LNCS (LNAI), vol. 4343, pp. 138–151. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74200-5_7

Brehm, J.W.: The intensity of emotion. Pers. Soc. Psychol. Rev. **3**(1), 2–22 (1999)

Cao, H., Verma, R., Nenkova, A.: Speaker-sensitive emotion recognition via ranking: studies on acted and spontaneous speech. Comput. Speech Lang. **29**(1), 186–202 (2015)

Chakraborty, R., Pandharipande, M., Kopparapu, S.K.: Analyzing Emotion in Spontaneous Speech. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-7674-9_6

Chan, K.W.: Speaker variability in the realization of lexical tones. Int. J. Speech Lang. Law **23**(2), 195–214 (2016)

Kassam, K.S., Mendes, W.B.: The effects of measuring emotion: physiological reactions to emotional situations depend on whether someone is asking. PLoS ONE **8**(6), 1–8 (2013)

Kim, J.: Bimodal emotion recognition using speech and physiological changes. In: Robust Speech Recognition and Understanding, pp. 1–18. InTech (2007)

Sethu, V., Epps, J., Ambikairajah, E.: Speaker variability in speech based emotion models-analysis and normalisation. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7522–7526. IEEE (2013)