A Survey on Paralinguistics in Tamil Speech Processing

Anosha Ignatius and Uthayasanker Thayasivam
University of Moratuwa
Sri Lanka
anoshai@uom.lk, rtuthaya@cse.mrt.ac.lk

Abstract

Speech carries not only the semantic content but also the paralinguistic information which captures the speaking style. Speaker traits and emotional states affect how words are being spoken. The research on paralinguistic information is an emerging field in speech and language processing and it has many potential applications including speech recognition, speaker identification and verification, emotion recognition and accent recognition. Among them, there is a significant interest in emotion recognition from speech. A detailed study of paralinguistic information present in speech signal and an overview of research work related to speech emotion for Tamil Language is presented in this paper.

1 Introduction

The field of Paralinguistics deals with how something is being spoken and is limited to non-verbal aspects of speech. It covers the manner of speaking and information characterizing speaker traits and states. Paralinguistic content present in the speech signal provides information which is mainly grouped into short-term traits and long-term traits. Short-term traits include speaking style, voice quality, and emotional states while long-term traits include biological trait primitives such as height, weight, age and gender, ethnicity, culture, and personality. Health state, intoxication, mood, and sleepiness are categorized as medium term traits between short term and permanent traits. (Schuller and Batliner, 2014)

The research on paralinguistics is beneficial in a wide range of speech processing applications. Performance of automatic speech recognition (ASR) is affected by variability in the speaking style due to speaker traits emotions. Analyzing the paralinguistic information in the speech help to compensate for the speaker variability through normalization techniques (Anosha and Uthayasanker, 2020). In addition to this, several paralinguistic tasks including speaker recognition, accent recognition, and emotion recognition have been investigated over the past years. Interest in paralinguistics has significantly grown in the past years and Interspeech conference hosts a computational paralinguistics challenge ComParE since 2009 with different sets of tasks each year. The most popular paralinguistic task is speech emotion recognition. Emotion detection can help in making the human computer interface adapt to user's emotional condition, thus improving the user satisfaction. Study of paralinguistics can also be applied in diagnosing and monitoring the disease progression in diseases like neurodegenerative disorders which show speech impairment as one of the early signs.

Selecting appropriate feature extraction techniques is important for discriminating between classes in paralinguistic tasks. Then most common features that capture paralinguistic information are low-level descriptors (LLD) (Schuller et al., 2013) which include mel frequency cepstral coefficients (MFCC), energy, pitch frequency, loudness, zero crossing rate, harmonicity, jitter, shimmer etc. Recent research works use deep neural networks (DNNs) to learn high-level acoustic features from utterance-level LLD or directly from raw speech signal. DNNs have shown significant performance in a variety of applications ranging from speech recognition to speaker identification and verification.

This paper presents the background of paralinguistics and focuses on the research work related to speech based emotion recognition for Tamil language. Tamil is a Dravidian language natively spoken by South Asia's Tamil people. Tamil is the official language of two sovereign states, Singapore and Sri Lanka, as well as the Indian province of Tamil Nadu and Puducherry (Chakravarthi et al.,

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 94–99

2018, 2019; Chakravarthi, 2020). With a history stretching back to 600 BCE, the Tamil language is one of the world's longest-surviving classical languages. Poetry dominates Tamil literature, especially Sangam literature, which consists of poems written between 600 BCE and 300 CE. The Tamil language accounts for more than 55 percent of the epigraphical inscriptions discovered by the Archaeological Survey of India (approximately 55,000) (Caldwell, 1875). The rest of the paper is organized as follows: Section II describes in detail about paralinguistic information in speech and explores various acoustic features extracted from the speech signal. Section III discusses speech emotion recognition and presents a survey of research work done for Tamil language. The paper is concluded in Section IV, with discussions in this research area.

2 Background

Speech signal carries two types of information: linguistic information which conveys the spoken content and paralinguistic information that covers the speaker attributes. Paralinguistic information can be categorized into three types as shown in Figure 1. The following sections discuss about paralinguistic features, its applications and acoustic features that represent them.

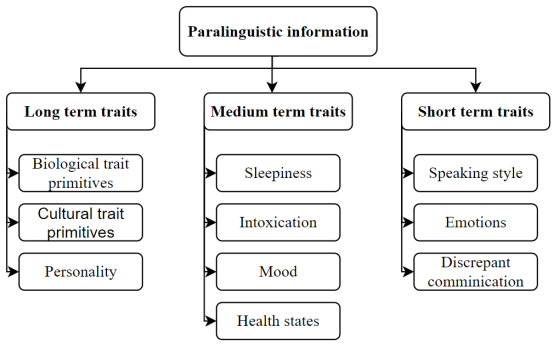

Figure 1: Types of paralinguistic information

2.1 Paralinguistic information in speech

Mainly paralinguistics deal with age, gender, personality, emotion, deviant speech, and discrepant communication. Biological trait primitives such as age and gender are designed by nature. Male and female voices differ due to physiological differences. In addition to this, age and cultural factors contribute to variations in speech. Cultural background determines the native language and regional dialect. Both the biological trait primitives and cultural trait primitives heavily influence the speech.

Deviant Speech is normally a long-term condition caused by disorders such as autism spectrum disorder (ASD), Alzheimer's disease, Parkinson's disease and motor neuron disease. Medium term states such as sleepiness and intoxication by alcohol consumption can also affect the normal speech. However, it returns to normal state quite soon. In the case of discrepant speech, speaker intentionally chooses to use deviant speech, deceptive speech, irony, or sarcasm. Detection of deceptive speech could be beneficial in therapeutic scenarios and forensic scenarios.

In speaker recognition, combined traits of a speaker are considered. Speaker recognition assigns a given speech segment to a particular speaker. Thus, speech features are used to find biological traits, thereby identifying the speaker. The typical application of speaker characteristics in forensics is speaker identification. The other field of speaker recognition is speaker verification, used in access control systems. The system has to verify that the given speech segment belongs to a speaker within a group of people who are allowed access.

The current emotional state of the speaker influences the tone, volume, and speech rate. Each emotion can be characterized by a unique acoustic pattern. Several studies on speech emotion recognition have been carried out over the time. Prosodic features such as pitch, duration, and quality are found to be important in emotion recognition. The way emotions are expressed depends on the personality of the speaker. Also, some aspects of emotions vary with language and culture. Therefore, it is challenging to develop emotion recognition systems across cultures.

2.2 Acoustic features representing paralinguistic information

Speech is a time varying signal and it is analyzed frame by frame. The most common features representing the speech signal are referred to as LLDs. Typical LLDs cover intensity, intonation, linear prediction cepstral coefficients (LPCCs), perceptual linear prediction (PLP), mel frequency cepstral coefficients (MFCCs), gammatone frequecny cepstral coefficients (GFCCs) formants, harmonicity, vocal cord perturbation etc. Breakdown of these features into three types namely prosodic, spectral, and voice quality features is shown in Table 1. These features are generally augmented by other descriptors computed from the raw LLDs such as

delta coefficients or regression coefficients.

MFCCs are the most widely used feature extraction technique in speech processing applications. MFCCs are computed as follows. Windowed speech signal transformed to frequency domain using discrete Fourier transform, mel filter bank is applied to the magnitude transform and discrete cosine transform is computed on the logs of powers at each mel frequencies.

Formants are resonant frequencies of the vocal tract. They vary according to the spoken content. In particular, the lower resonance frequencies of the vocal tract, that is, the first two formants are well correlated with the phonetic content while the higher formants describe speaker characteristics. Formants are mostly computed from LPCs. Fundamental frequency and formant frequencies are most important speech parameters and its detection has a significant influence in recognizing emotion from speech (Belean, 2013).

Vocal cord perturbation measures such as jitter and shimmer describe the quality of the voice. Jitter is the fluctuation in the length of the fundamental period from one cycle to the next and shimmer is the variation of the waveform amplitude from one cycle to the next. As they describe the pathological characteristics of voice, jitter, and shimmer measures are helpful in determining speaker age or voice pathology (Teixeira and Gonçalves, 2016).

Derived features can be computed from the LLDs and they could be combinations of the above-mentioned features. The most popular ones that capture the temporal information are the first and second order derivatives referred to as delta, delta-delta coefficients respectively.

Feature Type	Feature
Prosodic	Pitch
	Duration
	Energy
Spectral	MFCC
	GFCC
	LPCC
	PLP
	Formants
Voice quality	Jitter
	Shimmer
	Harmonics to noise ratio
	Normalized amplitude quotient
	Quasi open quotient

Table 1: Acoustic features

3 Speech Emotion Recognition for Tamil Language

Emotion recognition from speech is one of the major topics in the field of computational paralinguistics. It helps to understand the emotional condition and actual intentions of the speakers which would be beneficial in improving the speech based applications such as automatic speech recognition (Jose et al., 2012; Madhavaraj and Ramakrishnan, 2017; Lokesh et al., 2019) and spoken intent recognition (Yohan et al., 2019b,a). It can be used to enhance the human computer interaction and to identify the emotional state of the user in call centers. Several research studies have worked on developing speech emotion recognition systems using DNNs. Emotion expression varies across cultures and every emotion is expressed somewhat differently by each culture. Thus, there is a need to build and evaluate a Tamil emotional speech corpus to observe the representation of different emotions in Tamil speech. Some of the available Tamil speech datasets include mozilla common voice (Ardila et al., 2020), Open SLR - Tamil (He et al., 2020), Microsoft Speech Corpus (Indian languages) - Tamil (mic), and Tamil Speech Intent Dataset - UoM (Buddhika et al., 2018). However, there is no standard database available for Tamil emotional speech.

Emotional speech corpora can consist of three types of emotional speech: acted emotions which is simulated by actors, spontaneous emotions from real life situations, and elicited emotions stimulated through emotional movies, stories and games. It is desired to have good quality speech recordings of diverse set of emotions collected from a large group of speakers. Though spontaneous emotions are preferred, it is difficult to acquire. Therefore, most of the available speech corpora contain acted emotional speech.

Joe (2014) built a Tamil emotional speech corpus with speech data consisting of five emotions Happy, Sad, Anger, Fear, and Neutral. Speech recordings were collected from acted emotional speech in Tamil audio plays. Support vector machine (SVM)-based emotion recognition system was used to perform classification with MFCCs as input features. This corpus was extended in (Vasuki et al., 2020) and another emotion corpus was developed separately for children using the samples collected from Tamil movies. Having an

emotion corpus consisting of utterances of both the adults and children could help in investigating the influence of age in emotion expression.

In (Renjith and Manju, 2017), an acoustic feature-based emotion recognition system is presented. The database used in this work is Amritaemo (Poorna et al., 2015) which consists of speech recordings in Tamil and Telugu languages with three emotions: anger, happiness, and sadness. K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN) were used to classify emotions based on two features namely Linear Predictive Cepstral Coefficients (LPCCs) and Hurst Parameter. Hurst parameter is the decaying rate of the autocorrelation coefficient function of the speech signal. The experimental results showed that when individual features were considered, for both the languages Hurst parameter achieved higher accuracy, precision, and recall when compared to LPCCs while the combination of these features resulted in a better performance.

Murali et al. (2018) proposed a Gaussian Mixture Model (GMM) – Deep Belief Network (DBN) system for emotion recognition. It models GMM for each emotion independently using MFCC features extracted from the speech signal. The minimum distance between the distribution of features for each utterance with respect to each emotion model is derived as Bag of acoustic features and it is fed to DBN as the feature vector. The proposed system was evaluated on Berlin emotional speech corpus (Burkhardt et al., 2005) as well the emotional database for Tamil language. The database used in this work consists of four emotions namely anger, happy, sad, and neutral state for three languages Tamil, English, and Malayalam.

Ram and Ponnusamy (2018) developed a speech corpus to evaluate the Tamil speech emotion recognition of children with autism spectrum disorder. In addition, emotional speech recordings for Tamil and Telugu language were collected from movies and Berlin emotional speech corpus was used as well for training. The selected emotion classes were anger, neutral, happiness, sadness, and fear. MFCC, pitch frequency, and energy were used as the set of features fed to the SVM based classifier.

An emotion recognition system using weighted features is proposed in (Poorna et al., 2018). A speech dataset consisting of audio recordings expressing five emotions namely anger, surprise, disgust, sadness, and happiness was created for three South Indian languages Tamil, Telugu, and Malayalam. Mean and variance of energy contour, first five formant frequencies, and Linear predictive coding (LPC) coefficients were chosen as the speech features and classification was performed using KNN, SVM, and ANN. Classification performance using normal features and weighted features were compared and a significant increase in recognition accuracy was observed with weighted features. ANN achieved the best accuracy among the considered classifier models. The results indicated that the weight factors are language dependent for the input features.

The paper (Rajan et al., 2019) presents a multilingual emotional speech corpus TaMaR-EmoDB developed for Tamil, Malayalam, and Ravula. The corpus contains short speech utterances in five emotions: anger, anxiety, happiness, sadness, and neutral state. The corpus was built using simulated speech utterances where the subjects were asked to read a sentence, expressing a given emotion. It was evaluated using a DNN-based classifier and the classification was performed using the fusion of MFCCs and prosodic features such as short-time energy, zero crossing rate, and standard deviation, skewness, and kurtosis of pitch. Sowmya and Rajeswari (2020) created a speech dataset for Tamil language and built a classifier to predict emotions with MFCC and energy as the input features. The database consists of four emotions anger, sad, neutral, and happiness. KNN, SVM, and ANN were used for emotion classification where SVM and ANN achieved higher accuracy.

4 Conclusion

The presented study discussed about the paralinguistic information present in speech signal and its relevance in many speech processing applications. The Low-level descriptors characterize the underlying paralinguistic content and they can be used as input features to DNN based models. High level speech representations can be extracted directly from raw speech signal as well using DNNs. These models have shown significant performance in many paralinguistic tasks. Paralinguistic information can also be applied in improving speech recognition systems using normalization techniques. The most popular paralinguistic task, emotion recognition is challenging when applied across different languages and cultures since aspects of emotion expression are language and culture specific.

Therefore, a proper emotional speech database is required to build effective emotion detection models for Tamil language. Several research studies have created speech databases for Tamil language and evaluated them using SVM or DNN based classifiers which reported good performance. However, further improvements are still needed and a standard Tamil emotional speech database needs to be built with a diverse set of emotions and a large number of speakers.

Acknowledgment

This research was supported by Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Education, Sri Lanka funded by the World Bank.

References

Interspeech (2018) Low resource speech recognition challenge for Indian Languages.

I. Anosha and T. Uthayasanker. 2020. In *Tamil Internet Conference*.

R. Ardila, M. Branson, K.Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F.M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Bogdan Belean. 2013. Comparison of formant detection methods used in speech processing applications. *AIP Conference Proceedings*, 1565:85–89.

D. Buddhika, R. Liyadipita, S. Nadeeshan, H. Witharana, S. Jayasena, and U. Thayasivam. 2018. Voicer: A crowd sourcing tool for speech data collection. In *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 174–181.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A database of german emotional speech. volume 5, pages 1517–1520.

Robert Caldwell. 1875. *A comparative grammar of the Dravidian or South-Indian family of languages*. Trübner.

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2018. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global Wordnet Conference*, pages 77–86, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland. European Association for Machine Translation.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6494–6503, Marseille, France. European Language Resources Association (ELRA).

C. V. Joe. 2014. Developing tamil emotional speech corpus and evaluating using svm. In *2014 International Conference on Science Engineering and Management Research (ICSEMR)*, pages 1–6.

J. Melvin Jose, N. T. Vu, and T. Schultz. 2012. Initial experiments with tamil lvcsr.

S. Lokesh, Priyan Malarvizhi Kumar, M. Ramya Devi, P. Parthasarathy, and C. Gokulnath. 2019. An automatic tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. *Neural Computing and Applications*, 31.

A. Madhavaraj and A.G. Ramakrishnan. 2017. Design and development of a large vocabulary, continuous speech recognition system for tamil. In *2017 14th IEEE India Council International Conference (INDICON)*, pages 1–5.

Murali, Srikanth, Pravena, and Govind. 2018. Tamil speech emotion recognition using deep belief network(dbn). In *Advances in Signal Processing and Intelligent Recognition Systems*, pages 328–336.

S.S. Poorna, K. Anuraj, and G. Nair. 2018. A Weight Based Approach for Emotion Recognition from Speech: An Analysis Using South Indian Languages. In *Second International Conference, ICSCS*, pages 14–24.

S.S. Poorna, C. Jeevitha, S. Nair, S. Santhosh, and G.J. Nair. 2015. Emotion recognition using multi-parameter speech feature classification. In *2015 International Conference on Computers, Communications, and Systems (ICCCS)*, pages 217–222.

Rajeev Rajan, U.G. Haritha, A.C. Sujitha, and T.M. Rejisha. 2019. Design and Development of a Multi-Lingual Speech Corpora (TaMaR-EmoDB) for Emotion Analysis. In *INTERSPEECH*, pages 3267–3271.

C. Sunitha Ram and R. Ponnusamy. 2018. Toward design and enhancement of emotion recognition system through speech signals of autism spectrum disorder children for tamil language using multi-support vector machine. In *Proceedings of International Conference on Computational Intelligence and Data Engineering*, pages 145–158.

S. Renjith and K.G. Manju. 2017. Speech based emotion recognition in tamil and telugu using lpcc and hurst parameters — a comparitive study using knn and ann classifiers. In *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, pages 1–6.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayan. 2013. Paralinguistics in speech and language - state-of-the-art and the challenge. *Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language*.

Björn W. Schuller and Anton M. Batliner. 2014. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*.

V. Sowmya and A. Rajeswari. 2020. Speech emotion recognition for tamil language speakers. In *Advances in Intelligent Systems and Computing, Machine Intelligence and Signal Processing. MISP*, pages 125–136.

João Paulo Teixeira and André Gonçalves. 2016. Algorithm for jitter and shimmer measurement in pathologic voices. *Procedia Computer Science*, 100:271 – 279.

P. Vasuki, B. Sambavi, and Vijesh Joe. 2020. Construction and evaluation of tamil speech emotion corpus. *National Academy Science Letters*, 43.

K. Yohan, T. Uthayasanker, and R. Surangika. 2019a. Sinhala and tamil speech intent identification from english phoneme based asr. *2019 International Conference on Asian Language Processing (IALP)*, pages 234–239.

K. Yohan, T. Uthayasanker, and R. Surangika. 2019b. Transfer learning based free-form speech command classification for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 288–294, Florence, Italy. Association for Computational Linguistics.