ARABIC SPEAKER EMOTION CLASSIFICATION USING RHYTHM METRICS AND NEURAL NETWORKS

Ali Meftah[1], Yousef A. Alotaibi[2], Sid-Ahmed Selouani[3]

[1,2]College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[3]Université de Moncton, 218 bvd. J.-D.-Gauthier, Shippagan, E8S 1P6, Canada
{ameftah, yaalotaibi}@ksu.edu.sa, selouani@umcs.ca

ABSTRACT

In this paper, rhythm metrics are calculated and used to classify five Arabic speech emotions; namely, neutral, sad, happy, surprised, and angry. Eight speakers (four male and four female) simulated the five emotions in their speech by speaking three selected sentences two times each. A human perception test was conducted using nine listeners (male and female). The results of a neural network-based automatic emotion recognition system using rhythm metrics were similar to the human perception test results, although less accurate. Anger was the most recognized speaker emotion and happiness was the least. One of our findings is that the emotions of male speakers are easier to recognize than those of female speakers. In addition, we found that the neural networks and rhythm metrics can be used for speaker emotion recognition using speech signals, but only when the dataset size is large enough.

Index Terms— Emotion, Arabic, corpus, classification

1. INTRODUCTION

A human–machine interaction system is one important application of emotion recognition. Recognizing and classifying speech into different emotions is an area of research that has been developing over the past three decades.

Human emotions are a combination of psychological and physiological factors and can be expressed as a subjective experience, physiological changes in the body (such as tense muscles, a dry mouth, or sweating), and individual behavior (such as facial expressions, fleeing, or hiding). Non-linguistic information plays an important role in human communication and may be observed through facial expressions, the expression of emotions, and even punctuation marks in the cases of video, speech, and written text, respectively [1], [2]. Understanding the emotions present in speech and synthesizing desired emotions in speech according to the intended message are the basic goals of emotional speech processing [2].

Many speech emotion recognition systems have been designed, each focusing on different features, different speech units, and different approaches. In general, a speech emotion recognition system consists of two stages: a front-end processing unit that extracts the appropriate features that represent different speech information (i.e., speaker, speech, and emotion) from the available speech signal, and a classifier that decides the underlying emotion of the speech utterance [2][3].

The main purposes of the corpora were for recognition (the majority), synthesis, or both, while the methodologies used for data collection included simulated [4], [5], elicited [6], [7], and natural collection strategies [8]. Almost all of the existing corpora of emotional speech suffer from issues that make it difficult for automated emotion recognizers to make good use of the data. Some of these issues include the lack of proper contextual information, low recording quality, low quality of simulated emotions, small sample size, and the lack of phonetic transcriptions [3], [9]. Arabic language resources for emotional speech processing are facing lack of dependable corpora. According on our best knowledge, there is no publicly existing Arabic emotional speech corpus that is suitable for digital emotional Arabic speech processing that passes the design and language criteria [3].

It has been reported that rhythm metrics are improving the system accuracy on English language emotion classification [10]. In a past effort for our research team [11], we classified Arabic speaker emotion using rhythm Interval Measures (IM) metrics using graphical scheme in order to find the similarity and dissimilarity between different speakers' emotions. This paper is a continuation to the past effort where our goal is to investigate and automatically classify the effect of rhythm metrics on Arabic speech with respect to speaker emotions. In this current effort we used both IM, Pairwise Variability Index (PVI), and automatic classification using artificial neural network classifier. We targeted this point of research because none of the past researcher covered it for Arabic language depending our survey.

This effort presents a study of rhythm metrics of Arabic speech for the purposes of speaker emotion classification using an MLP classifier.

The rest of the paper is organized as follows: Section 2 presents the Arabic speech rhythm and emotion related work Section 3 explains the speech corpus, the King Saud University Emotions (KSUEmotions) Corpus that is used in this study; Section 4 presents the experimental setup; Section 5 presents the results and discussions; and the conclusion is given in Section 6.

2. RHYTHM METRICS AND ARABIC SPEECH EMOTION RELATED WORK

2.1. Rhythm metrics

Rhythm has been defined as an effect involving the isochronous (i.e., the property of speech to organize itself

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

into pieces of equal or equivalent duration) recurrence of some types of speech units [12].

Rhythm metrics consist of three types [13]. Metrics of the first type are Interval Measures (IMs), ΔV, ΔC, and %V, where ΔV is the standard deviation of vocalic (i.e., the total duration of all adjacent vowels as one reading) intervals, ΔC is the standard deviation of consonantal (i.e., the total duration of all adjacent consonants as one reading) intervals, and %V is the percent of utterance duration composed of vocalic intervals. Metrics of the second type are Pairwise Variability Indices (PVIs), nPVI-V, rPVI-C, nPVI-VC, and rPVI-VC, where nPVI-V is the normalized pairwise variability index for vocalic intervals (the mean of the differences between successive vocalic intervals divided by their sum), rPVI-C is the pairwise variability index for consonantal intervals (the mean of the differences between successive consonantal intervals), nPVI-VC is the normalized pairwise variability index for vocalic plus consonantal intervals (the mean of the differences between successive vocalic plus consonantal intervals divided by their sum), and rPVI-VC is the pairwise variability index for vocalic and consonantal intervals (the mean of the differences between successive vocalic and consonantal intervals).

The final type comprise VarcoV, the standard deviation of vocalic intervals divided by the mean vocalic duration (× 100); VarcoC, the standard deviation of consonantal intervals divided by the mean consonantal duration; and VarcoVC, the standard deviation of vocalic plus consonantal intervals divided by the mean vocalic plus consonantal duration.

2.2. Rhythm and emotion related work

The subject of emotion digital processing via speech signal as an input is a new and active research topic.

Consequently, there are several works that deal with Arabic speech emotion recognition and classification. One of these efforts is the research conducted by Khan et al. [14]. They attempted to classify Arabic sentences into either question or non-question sentences by segmenting them from continuous speech using intensity and duration features as well as by extracting the prosodic features from each sentence. Their approach achieved an accuracy of 75.7%.

Al Dakkak et al. [15] developed an automated tool for emotional Arabic synthesis based on an automatic prosody rough generation model as well as the number of phonemes in a sentence. They claimed that their system's model worked successfully in tests.

In a previous work [16] we extracted the acoustic features of pitch, intensity, formants, and speech rate and used them to classify five Arabic speaker emotions: neutral, sad, happy, surprised, and angry, using three sentences read by four male and four female native Arabic speakers. Furthermore, in [17] we investigated the relationship between speaker gender and rhythm metrics (i.e., %V, ΔC, and ΔV) for three Arabic dialects, namely Modern Standard Arabic, Saudi Arabic, and Levantine. In another related effort, Alotaibi et al. in [11] investigated the relationship between IM rhythm metrics (i.e., %V, ΔC, and ΔV) and speech emotions.

3. KSUEMOTIONS CORPUS

The speech corpus used in this study was recorded using 23 (10 male and 13 female) speakers from Syria, Yemen, and Saudi Arabia in two phases to simulate six emotions, specifically, neutral, sad, happy, surprised, questioning, and anger. Questioning was considered as an emotion because it was incorporated in our original used corpus [18]. Moreover, acted emotional speech was considered in creation of this corpus because we found it is impossible to get real Arabic emotional speech with the targeted dialect. All speakers were 20–30 years old. This corpus was recorded for Modern Standard Arabic (MSA). In Phase 1 of the corpus, 10 male speakers (from Syria, Yemen, and Saudi Arabia) and 10 female speakers (from Syria and Saudi Arabia only) emulated the first five emotion classes by reading 16 MSA sentences. The total size of Phase 1 is about 2 hours and 55 minutes in duration of speech signal. A human perceptual test was performed using nine listeners (acting as independent reviewers) (six males and three females) to evaluate Phase 1 recordings by testing whether normal listeners were able to identify the emotion in the recorded sentence.

According to the results of the human perceptual test, seven male speakers (from Syria, Yemen, and Saudi Arabia) and four female speakers (from Syria, and Saudi Arabia) in Phase 1 were used to produce Phase 2 of this corpus. The criteria of selecting these subsets from the original corpus were based on avoiding corrupted or mispronounce speech files, and also to gain the uniformity among different variables such as speakers' gender. In addition, three new female speakers from Yemen were added to this new phase of the corpus. Also depending on the same criteria, ten sentences were chosen for Phase 2. Each sentence was spoken over two trials. In Phase 2, anger was added to the emotions tested in Phase 1, but the questioning emotion was excluded. Action of adding anger and eliminating questioning is to be more consistent with most of the similar

Table 1. Perceptron test results

		Total No. of files	**Recognized files**	**%**
Male	Phase 1	800	666	83.25
	Phase 2	840	767	91.31
	All	1640	1433	**87.4**
Female	Phase 1	800	643	80.38
	Phase 2	920	809	87.93
	All	1720	1452	**84.42**
Phase 1		1600	1309	81.81
Phase 2		1760	1576	89.54
Phase 1+ Phase 2		3360	2885	85.86

corpora in the field. The size of Phase 2 corpus in speech duration is about 2 hours and 21 minutes.

A human perceptual test was performed again on the new phase of the corpus using the same nine listeners who reviewed Phase 1. The recording process used PRAAT software [19].

Table 1shows the human perceptual test results for the KSUEmotions Corpus for Phases 1 and 2.

4. EXPERIMENT

4.1. Data preparation

Based on the results of the human perceptual test for Phase 2, we selected three sentences, specifically, S09, S11, and S12, as listed in Table 2. In addition, we decided to keep a subset of four male and four female speakers along with the following five emotion classes: neutral, sad, happy, surprised, and angry and this selection is governed by the criteria mentioned above. Each speaker in Phase 2 recorded each sentence two times (i.e., Trials 1 and 2). The total number of selected files in the yielded dataset is 240 (8 speakers × 3 sentences × 5 emotions × 2 trials).

Table 2. Selected sentences

S9:	جُورْجِ بُوش، يُقَدِّمُ وَسَاطَتَهُ لِحَلِّ الْأَزْمَة، بَيْنَ رُوسْيَا وَجُورْجِيَا. ʒuurʒi buuʃ juqaddimu wasaatʔatahu liħallil ʔazmah bajna ruusjaa waʒuurʒijaa
S11:	أَلسَّادَاتْ، بَطَلُ اَلْحَرْبِ وَالسَّلَامْ. ʔassaadaat batʔalul ħarbi wassalaam
S12:	كَامْبِ دِيفِيدْ، إِتِّفَاقِيَّةٌ مُلْزِمَةٌ بَيْنَ فِلَسَطِين وَإِسْرَائِيلْ. kaambi diifiid ʔittifaaqijjatun mulzimatun bajna filasatʔiin waʔisraaʔiil

4.2. Rhythm metrics calculations

To compute the rhythm metrics, it is important to label and segment the speech signals of the corpus being analyzed. Segmentation must identify and separate consonants, vowels, and non-speech portions such as silences and short pauses. This is because rhythm metric computation depends heavily on various comparisons between consonantal and vocalic intervals in the given speech utterance. In our research, labeling was performed manually while segmentation and alignment were performed automatically. To perform this task in a time-efficient manner, we used the HTK toolkit [20] parallel accumulator of the HERest tool for HMM re-estimation in combination with GNU parallelization [21]. The master label file was divided into N parts to enable parallel time-alignment using the HVite tool.

Speech rhythm can be defined as the temporal scattering of linguistic information of a language. It can reflect the closeness of the relationship between spoken speech units, and it plays a central role in the syntax and semantics of many languages. However, current acoustical speech analyzers do not include these aspects when constructing speech processing systems. The schemes illustrated in [22] and [12] were used to calculate rhythm metrics.

4.3. Classifier

All the classification results in this work were carried out using the Waikato Environment for Knowledge Analysis (WEKA) Software [23]. The Multilayer Perceptron (MLP) neural network classifier is based on a back propagation algorithm and is best used for classification purposes. MLP can be monitored and modified during training time. The nodes in this network are all sigmoid activated except when the class is numeric, in which case the output nodes become unthresholded linear units [23].

We semi-exhaustively tested all values of the MLP classifier parameters as follows: We first varied the percentage split of the data training-testing subsets from 66 to 75%. It is important to stress on the issue of randomness and uniformity distribution among speakers and emotion classes of splitting the whole corpus into the two above subsets depending on the considered percentage above. It is important to affirm that the training and testing data subsets are completely disjoint and there is no overlap. Second, we tested the parameters of the neural network hidden layer defined in the WEKA tools {"a," "i," "o," "t"}, where "a" = (attribs + classes) / 2, "i" = attribs, "o" = classes, and "t" = attribs + classes. Third, we varied the learning rate (the amount the weights are updated) from 0.1 to 0.8.

Table 3. Classification performance (%)

	IM metrics		**PVI metrics**		**ALL Rhythm metrics**	
	Best MLP Parameters value	Accuracy (%)	Best MLP Parameters value	Accuracy (%)	Best MLP Parameters value	Accuracy (%)
Male	(L 0.7 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a)	51.72	(L 0.3 -M 0.4 -N 3000 -V 0 -S 0 -E 20 -H t)	50	(L 0.1 -M 0.4 -N 1500 -V 0 -S 0 -E 20 -H i)	60.71
Female	(L 0.5 -M 0.4 -N 1500 -V 0 -S 0 -E 20 -H a)	43.75	(L 0.7 -M 0.0 -N 500 -V 0 -S 0 -E 20 -H a)	54.29	(L 0.3 -M 0.6 -N 2000 -V 0 -S 0 -E 20 -H o)	58.62
Male&Female	(L 0.7 -M 0.4 -N 500 -V 0 -S 0 -E 20 -H a)	55.71	(L 0.3 -M 0.6 -N 1000 -V 0 -S 0 -E 20 -H t)	63.79	(L 0.1 -M 0.0 -N 500 -V 0 -S 0 -E 20 –H t)	72.41

Fourth, the momentum applied to the weights during updating was varied from 0–0.6.

Finally, the number of tested epochs ranged between 500 and 2,000. The other parameters were set to the WEKA defaults.

All rhythm metrics discussed above were used to design five emotion MLP-based recognition systems for male, female, and a combination of male and female speakers. None of the rhythm metrics were sufficiently accurate for emotion recognition when used individually.

5. RESULTS AND DISCUSSION

Table 3 shows the results of the Arabic speech emotion classification performance with the values of the MLP parameters determined by the semi-exhaustive testing. The percentage portion of training subsets that were split from the whole corpus for IM, PVI, and all rhythm metrics experiments, respectively, as follows: 76%, 70%, and 77%; 73%, 71%, and 76%; and 71%, 76%, and 76% for male, female, and combined gender speakers, respectively. That is, the other disparate parts of the corpus were reserved for testing. The way different features are combined is to concatenated features in vectors with larger size before applying them to MLP system.

Generally, classification accuracies using IM or PVI rhythm metrics alone were not high, but when we used all rhythm metrics, the accuracies were improved.

Male speaker emotion MLP-based automatic classification performance is better than that for female speakers, as can be inferred from Table 3. This supports results obtained from the human perceptual test, where the recognition of the emotion of male speakers was superior to that of female speakers for all group features, as listed in Table 1. The best training subset split for this system was between 70– 76%.

The highest value of the emotion recognition accuracy is 72.41% for male and female together using all rhythm metrics. The accuracy improved thanks to the increase in dataset size.

The next step was to study the effect of different features on the performance of the emotion recognition system with fixed MLP parameters. From above results, we selected the best MLP parameters by adopting the ones with the best accuracy in the mixed gender speaker experiment. These parameters are L = 0.7, M = 0.4, N = 500, V = 0, S = 0, E = 20, H = "a," and a split of 76% and 24% for training and testing subsets. Unfortunately, the accuracy worsened with these fixed system parameters.

Table 4 shows the best and worst emotion classification accuracies using fixed and variable MLP parameters for different subsets and set of rhythm features. For male speakers, the most accurately recognized emotion was anger and the least was happy, while the results for female speakers were lower than those for male speakers. In general, anger and sadness were the emotions most easily identified, while happiness was the most challenging to recognize, as inferred from the three experiments. In addition this was confirmed in our human perceptual test, as discussed before.

6. CONCLUSION

In this paper, rhythm metrics were used to recognize five Arabic speaker emotions (neutral, sad, happy, surprised, and angry) using an MLP-type neural network classifier.

The results of the automatic recognition system were similar to the human perceptual test. The emotion of happiness for both male and female speakers was the most challenging and had the least accurate emotion recognition results. In addition, the ability to derive speaker emotion from speech improves for male speakers compared with female speakers. We conclude that ANN classifiers and rhythm metrics can be used for speaker emotion recognition when the dataset to be classified is sufficiently large.

In our planned next research continuation validation and verification of the results, using another public emotional speech corpus probably for English language. This is to test our system and parameters correctness. Also, one of our future objectives is to conduct another research to compare rhythm metrics with signal acoustic metrics such as energy and pitch information.

ACKNOWLEDGMENT

This project was supported by the NSTIP Strategic Technologies Program number (11INF1968-02) in the Kingdom of Saudi Arabia.

Table 4: Best and worst accurate emotion classification (E00 = neutral, E01 = happy, E02 = sad, E03 = surprise, E05 = anger)

		Male		Female		Male and Female	
		Fixed	Variable	Fixed	Variable	Fixed	Variable
Rhythm Metrics	**Classification**	Emotion	Emotion	Emotion	Emotion	Emotion	Emotion
IM	**The Best**	E00	E00	E00	E00	E05	E02
	The Worst	E01	E01	E01	E02	E00 & E01	E01
PVI	**The Best**	E05	E05	E01	E00	E05	E05
	The Worst	E01 & E03	E01	E00 & E02	E01	E01	E01
All	**The Best**	E02	E05	E00	E00	E02	E02
	The Worst	E01	E02	E02	E05	E01	E00 & E01

REFERENCES

[1] A. Hassan, "On automatic emotion classification using acoustic features." University of Southampton, Faculty of Physical and Applied Sciences, p. 204 pp, 2012.

[2] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, Jan. 2012.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[4] Ambrus, D. C. (2000). Collecting and recording of an emotional speech database. Tech. rep., Faculty of Electrical Engineering, Institute of Electronics, Univ. of Maribor

[5] Alpert, M., Pouget, E. R., & Silva, R. R. (2001). Reflections of depression in acoustic measures of the patient's speech. Journal of Affective Disorders, 66, 59–69.

[6] Batliner, A., Hacker, C., Steidl, S., Noth, E., Archy, D. S., Russell, M., & Wong, M. (2004). You stupid tin box children interacting with the Aibo robot: a cross-linguistic emotional speech corpus.In Proc. language resources and evaluation (LREC 04), Lisbon

[7] Pereira, C. (2000). Dimensions of emotional meaning in speech. In Proc. ISCA workshop on speech and emotion, Belfast, Northern Ireland, 2000 (pp. 25–28).

[8] Cowie, R., & Douglas-Cowie, E. (1996). Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In Fourth international conference on spoken language processing ICSLP96, Philadelphia, PA, USA, October 1996 (pp. 1989–1992).

[9] Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. International Journal of Speech Technology, 15(2), 99–117.

[10] Ringeval et al. (**2012**) Novel metrics of speech rhythm for the assessment of emotion.

[11] Y. A. Alotaibi, A. H. Meftah, and S.-A. Selouani, "Investigation of emotion classification using speech rhythm metrics," in *Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE), 2013 IEEE*, 2013, pp. 204–209.

[12] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Pap. Lab. Phonol.*, vol. 7, no. 515–546, 2002.

[13] J. M. Liss, L. White, S. L. Mattys, K. Lansford, A. J. Lotto, S. M. Spitzer, and J. N. Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *J. Speech, Lang. Hear. Res.*, vol. 52, no. 5, pp. 1334–1352, 2009.

[14] O. Khan, W. G. Al-Khatib, and C. Lahouari, "detection of questions in Arabic audio monologues using prosodic features," in *Multimedia, 2007. ISM 2007. Ninth IEEE International Symposium on*, 2007, pp. 29–36.

[15] O. Al-Dakkak, N. Ghneim, M. Abou Zliekha, and S. Al-Moubayed, "Prosodic Feature Introduction and Emotion Incorporation in an Arabic TTS," in *Information and Communication Technologies, 2006. ICTTA'06. 2nd*, 2006, vol. 1, pp. 1317–1322.

[16] A. H. Meftah, S.-A. Selouani, and Y. A. Alotaibi, "Preliminary Arabic Speech Emotion Classification," in *IEEE International Symposium on Signal Processing and Information Technology*, 2014.

[17] A. H. Meftah, S.-A. Selouani, and Y. A. Alotaibi, "Investigating speaker gender using rhythm metrics in Arabic dialects," in *8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA), 2013*, 2013, pp. 347–350.

[18] KTD Corpus, KACST Unpublished Technical Report.

[19] D. Boersma, Paul & Weenink, "Praat: doing phonetics by computer." 2014.

[20] S. J. Young and S. Young, *The HTK hidden Markov model toolkit: Design and philosophy*. Citeseer, 1993.

[21] O. Tange, "Gnu parallel-the command-line power tool," *USENIX Mag.*, vol. 36, no. 1, pp. 42–47, 2011.

[22] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.