Indian Institute of Technology,
Kanpur

Department of Electrical Engineering

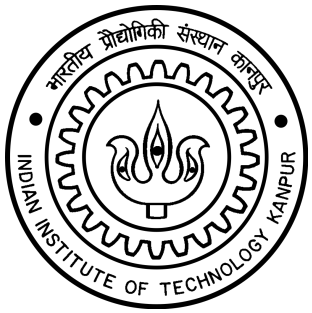
A report submitted in partial fulfillment for the course Undergraduate Project-I (EE391A)

Speech-based synchrony measurement in dyadic interaction

Author:
Karttikeya Mangalam
14311

Supervisor:
Prof. Tanaya Guha

April 15, 2017



Introduction

Synchrony detection and measurement is a recently active field of research in human centered Signal Processing. Of its many objectives, a major one is to develop a measure of "Synchrony" present between two or more individuals during their are interaction.
For our means, synchrony is understood as a notion of mutual interest that the persons concerned are displaying in their interaction with others. For an example : Consider an adept story-teller narrating an interesting anecdote to an interesting audience over dinner. Clearly, by virtue of his voice modulations and expressions he has maintained a stron grip over his audience and they are responding to the stimulus presented by the speaker. For another setting, consider two men quarreling with each other over a road accident; naturally, one would respond with great vigor and energy to the allegations put forward by the other. In both of the above examples, the interaction is said to have high synchrony because the humans concerned are deeply involved in their respective interactions.
On the other hand, consider a post lunch high-school class with an extremely boring topic at hand in a primary school. As one can naturally imagine, the students would be quite disinterested in the class proceedings and consequently, the interaction would have low synchrony. Or, consider a person suffering from Attention Deficit Disorder (ADD) in a formal interaction at a meeting not being able to concentrate at the matter at hand. This again would be a low synchorny situation. On a rough scale, synchrony is the measure of "attention" the speakers give to the other persons involved in the situation.
Synchrony detection is a fairly recent field and consequently does not have many standard benchmarks and datasets to compare results with. On the onset, synchrony itself while being formally defined in psychology, doesn't have a definite measure to be assigned in the present scenario.Concurrently, We have used Discrete Time Warping (DTW) as a measure for the same.
Also, no datasets are available with synchrony annotations and all results have to be empirically verified with human effort. However, this is a very exciting endeavour with rich applications to many social contexts and situations.

Dataset description

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an acted, multimodal and multi-speaker database, recently collected at SAIL lab at USC. It contains approximately 12 hours of audiovisual data, including video, speech, motion capture of face, text transcriptions. It consists of dyadic sessions where actors perform improvisations or scripted scenarios, specifically selected to elicit emotional expressions. IEMOCAP database is annotated by multiple annotators into categorical labels, such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation and dominance.[1]
For our purposes, we have used over 12 hours of speech signals divided into five sessions and with over 2000 dialagoues in each session. All this data is processed using openEAR, a Munich Open-Source Emotion and Affect Recognition Toolkit developed at the Technische Universität München (TUM). It provides efficient (audio) feature extraction algorithms implemented in C++, classfiers, and pre-trained models on well-known emotion databases.

Methodology Used

Feature extraction and pooling

We have started by processing the entire dataset to extract features from each dialogue using openEAR. Three different set of features were extracted for experimental purpose described as follows:

- The feature set used for the Interspeech 2009 Emotion Challenge: consisting of 384 features containing the following low-level descriptors (LLD): Intensity, Loudness, 12 MFCC, Pitch (F0), Probability of voicing, F0 envelope, 8 LSF (Line Spectral Frequencies), Zero-Crossing Rate. Delta regression coefficients are computed from these LLD, and the following functionals are applied to the LLD and the delta coefficients: Max./Min. value and respective relative position within input, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1-3, and 3 inter-quartile ranges.

- A baseline set of 988 features used for Emotion recognition.

Table 1: Data distribution in different classes

Emotion	#Examples in IEMOCAP	% of Total Example	Emotion Classified as
Frustration	2901	29.3	Negative
Anger	1199	12.11	Negative
Excited	1934	19.54	Positive
Fear	101	1.02	Negative
Happiness	652	6.58	Positive
Sadness	1249	12.62	Negative
Neutral State	1720	17.38	Neutral
Surprise	0100	1.02	Positive
Others	26	0.20	Positive

- A massive all-permutation combined feature set of 6552 features extracted from each audio file. However, this turned out to contain many features having similar information content and was later abandoned in favor of faster testing and training times.

After such features were extracted from every audio dialogue, they were pooled and labelled using the emotion annotations from the IEMOCAP database; finally creating a dataset of nearly 10,000 examples each with feature size either 384 or 988 features.

Emotion Classification and Learning model

The original dataset contained a total of 9 different emotion categories with distribution as indicate in Table 1. Different classes were pooled into a 3-class dataset as indicated in the table with classes : Negative (55%), Positive (23%) and Neutral (22%).
A Support Vector Machine (SVM) classifier was trained on this using [2] using a 80-20 split between the training and testing data. The test emaxple were chosen uniformly from the distribution and no examples were excluded during the training/testing phase. After, hand tuning the parameters for various different kernels, a highest of 62.6% accuracy was achieved in the 3-way classification using $C = 0.0056$ and $\gamma = 180$ using RBF Kernel and class-weights 1.0, 1.95 and 2.0 for Negative, positive and the neutral classes

respectively. For this setting, the traning accuracy is 71.1 % yielding a overall correct prediction percentage of 69.4 % over the entire dataset.

Predicted emotion signals

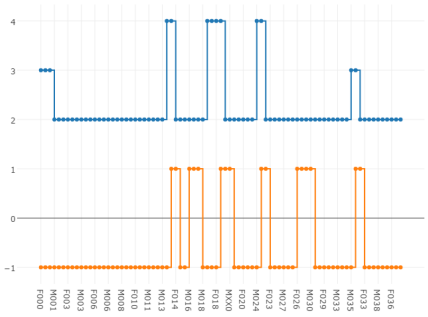
The SVM classifier trained takes as input a feature vector for each dialogue and outputs the corresponding predicted emotion label (positive/negative/neutral state). These predicted emotions are put together in a time-series fashion in the order in which the dialogues occurred in time. Since, each interaction is dyadic in nature, so, emotion signal for both the speakers have been plotted separately and while the first speaker is speaking, the emotion for the second speaker is assumed to remain same until he/her himself/herself speaks. Kindly see signal example I- IV for some of the predicted emotion signals and the true emotion signals.
The red signal is the emosignal for the female speaker and the blue signal corresponds to the male speaker.Also, for the female speaker a value of 0 represents a neutral state; -1 represents the negative state and +1 the corresponding positive state. For the male speaker, +3,+2 and +4 repsresent the neutral, negative and positive states respectively. These plots are generated using MatPlotlib library in scikit-learn environment in Python.[3]

Measurement of Synchrony

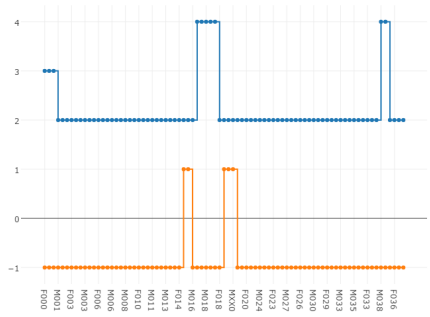
Herein, lies the cornerstone idea for the project : We can developed a reliable model for synchrony in a dyadic setting by a measure of, in a crude sense, causality between the speaker emotion signals. The intuition behind this is the natural behavior that causes a change in the emotional state in respond to stimuli from the speaker in a high synchrony situation and a relative independence between the emotional states for a low synchrony situation.

Clearly, for our purposes the actual change, in the emotional state of the listener in response to a change in speaker is immaterial. The only important factor is the existence of such a change. For measuring this, we are currently employing a Dynamic Time Warping (DTW) model that operates on either the true (in case it's known) or the the predicted signal to output the synchorny measure. Dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in speed and polarity. Temporarily, we are operating the DTW algorithm

on the original emosignals because the model accuracy is not very good and that would consequently effect the DTW response which in any case has to be validated by human intuition.

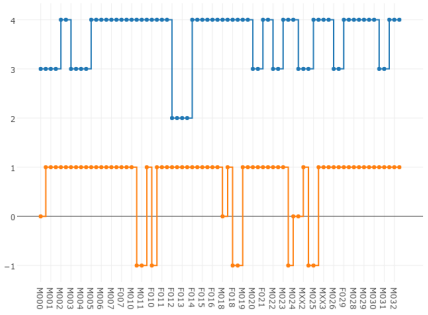


(a) Predicted label signal

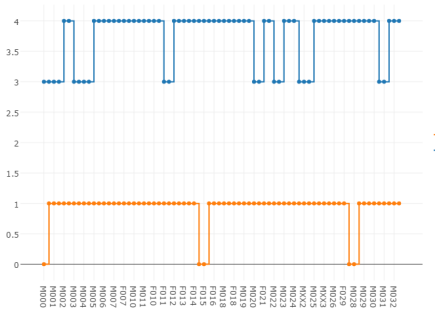


True label signal

Example I

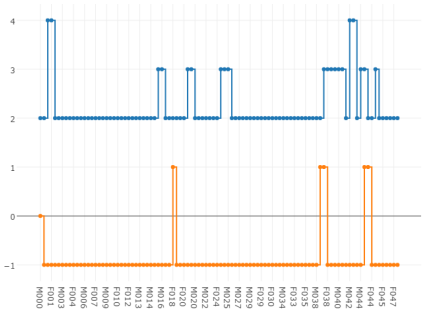


(a) Predicted label signal

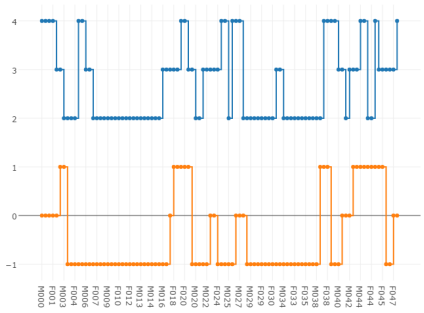


(b) True label signal

Example II

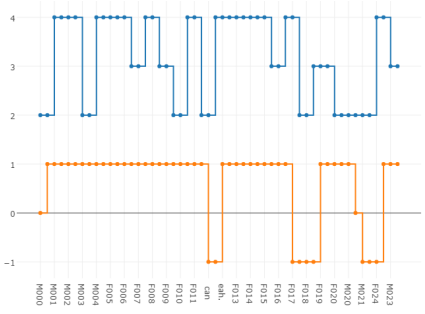


(a) Predicted label signal

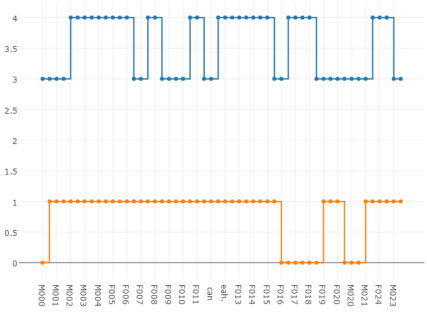


(b) True label signal

Example III



(a) Predicted label signal



(b) True label signal

Example IV

Results & Future Work

The key conclusions from the project can be summarized as:

- A testing accuracy of 62.6 % is achieved in 3-way classification over the whole dataset using Support Vector Machines as classifiers.
- With SVM tuned at the optimality, the kernel chosen is RBF with parameters $\gamma = 180$ and $C = 0.0056$. This has 71.1% accuracy over training data and a overall 69.4 % accuracy over the whole dataset.
- Dynamic time warping was implemented as a measure for measuring causality between the emotion signals. The results however, do not match human intuition and a need for better measure is felt. For examples shown, the (normalized) DTW cost is in 0.2-0.4 with 1 being the maximum possible for two signals allowed to take value either -1, 0 or 1 and with same signal length.
- DTW cost differs by less than 10% between the predicted emo-signals and the annotated ones in 85 % of the given data. This however, is unreliable because of high variance and little correlation with human intuition.

The approach adopted represents a new perspective on measuring synchorny which previously has been mostly unchartered. The accuracy for emotion signals is expected to boost significantly by using appropriately tuned neural networks which will reduce the variance between the DTW cost from true signals and predicted signals.
A major improvement is needed in using a more robust and closer to actual experience causality measure such as Granger causality [4]. A better measure is expected to provide a good measure of the actual dependence between the signals instead of a 'distance' measure between them.

Bibliography

[1] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.

[2] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[3] John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.55

[4] C. W. J. Granger, Investigating Causal Relations by Econometric Models and Cross-spectral Methods, Econometrica, Vol. 37, No. 3. (Aug., 1969), pp. 424-438