Degree of anger in speech signals

Gonzalo Sosa[1]

Ingeniería de Sonido, Universidad Nacional de Tres de Febrero.
sosa40730@estudiantes.untref.edu.ar[1]

Abstract - *In this research, the perception of the degree of anger in the voice is studied. The study focuses on the objective parameter of the variation of the fundamental frequency on the Mel scale. A subjective test is carried out on 37 subjects in which they must choose the voice that they perceive with the greatest anger from among female voices of the Greek and German languages. Then a validation process is carried out in which 33 consistent tests are obtained. With the 33 validated tests, it is shown that the stimulus and the language are influential in the perception of the degree of anger. Furthermore, it is shown that there is a correlation between the objective and subjective parameters.*

1. INTRODUCTION

Speech emotion recognition (SER) is the task of extracting the emotions of a speaker from his or her speech signal. Emotions are forms of expression for humans and are used in everyday speech. Emotion recognition is useful in many tasks such as detecting the emotion of a telephone operator in a call center [1] or detecting the mental state of a driver [2]. In a speech signal, there are two important factors to be considered, the linguistic and the non-linguistic information. This work focus on the non-linguistic information.

Many of the previous studies use Machine Learning techniques to classify a speech signal emotion. The usual workflow consists of:

1. Select a voice database with an emotion already assigned (labelled data)
2. Extract acoustic features from each signal
3. Train classification models using the extracted acoustic features (inputs) and labels (outputs)
4. Use the best model to predict emotions from new data

The studies differ between them in aspects such as the selection of acoustic features to train the models. One of the first works on this topic is the one presented by Chen et al. [3]. A classification model was used to classify six emotions (Sadness, Anger, Surprise, Fear, Happiness and Dislike). They used features such as Energy, Zero Crossing Rate, Pitch, Spectrum Centroid, Correlation Density, Fractal Dimension and MFCCs. The results showed a precision score of 0.74 for classifying anger but low scores for the other emotions. The conclusion was that they need features that better correlated with the other emotions.

A recent work, made by Móstoles et al. [4] outperformed previous ones with results of 0.94 (accuracy score). The emotions were the same as Chen's but they also added two more (Neutral and Calm). Features were Zero Crossing Rate, Spectral Flux, Short Time Energy Deviation, Short-Term Spectral Power Density, Band Entropy and Band Periodicity. They implemented a four-step model. The first step focus on removing the silence, the second step takes care of transforming the power spectrum to a mel scale, in the third step they use a Convolutional Neural Network (CNN) to obtain a vector of features and in the last step they use a K -Nearest Neighbors (KNN) classifier to assign an emotion. Despite the good results they conclude that an evaluation of the model with another dataset has to be done to ensure the model was not doing overfitting to the training data.

In another study made by Patel et al. [5] the authors used autoencoders to reduce the dimensionality of the input features of the classification models. They found out that the introduction of autoencoders can improve the classification accuracy of the emotion. They concluded that the choice and application of dimensionality reduction of audio features impacts the results that are achieved and by working on this aspect of the general speech emotion recognition model it may be possible to make improvements in the future.

The data usually used to train the models is recorded to intentionally recreate an emotion. Sainz et al. [6] did subjective test to validate a dataset recorded by actors for speech analysis tasks. The conclusion was that the database is a valid resource for the research and development purposes it was designed for.

Another study, published in 2006 by Navas et al. [7] performed subjective and objective analysis. Subjectively they check the ability of human to identify emotions and objectively they analyzed a set of acoustic prosodic parameters. Prosodic features included were the fun-

damental frequency and energy. They made two data corpus. For one of them an actress recorded emotions reading neutral semantic texts. For the other one the same actress read emotional semantic texts that match with the emotion. Then in the subjective test users had to select one of the proposed emotions. The data was recorded by an actress that spoke in Basque. They used two groups of participants, Basque language speakers and non-Basque language speakers. As expected, the results showed better recognition scores for Basque language speakers. Both groups got better scores with the second corpus. The data was recorded in an over exaggerated way to recreate the emotion. They concluded that the results may be influenced by the data and that it would be interesting to have a stimuli with multi-speaker and with real recordings.

An aspect that most of subjective tests failed to achieve is independence of linguistic content. A subject might respond that a speaker is angry only because of the content of what he is saying despite the acoustic characteristics of the signal. One of the objectives of this study is to know the acoustic characteristics that best correlate with an emotion regardless of the linguistic information. A proposal to overcome this problem is to test with data using languages that the person does not know as did the study mentioned above.

This study aims to measure the degree of emotion in the speech signal. Previous studies only focus on classifying an emotion from speech but no one reported the degree of that emotion. In this work a subjective test is done to evaluate the degree of emotion in a speech signal. Also a deep analysis on the correlation of acoustic features with the emotion is done.

2. METHODS

2.1 Stimuli selection and Test Stimuli

As stated in the introduction, the purpose of this investigation is to evaluate the degree of an emotion. Therefore, the selection of audio files from public databases was made. The databases used were the *Acted Emotional Speech Dynamic Database* (AESDD) [8] and the *Berlin Database of Emotional Speech* (Emo-DB) [9]. Both are publically available speech emotion recognition databases. They contain utterances of acted emotional speech in the Greek and German language respectively. In each databases the audios are classified by an emotion (e.g., AESDD contains audios of Anger, Disgust, Fear, Happiness and Sadness).

For each audio, temporal windows of 25 ms [10] were taken. For each window fundamental Frequency (F0) was obtained using Librosa [11]. Then the parameter was transform from Hertz to a Mel scale which is a perceptual scale of pitches judged by listeners to be equal in distance from one another [12]. Then mean, median and standard deviation were calculated using the values of the windows.

To achieve a subjective test with the appropriate time duration a decision about selecting a specific emotion was made. Because of a previous study [3] already mentioned in the introduction, anger was chosen as the emotion to test.

Then the audios for the subjective test were chosen. AESDD (Greek language) contains groups of three audios where in each group a sentence is recorded with three different intonations and time duration. From the global parameters the standard deviation of the Mel fundamental frequency (F0 variation) was chosen to represent the variation of the tonality. A group of three audio was chosen with the consideration of having approximately equal intervals of F0 variation. Also three audios were selected from the German database with the same criteria. The only difference was that in this database the audios corresponds to different sentences. Figure 1 shows the value of F0 variation.

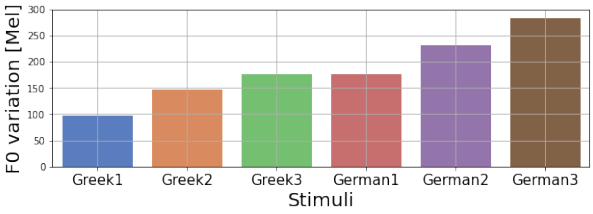

Figure 1: *F0 variation.*

Table 1 shows the semantic content of the spoken sentences translated to english. Figure 7 shows the selected audio files in the time domain.

Table 1: *Spoken sentences translated to English*

Audio Files	Sentences in English
Greek1	*We were paid just to do nothing, just only to be there.*
Greek2	*We were paid just to do nothing, just only to be there.*
Greek3	*We were paid just to do nothing, just only to be there.*
German1	*The tablecloth is lying on the frigde.*
German2	*She will hand it in on Wednesday.*
German3	*Tonight I could tell him.*

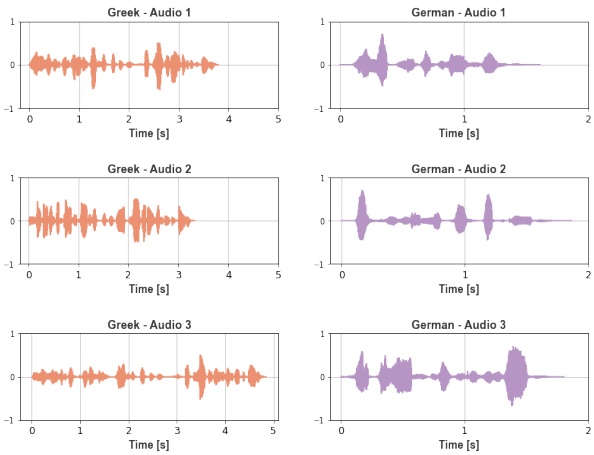

Figure 2: *Selected audio files for the test.*

Table 2 shows the mean, standard deviation and median of the Fundamental frequency for each audio. The audios were ordered in ascending order with respect to the F0 variation. It is interesting to note that for German audios the median and the mean have a considerable difference. This shows that the audios in this language have outliers in this parameter. This behavior does not occur for Greek audios as the mean and median have similar values.

Another important aspect is that between Greek3 and German1 the difference in fundamental frequency variation is minimal thus is something to pay attention to.

Table 2: *Mel Fundamental Frequency (F0) parameters*

Audio Files	Mean F0	Std F0	Median F0
Greek1	463.20	98.90	471.71
Greek2	384.63	148.08	358.21
Greek3	486.40	177.54	468.00
German1	891.55	178.31	421.83
German2	967.54	233.57	449.26
German3	1008.27	284.71	462.80

2.2 Test Environment

An online test was carried out due to the health emergency generated due to
COVID-19 virus. The conditions for the test were to be in a quite place, to use headphones and to have a computer with internet connection. To make sure that the levels of the signals are appropriate for the the listening level was adjusted according to the sensation level (SL). An audio file from the greek database which is not used in the subjective test was used to calibrate the system. The signal was processed to have a difference of 40 dB to the stimulus signals. Then, before the test, the person had to adjust the volume of their device so that the calibration signal is just audible (threshold). If the subject performed the procedure correctly, stimuli would be at 40 dBSL which is an appropriate level for the test.

2.3 Subjective Test

The subjective test was performed using a paired comparison test. Thirty-seven listeners were selected randomly with the condition that they must be native Spanish speakers from Argentina. This condition is because the perception of emotion may change depending on the culture. Also, the listeners were not supposed to understand the Greek and German language.

Participants were asked to choose between two stimuli based on which one they perceived as more angry. The exact question translated to english was:

In which of the audios do you think the person showed a greater degree of anger?

The original question was asked in spanish. The question was:

¿En cuál de los audios cree que la persona mostró un mayor grado de enojo?

As mentioned in Section 2.1 there are six audio files selected. Despite having three audios from each language, the idea was to compare each stimulus againts the rest to be able to make a particular and global analysis. Therefore, there were fifteen comparison pairs. The duration of each audio is approximately 2 s for the Greek database and 5 s for the German database, so the total duration of audio, taking into account an interval of 1 s between each signal, was around two minutes. Stimuli was presented in a random order.

Also the test includes the calibration part, so taking into account this and the time to answer the total duration of the test was around five minutes. It should be clarified that subjects were allowed to repeat the audios again as many times as they wanted.

3. RESULTS

3.1 Validation

First, a validation process was carried out with the results obtained from the subjective test. For this, a consistency test and an agreement test were performed. For the consistency test, the circular error of each of the participants was calculated. An accepted circular error of 0.4 was established, for which the subjective tests of the participants who had a circular error greater than 0.4 were discarded. By this process 33 out of 37 participants remained. Then, with the 33 responses, it was found that there was a significant agreement $(180 > \chi^2(0.05, 16))$.

3.2 Test Scores

3.2.1 Global Scale Values

The global test scores are the scaled values obtained from the Thurstone method. Figure 3 shows the results of the scaled value for each stimulus. In addition, Figure 4 shows a scatterplot between the scale values and F0 variation which was the objective parameter. The hypothesis of the test was that the perception of anger increases with F0 variation. It can be seen that the result is similar to the expected one.

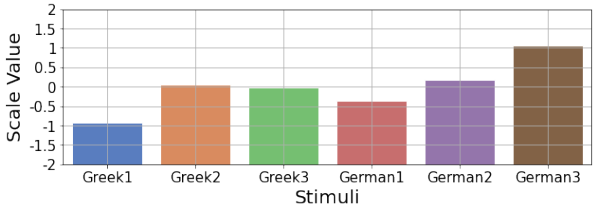

Figure 3: *Global Scale values obtained by Thurstone method*

The biggest difference is between Greek2 and Greek3 comparisons. Figure 5 shows the results that presented the greatest disparity in their responses. This behavior also occurred between the comparison between German2-Greek2 and German2-Greek3. Later, the influence of the language factor on the perception of anger will be studied.

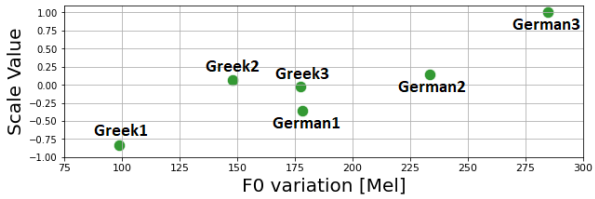

Figure 4: *Fundamental frequency variation Vs Scale Values*

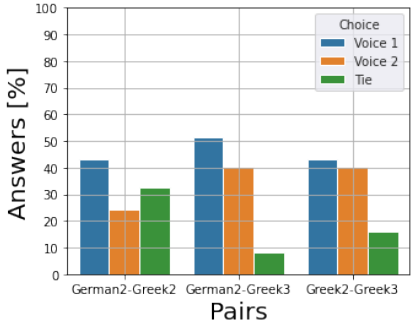

Figure 5: *Pairs with disparity answers*

3.2.2 Individual Scale Values

Figure 6 shows a boxplot of the scale values of each participant. The presence of outliers can be observed in Greek1, Greek2 and German3, also it can be seen that Greek3 shows a considerable deviation.

For each of the stimuli, a Shapiro Wilk [13] test was performed to check the normality of the distribution. The results are shown in Table 3. Normality could be confirmed for Greek2, Greek3 and German1 with a P value less than 0.05. For the rest of the stimuli the P value was greater than 0.05 thus normality could not be confirmed.

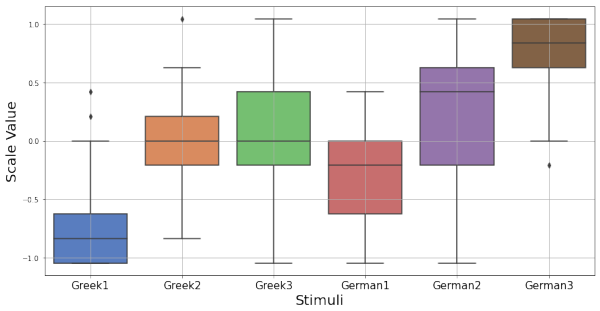

Figure 6: *Boxplot of individual Scale Values*

Table 3: *Shapiro Wilk test results for individual Scale Values (Sig. for P value less than 0.05)*

Stimuli	P_value
Greek1	No Sig.
Greek2	Sig.
Greek3	Sig.
German1	Sig.
German2	No Sig.
German3	No Sig.

Then a Levene test [14] was performed to check the homogeneity of the variances. With a P value less than 0.05, homogeneity could not be confirmed.

3.3 Analysis of variance

3.3.1 Factor: Stimuli

To study whether the variation of the stimulus is influencing the perception of anger, the Kruskal Wallis [15] test is used. This test is chosen over an ANOVA due to the results of normality and homogeneity. The results show that the variation of the stimulus is influential ($\chi^2 = 100.71, df = 5, p < 0.01$).
As the factor is made up of six variables, it is necessary to carry out a post-hoc analysis to find out which variables are being influential.

Dunnet's [16] test is used to carry out the analysis. The results are shown in Table 4. As can be seen, the pairs Greek2-Greek3, German2-Greek2 and German2-Greek3 are the ones that did not have significant results in the perception of anger. These pairs are also the ones that had the most discrepancy responses.

Table 4: *Results of Dunnet's test (Sig. for P value less than 0.05)*

	Greek1	Greek2	Greek3	German1	German2	German3
Greek1		Sig.	Sig.	Sig.	Sig.	Sig.
Greek2	Sig.		No Sig.	Sig.	No Sig.	Sig.
Greek3	Sig.	No Sig.		Sig.	No Sig.	Sig.
German1	Sig.	Sig.	Sig.		Sig.	Sig.
German2	Sig.	No Sig.	No Sig.	Sig.		Sig.
German3	Sig.	Sig.	Sig.	Sig.	Sig.	

3.3.2 Factor: Language

The influence of language on the perception of anger was also analyzed. For this, the results were grouped into two groups, Greek and German. The normality and homogeneity of the variance was re-studied. As in the previous case, the results of the tests did not confirm normality or homogeneity, so the analysis was carried out through the kruskal wallis test. With a P value less than 0.01, it was found that there is an influence of the language factor in the perception of anger ($\chi^2 = 21.58, df = 1, p < 0.01$).

3.4 Correlation and Linear Regression

The Spearman correlation between the objective parameter F0 variation and the scale values was studied. The results are shown in Table 5. As can be seen, there is a correlation between the objective parameter studied and the perception of anger.

Table 5: *Spearman correlation results*

Spearman	Correlation	P Value
F0 Variation Vs Scale Values	0.91	0.047

Since it was possible to verify that there is a correlation, a linear regression model was made between the variables. Table 6 shows the results of the model. It can

be seen that the value of R squared is close to one and that means that the model fits well. The model could be used to make predictions of the degree of anger from the variation of the fundamental frequency.

Table 6: *Linear Regression model*

Linear Regression	R2	Coefficient	Intercept
F0 Variation Vs Scale Values	0.82	0.008	-1.59

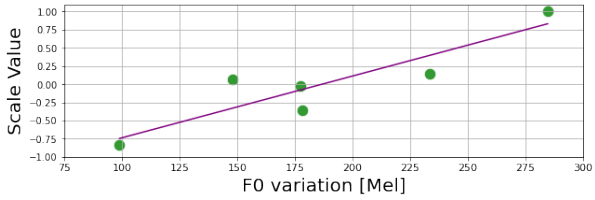

Figure 7: *S*

4. DISCUSSION

The test results showed that there is a correlation between the variation of the fundamental frequency and the perception of the degree of anger. However, not in all cases it was true that the higher the fundamental frequency, the greater the perception of anger. To know the relationship in more depth, the study should be carried out incorporating more languages. In addition, it would be interesting to use synthesized voices in such a way as to be able to have greater control of the objective parameters.

The variation of the fundamental frequency used in this work represents in a global way how much the tone of the voice varied in the audio, but it would be interesting in future research to give importance to temporality because the moment in which these variations occur could change the perception of anger.

Although the variation of the fundamental frequency turned out to be a strong parameter in the perception of anger, it would be useful to carry out the study with more parameters like those mentioned in the introduction.

Furthermore, if there had been a greater number of participants, normality might have been confirmed for the stimuli that did not do so.

5. CONCLUSIONS

The study of the degree of anger in the voice can be useful in various applications. Throughout this work, the focus was placed on the objective parameter of the fundamental frequency variation and its influence on the perception of the degree of anger. Only female voices from the Greek and German languages were used. A subjective test was carried out in which 37 subjects participated. Of the 37 tests, 33 were used that were consistent with a CER < 0.4.

Then it was found that the different stimuli and the language were influential in the perception of anger. A high correlation was also obtained between the objective and subjective parameters. For future research it would be interesting to incorporate more languages and male voices. By doing a deeper investigation and evaluating more subjects, a model could be made that predicts the degree of anger in a voice. This would be useful for various applications, such as telephone calls where the emotion of the interlocutor could be important.

REFERENCES

[1] D M Litvinov. Speech analytics architecture for banking contact centers. Moscow, Russia, 2021. International Scientific and Practical Conference.

[2] Norhaslinda Kamaruddin and Abdul Wahab. Driver behavior analysis through speech emotion understanding. In *2010 IEEE Intelligent vehicles symposium*, pages 238–243, La Jolla, CA, USA, 2010. IEEE.

[3] Lijiang Chen, Xia Mao, Yuli Xue, and Lee Lung Cheng. Speech emotion recognition: Features and classification models. *Digital signal processing*, 22:1154–1160, 2012.

[4] Roberto Móstoles, David Griol, Zoraida Callejas, and Fernando Fernández-Martínez. A proposal for emotion recognition using speech features, transfer learning and convolutional neural networks. 2021, Valladolid, Spain.

[5] Nivedita Patel, Shireen Patel, and Sapan H Mankad. Impact of autoencoder based compact representation on emotion detection from audio. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–19, 2021.

[6] Iñaki Sainz, Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, Jon Sanchez, Iker Luengo, Igor Odriozola, and Imanol Madariaga. Subjective evaluation of an emotional speech database for Basque. In *LREC*, 2008.

[7] Eva Navas, Inma Hernáez, and Iker Luengo. An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *IEEE transactions on audio, speech, and language processing*, 14:1117–1127, 2006.

[8] Nikolaos Vryzas, Rigas Kotsakis, Aikaterini Liatsou, Charalampos A Dimoulas, and George Kalliris. Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society*, 66:457–467, 2018, Aristotle University of Thessaloniki, Thessaloniki, Greece.

[9] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A

database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005.

[10] Vlado Keselj. Speech and language processing daniel jurafsky and james h. martin (stanford university and university of colorado at boulder) pearson prentice hall, 2009, xxxi+ 988 pp; hardbound, isbn 978-0-13-187321-6, 115.00, 2009.

[11] Brian McFee, Vincent Lostanlen, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana, Kyungyun Lee, Oriol Nieto, Jack Mason, Dan Ellis, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, Keunwoo Choi, viktorandreevichmorozov, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, nullmightybofo, Darío Hereñú, Fabian-Robert Stöter, Pius Friesch, Adam Weiss, Matt Vollrath, and Taewoon Kim. librosa/librosa: 0.8.0, July 2020.

[12] Douglas O'Shaughnessy. Speech communication, human and machine addison wesley. *Reading MA*, 1987.

[13] S Shaphiro and M Wilk. An analysis of variance test for normality. *Biometrika*, 52(3):591–611, 1965.

[14] Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pages 279–292, 1961.

[15] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[16] Charles W Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.