On the use of Stress Information in Speech for Speaker Recognition

Laxmi Narayana M
TCS Innovation Lab - Mumbai,
Tata Consultancy Services, Yantra Park,
Thane (West), Maharastra, India.
Email: M.Laxminarayana@Gmail.Com

Sunil Kumar Kopparapu
TCS Innovation Lab - Mumbai,
Tata Consultancy Services, Yantra Park,
Thane (West), Maharastra, India.
Email: SunilKumar.Kopparapu@TCS.Com

***Abstract*—The performance of a speaker recognition system decreases when the speaker is under stress or emotion. In this paper we explore and identify a mechanism that enables use of inherent stress-in-speech or speaking style information present in speech of a person as additional cues for speaker recognition. We quantify the the inherent stress present in the speech of a speaker mainly using 3 features, namely, pitch, amplitude and duration (together called PAD) We experimentally observe that the PAD vectors of similar phones in different words of a speaker are close to each other in the three dimensional (PAD) space confirming that the way a speaker stresses different syllables in their speech is unique to them, thus we propose the use of PAD based speaking style of a speaker as an additional feature for speaker recognition applications.**

I. Introduction

Several speech features for speaker recognition and speech recognition applications have been proposed in literature. Mel Frequency Cepstral Coefficients (MFCC) [1] by far, have been the most commonly used speech features [2] [3] [4] [5] [6]. Other speech features like Linear Frequency Cepstral coefficients (LFCC) [1], wavelet octave coefficients of residues (WOCOR) [7] are also used for speech and speaker recognition applications and good performance of these systems have also been reported. Combination of two or more speech features for speaker recognition applications is also in practice and the literature reports an improved performance with combination of multiple speech features [7]. Nevertheless automatic speech and speaker recognition systems function less efficiently when the speaker is under stress or emotional state. In this paper, we assume that the way people express these emotions and stress certain syllables in their speech is unique to them and this style of speaking is consistent for a speaker. We make this assumption based on observations and this motivates the current work. We further explore to identify a mechanism to use the inherent stress information in the speech as an additional feature for speaker or speech recognition[1].

Prosody is the melody of speech [8] and emotion-initiated gestures and stressed syllables in human speech communication help in the improvement of speech understanding. However, stress and emotional expressions in human speech have been found to be the source of difficulty for some applications like automatic speech recognition (ASR) and speaker recognition (SR). The reason behind the manifestation of stress in human speech is due to several factors. The first and foremost reason is that emphasis of some syllables in human speech is natural and this becomes very prominent when the speaker is emotional or under stress. Many studies that consider stress in speech as distortion introduced by emotion have shown that these factors can severely reduce speech recognition accuracy. Techniques for detecting or assessing the presence of stress could help neutralize stressed speech and improve robustness of speech recognition systems. Although some acoustic variables derived from linear speech production theory have been investigated as indicators of stress, they are not consistent.

Milan [9] mentions in his work on spectral analysis of stressed speech that stress is a psycho-physiological state characterized by subjective strain, dysfunctional physiological activity and deterioration of performance. The accepted term for speech signal carrying information on the speaker's physiological stress is 'stressed speech'. This refers to the imprints of stress in the speech of person when under stress. In other words, the stress experienced by the speaker is reflected in their speech. In contrast to this [10] reports the work of stress labeling of syllables, analyzes the stress in normative speech - uttered in a neutral speaking style or when the speaker is not stressed. In this case, the 'stressed speech' refers to that portion of speech, which is stressed naturally with no influence of the speaker's mental or psycho-physiological state.

As mentioned earlier, our motivation for using the natural 'stress' information in human speech as a feature for speaker recognition is based on the assumption that the *speaking style* of a speaker is consistent and unique to the speaker. One can also extend this idea to say that the people of same geographical area have a similar speaking style (or accent). This paper is motivated by the question *"If the way a speaker speaks is unique to the speaker, then why not use that 'speaking style' information specifically as an additional feature for speaker recognition?"* We interpret that the speaking style of a speaker is a reflection of the way the speaker stress certain syllables in a sentence or phrase or a word. We believe that the speaking style of a speaker can be parametrized by identifying the parameters related to *stress* in speech.

The rest of the paper is organized as follows. Section II

[1]These features are in addition to the traditionally used speech features

978–1–4244–4547–9/09/$26.00 ©

gives a summary of literature on the stress related features used in speech. Section III discusses our approach of using the inherent stress information in speech specifically for speaker recognition. Section IV gives the details of the experiments conducted to quantify the speaking style of a speaker. and we conclude in Section V.

II. Stress related features in Speech

Stress and its manifestation in the acoustic signal have been the subject matter of many studies in literature [10] [11] [12]. Researchers have attempted to determine reliable indicators of stress by analyzing certain variable parameters of speech such as fundamental frequency (pitch), amplitude, concentration of spectral energy, duration and several others [12] [13] [14] [15] [16]. In literature, analysis of stress is performed through analysis of some parameters of stress like fundamental frequency (F0), pitch, vowel duration and formants in recorded emotional speech, namely, analyzing a speaker's speech when they are under stress, fatigue, heavy workload, environmental noise, sleep loss or expressing some emotion like happiness, anger or sorrow.

Literature clearly distinguishes the speech as (a) uttered when the speaker is under stress or expressing some emotion and (b) uttered in a neutral speaking style, namely, when the psychological state of the speaker does not seriously affect the speech, for example, reading news or even normal conversations. To our best knowledge the available literature talks only about the speech of type (a) to analyze stress in speech. The inherent presence of stress in some syllables of speech whether it is emotional or normative, is natural in accordance with the influencing factors like language, accent and the geographical location to which a speaker belongs to. If the stress is absent in non-emotional speech, the corresponding pitch and amplitude contours would have been absolutely flat. We believe that the manifested stress in speech whether the speaker is under stress or in normal condition, is distinguished by the intensity of the parameters of speech (or stress), but, nevertheless, the set of parameters that quantify stress is same in both the cases. A speaker naturally and unintentionally stresses some syllables while speaking; there exists an inherent mechanism behind this unintentional occurrence of stress in speech, which is unique to a person or people belonging to a geographical region. We believe that this information can be used as an additional feature in speaker recognition applications. We intend, in this paper, to study the parameters of stress and seek to use the stress information in normative speech in speaker identification or verification. We summarize our study of literature on how 'stress' in speech is parameterized.

Higher intensity, greater duration and higher F0 are believed to be the primary acoustic cues for stressed syllables, although how these three factors work together to make a syllable more prominent than the surrounding ones is still not very clear. Therefore, these cues are used as the main acoustic features in the stress detection task in some studies (example, [13]). Stressed syllables [11] are usually indicated by high sonorant energy, long syllable or vowel duration and high and rising F0. Stress is found to be correlated with voice quality as well. Usually, stressed vowels are pronounced clearer and unstressed vowels tend to have reduced clarity. When listening to an utterance, people not only use acoustic cues but also syntactic and/or semantic cues to help identify the location of stress in speech. Therefore, features derived from the text, such as part of speech (POS) and the position in the phrase are as well used for detection of stress [14]. Many works in literature mention that the assignment of stress is based on a relative comparison of the syllables within a word and does not rely on a global model of a stressed or unstressed syllable.

Stress, accent and/or emphasis detection all deal with the detection of the relative prominence of a syllable within a word. A discussion of the different detection methods is difficult, because the word 'stress' has been used ambiguously to refer to several types of prominence, including strong versus weak syllable distinctions as indicated by lexical stress marking, as well as phrasal prominence as indicated by a pitch accent [14]. The cues for stress discussed above have been found mainly in English and Dutch languages based on the work carried out for detection of stressed syllables. Further, the applications for which such analysis of stress were carried out were automatic speech recognition (ASR) and speaker recognition. Also the speech corpus analyzed for the study was, in most cases, a biased one; for example, emotional speech was recorded (with happiness, anger, sadness) and used for stress analysis. Not much literature is found on stress labeling of syllables from neutral (speaking style) speech.

In our work, the attributes chosen to quantify stress in speech (whether emotional or non-emotional) are pitch, amplitude and duration of a phone or syllable. Hereafter, the combination of these three parameters - pitch, amplitude and duration of a phone/syllable will be referred to as PAD[2].

III. The Approach

Our approach of using the inherent stress information in speech for speaker recognition is as follows. We first collect a database of some spoken words in which all the phonemes in a language occur. We make sure that for a given a phoneme, we can find several instances of it in the database. For example, a phoneme `/a/` should occur in at least a few words in the database, say, `/jar/`, `/ball/`, `/walk/`, `/mark/`, `/hard/` etc. These words are recorded, say, m times from say, n different speakers at different times of the day over a period of several days. The recorded speech samples are then segmented and phoneme labeled[3]. We use PRAAT [17] for manually segmenting and phoneme labeling the speech samples. We then extract the characteristics, namely, pitch, amplitude and duration (PADs) of different phones in different words and observe the variation of PADs across different instances of phones.

[2]A syllable / phone has certain PAD means that its pitch is P Hz, its amplitude is A dB and it exists for D seconds.

[3]Segmenting and phoneme labeling is a process of marking the starting and ending times of phones uttered in the speech sample. This process could be automatic or manual.

The expectation is that the pitch (amplitude and duration) contours of different instances of the same word of the same speaker appear more or less identical. This can be thought of as the 'speaking style of the speaker' which we are interested to capture. One may also be interested in observing the influence of the adjacent phones on the characteristics of a phone (especially vowels); for example, how the characteristics of the vowel /a/ change in different phonetic contexts. The idea is to come up with a mechanism where one can represent a speech utterance with a stress or accent contour that represents the 'speaking style' of a person, and to use this information in speaker recognition to recognize a speaker.

The mean of the PADs of different instances of each phone uttered by a speaker is calculated. For each phone uttered by a speaker, we have a corresponding mean PAD vector

$$PAD_m = [P_m, A_m, D_m]$$

where P_m, A_m and D_m are the average pitch, average amplitude and average duration of different instances (recordings at different timings) of a phone. The variance of the parameters of each phone or the percentage deviation of each parameter about the mean are also calculated. The acceptable deviation range of a parameter of a particular phone is determined. Now the speech feature database of a speaker contains the mean PAD vectors of all the phones in a given language, the variance or the percentage deviation (deviation range) of each parameter (PAD) about the mean.

In the training phase, each speaker has a set of phones and their corresponding mean PAD vectors. In the testing phase, we extract the PAD contours of the phones in the sentence uttered by the speaker. For a verification system, we construct the corresponding PAD contours of the sentence from the training database and calculate the distance between the contours of the uttered sentence and that which is constructed from the training database. If the difference falls within the 'deviation range' then the speaker is genuine, else an imposter. For an identification system, we construct the corresponding PAD contours of the sentence from the training database for all the speakers and calculate the distance between the contours of the uttered sentence and those which are constructed from the training database. The identified speaker is the one whose constructed PAD contours have minimum distance from the PAD contours of the uttered sentence. Note that the minimum distance should also fall within the deviation range.

IV. Experiments

For our initial experimentation, we chose 4 speakers and recorded from each of them, 6 English words. Speakers were asked to speak the words multiple number of times to check the consistency and range of the variance of the characteristics of their speech samples over the recordings at different timings. The words are chosen such that they have a common phoneme /a/ in different phonetic contexts. The chosen words are ball, car, example, hard, mark and wall.

The recorded speech samples are manually segmented and labeled using PRAAT [17]. After segmentation of the speech samples recorded from different speakers, the values of pitch, amplitude and duration (PADs) are obtained for each instance of the phone /a/ from a speaker. A three-dimensional vector $PAD_{/a/,S_i} = [P, A, D]$ is formed[4]. The values of pitch, amplitude and the duration of a phone are obtained as follows. PRAAT software has a provision to get the 'pitch listing' of a speech file, which gives the values of the pitch for every k ms[5]. We obtained the values of pitch for every 10 ms and P is the maximum value among those pitch-values that fall within the duration of a phone. A is obtained from the 'intensity listing' of the PRAAT. We obtained the values of amplitude for every 10 ms and A is the maximum value among those intensity-values that fall within the duration of a phone. The duration, D of each phone is obtained from the phone boundaries (on the time scale) established by the manual speech segmentation process. The PAD values and their corresponding mean, variance or the percentage deviation about the mean are obtained for the samples of all the speakers. Values of the above mentioned parameters for one set of speech samples of a male speaker and are shown in Table I.

TABLE I
PAD statistics of the phone /a/ recorded from a male speaker, Sampling frequency: 8 kHz

Word	Duration (sec)	Amplitude (dB)	Pitch (Hz)
ball	0.178	52.787	110.239
car	0.170	47.400	113.772
example	0.128	50.657	116.262
hard	0.205	52.001	127.806
mark	0.124	50.629	118.914
wall	0.159	50.618	110.233
Mean	0.161	50.682	116.205
Variance	0.0008	2.825	36.518
%deviation	17.554	3.316	5.200

PADs are collected in a similar manner for the sets of speech samples recorded from the other speakers. As mentioned in Section III, we form the mean PAD vectors $PAD_m = [P_m, A_m, D_m]$ corresponding to the different phones of a particular language for each speaker. We compute the other parameters namely, variance or percentage deviation about the mean also for the different phones (in a language) uttered by each speaker.

After this process, we have the corresponding mean PAD vectors for each phone for each speaker. For speaker recognition applications, we extract the PAD contours of the test sentence uttered by the speaker.

- For speaker verification system, the corresponding PAD contours of the sentence are constructed by concatenating the already available mean PAD values from the training database of the corresponding speaker. We can then calculate the Euclidean distance between the PAD contours of the uttered sentence and that which is constructed from the training database. If the difference falls within the 'deviation range' calculated above, then the speaker is declared to be genuine, else an imposter.

[4] i is the index of the speaker
[5] k is adjustable

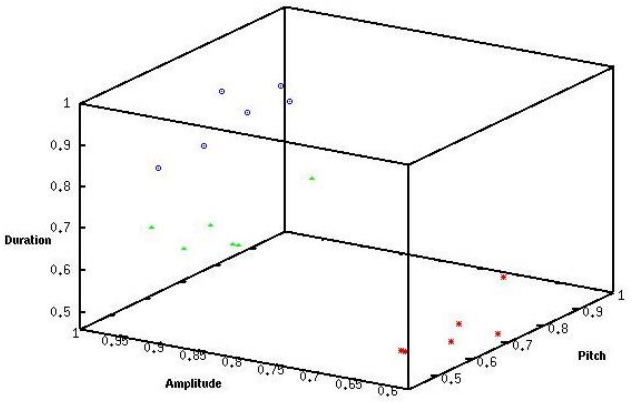

Fig. 1. 3 dimensional PAD vectors of the phone /a/ for three different speakers; The values of P, A and D are normalized with respect to the maximum value

- For speaker identification, we obtain the PAD contours of the sentence uttered by a speaker. We then construct the corresponding PAD contours of the same sentence from the training database, for all the speakers. We then calculate the Euclidean distance between the contours of the uttered sentence and those which are constructed from the training database. The identified speaker is the one whose constructed PAD contours have minimum distance from the PAD contours of the uttered sentence. Note that, another criteria is that the minimum distance should also fall within the deviation range.

An an initial check of the proposed idea, we plotted the three-dimensional PAD vectors obtained for different instances of the phone /a/ uttered by three different speakers in the three dimensional space. We found that the PAD vectors of the same phone uttered by there different speakers are well separated (see Figure 1). This gives us hope that we can record speech samples of all the phones from different speakers construct the training database completely and add the additional feature set to our existing speaker recognition system.

Along with the PADs which are identified as features for stress in speech, we are also collecting other features like formants, number of cycles per second for further analysis. We hope that this analysis enables us to strongly parametrize the speaking style of a speaker and thus use this as an efficient feature for speaker recognition.

V. Conclusion

Speaker recognition systems show a degraded performance when the speaker is under stress or emotion. We made an assumption that there is inherent stress or emotion related characteristics present in spoken speech of a person all the time; in addition the stress related features are consistent. Using this assumption as the base, we proposed a mechanism that enables the use of inherent stress in speech (speaking style) for speaker recognition. We propose that the stress related features be used in addition to the regular features used for speaker recognition. We identified 3 features which capture stress in speech, namely, pitch, amplitude and duration. We experimentally observe that the PAD vectors of the similar phones of a speaker are close to each other in the three dimensional space confirming our assumption that the way a speaker stresses different syllables in their speech is unique to themselves. Having observed experimentally that our assumptions are valid, we further proceed to construct a training database of PAD values of all the phones in a language for several speakers and incorporate the proposed mechanism in our speaker recognition system.

References

[1] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, No. 1, January 1995.

[3] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, No. 1-2, pp. 91–108, 1995.

[4] M. R. Hasan, M. Jamil, M. G. Rabbani, and M. S. Rahman, "Speaker identification using mel frequency cepstral coefficients," *3rd International Conference on Electrical & Computer Engineering ICECE 2004*, 28-30 December 2004, Dhaka, Bangladesh.

[5] H. Seddik, A. Rahmouni, and M. Sayadi, "Text independent speaker recognition using the mel frequency cepstral coefficients and a neural network classifier," *First International Symposium on Control, Communications and Signal Processing*, pp. 631–634, 2004.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, No. 1, pp. 19–41, 2000.

[7] N. Zheng, T. Lee, and P. C. Ching, "Integration of complementary acoustic features for speaker recognition," *Signal Processing Letters, IEEE*, vol. 14, Issue 3, pp. 181–184, March 2007.

[8] T. F. Quatieri, "Speech signal processing, theory and practice," *Pearson Education*, vol. 2, 2004.

[9] M. Sigmund, "Spectral analysis of speech under stress," *International Journal of Computer Science and Network Security*, vol. 7, No.4, pp. 170–173, April 2007.

[10] L. N. M and A. G. Ramakrishnan, "Defining syllables and their Stress labels in Tamil TTS corpus," *Proc. of Workshop in Image and Signal Processing (WISP-2007), IIT Guwahati*, vol. 2, pp. 92–95, Dec 28-29, 2007.

[11] C. Wang and S. Seneff, "Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER Domain1," *EUROSPEECH*, September 2-7, 2001, Aarlborg, Denmark.

[12] Cairns, A. Douglas, Hansen, and H. L. John, "Nonlinear analysis and classification of speech under stressed conditions," *The Journal of the Acoustical Society of America*, vol. 96, Issue 6, pp. 3392–3400, December 1994.

[13] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. on Speech and Audio Processing*, vol. 2, No.4, pp. 469–481, 1994.

[14] M. Lai, Y. Chen, M. Chu, Y. Zhao, and F. Hu, "A hierarchical approach to automatic stress detection in English sentences," *ICASSP*, pp. 753–756, 2006.

[15] R. Balusu, "Acoustic correlates of stress and accent in Telugu," *21st South Asian Languages Analysis Roundtable, University of Konstanz*, October 7-10, 2001.

[16] L. Astruc and P. Prieto, "Acoustic cues of stress and accent in Catalan," *University of Cambridge and ICREA-UAB*.

[17] PRAAT, "A tool for phonetic analyses and sound manipulations by Boersma and Weenink," *www.praat.org*, 1992-2001.