Classifying *Style* of Spoken Speech

Sunil Kumar Kopparapu, Sathyanarayana, Akhilesh Srivastava, P.V.S. Rao

Abstract— **The ability to classify spoken speech based on the style of speaking is an important problem. With the advent of BPO's in recent times, specifically those that cater to a population other than the local population, it has become necessary for BPO's to identify people with certain style of speaking (American, British etc). Today BPO's employ accent analysts to identify people having the required style of speaking. This process while involving human bias, it is becoming increasingly infeasible because of the high attrition rate in the BPO industry. In this paper, we propose a new metric, which robustly and accurately helps classify spoken speech based on the style of speaking. The role of the proposed metric is substantiated by using it to classify real speech data collected from over seventy different people working in a BPO. We compare the performance of the metric against human experts who independently carried out the classification process. Experimental results show that the performance of the system using the novel metric performs better than two different human expert.**

I. Introduction

BPO's (Business Process Outsourcing) centers are increasingly finding their way because of the increased quality consciousness, particularly in the service industry segment. Development in the area of telecommunications make it feasible for the BPO's to be located in regions which it is servicing other than the local population. In addition socio-economic reasons justify the geographical location of BPO's anywhere without the people being serviced being aware of it. This has led to a spate of BPO's cropping up in developing countries where there exists a large population that can speak the language of the people not necessarily in the same style. For this reason, there is no definite recruitment qualification that one should possess to join a BPO, except that, one be able to speak in the style of the population that the BPO services. The increase in number of BPO's and the no specific qualification requirement, leads to a situation of total influx, people are always on the move (high attrition). This leads to the requirement of a constant recruitment process at the BPO's. Today, BPO's with no exception, employ accent analyst to select candidates. The accent analyst judges the suitability of a candidate by analyzing the speaking style of the candidate. The process of recruitment is time consuming (on an average only about 7% of the candidates appearing for the interview are selected) and is prone to human bias. There is a need for an automatic system that can measure the candidates speaking style or more precisely, classify the candidates speaking style as being suitable (good), trainable (average) or unsuitable (bad).

Often one is able to make out the speakers background (American, British, Indian etc) by just listening to the spoken speech of the person. In addition, one is also able to tell if the person is speaking well or not, even in the absence of knowledge of the language being spoken. Thus, it is possible for a human to categorize speakers based of their speaking style by listening to their speech. A trained human is able to perform this task of classification better because he is aware of the nuances of what to be on the lookout for which identifies a well spoken speech. An ideal system[1] would be the one that has the ability to classify people based on their speaking style by looking at their free-spoken speech. While work is on by several researchers the development of such a system is still premature [1].

In this paper, we propose a system that can be used to classify people based on their speaking style[2]. The system captures the speaking style of a person by analyzing the spoken speech samples of a person. The analysis is carried out on a predetermined set of words and sentences[3].

II. Our Approach

In our approach a speaker to be accessed is asked to speak $\mathcal{W}$ predefined words (or sentences). The speaking style and articulatory capability of spoken speech can be assessed automatically by *comparing* the test speech sample of the speaker for every word $w \in \mathcal{W}$ with the corresponding statistical models (HMMs) of the w of $\mathcal{G}$ classification groups (example, Very Good, Good, Average, Bad, Very Bad). The comparison would give a measure of the closeness of the test sample to each of the $|\mathcal{G}| = 5$ categories. These closeness scores for all the $w \in \mathcal{W}$ to each of the classification groups is combined to come out with an overall category classification. The HMM [2] of each word w for each classification category $\mathcal{G}$ is constructed us-

Cognitive System Research Laboratory, Tata Consultancy Services Limited, Navi Mumbai. Email: SunilKumar.Kopparapu@TCS.Com

[1] Essentially the system would be built by first analyzing and deriving rules by listening to spoken speech samples. These rules would enable development of the system to determine the quality of speech.

[2] To assist BPO recruitment process.

[3] The speaker would be asked to speak a carefully selected list of words and sentences, which would be used by the system to analyze the style of speaking.

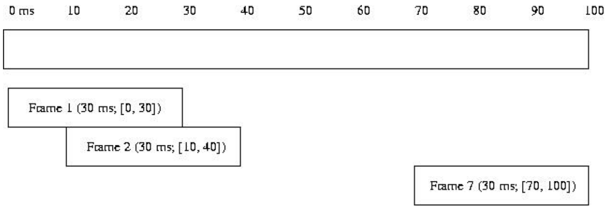

Fig. 1. If Speech signal duration is T (= 100 ms), then total total number of frames is given by (T - F_{size})/F_{shift} = (100 - 30)/10 = 7, provided frame size (F_{size}): 30 ms and frame shift (F_{shift}): 10 ms. Typically, F_{shift} = F_{size}/3, F_{size}/2

ing training samples. The training samples are collected from several people in different acoustic conditions and are classified by a human expert into one of the $\mathcal{G}$ categories.

A. Building Statistical Models (Training)

The training phase (model development phase) consists of collecting samples of speech data from several users in different acoustic conditions and at different time of the day and using these speech samples to develop HMM. Specifically, reference speech data is collected for $\mathcal{W}$ predefined words (or sentences) from $\mathcal{G}$ classification groups (example, Very Good, Good, Average, Bad, Very Bad). Each classification group has data from several different people.

Assume that each group $\mathcal{G}$ has $N_{\mathcal{G}}$ number of persons in it. For each $w \in \mathcal{W}$ and $g \in \mathcal{G}$, construct an HMM $\mathcal{H}_{wg}$ using the reference speech samples belonging to word w and category g collected from N_g persons.

The $\mathcal{H}_{wg}$ is estimated by first preprocessing all the reference speech samples for the word w and for the group g. The preprocessed speech samples are used to build HMM models using Baum-Welch training[4]. In speech signal processing preprocessing the speech sample usually involves

1. Dividing the speech signal into overlapping frames of 30 ms (see Figure 1).
2. Removing non-speech signal at the beginning and the end of the utterance using energy based algorithms. An energy based thresholding ($\mathcal{T}$) is used to determine if each frame window of the speech signal is a speech frame or a non-speech frame.

For N frames compute. Compute $\mathcal{A}^1_{max}, \mathcal{A}^1_{max}, \cdots \mathcal{A}^N_{max}$ (maximum amplitude in each frame). Compute,

$$\mu = \frac{1}{N}\sum_{i=0}^{N} \mathcal{A}^i_{max};$$

$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{N} (\mathcal{A}^i_{max} - \mu)^2;$$

$$\mathcal{T} = \frac{(\mu + 2\sigma^2) + \mathcal{A}_{max}}{\gamma}$$

Frame i is a speech frame if ($\mathcal{A}^i_{max} > \mathcal{T}$). The speech frames between the first identified speech frame from the start of the speech file and the last speech frame is the end silence detected speech. In all our experiments we choose $\gamma = 15$.

3. Smoothing the signal to remove noise (goes by the name pre-emphasis). Pre-emphasizing[5] the signal is necessary to spectrally flatten the signal to make it less susceptible to finite precision effects in signal processing. This is done by a 1^{st} order Finite Impulse Response (FIR) filter. The impulse response $H(z)$ of a pre-emphasis filter is

$$H(z) = 1 - \phi z^{-1} \quad \text{where} \quad \phi \in [0.9, 1.0]$$

In the time domain this is equivalent to the difference equation

$$s_p(n) = s(n) - \phi s(n-1)$$

where, $s_p(n)$ is the n^{th} sample of the pre-emphasized signal, $s(n)$ is the n^{th} sample of the original signal and ϕ is the pre-emphasis factor.

4. Tapering the frames (Hanning window) to remove introduction of high frequency noise. Windowing is done on each frame of the speech signal to minimize signal discontinuities at the beginning and the end of each frame. The signal s_f (N is the number of speech samples in a frame) is multiplied by a Hamming window $w(n)$ of length N (see Fig. 2).

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \le n \le N-1$$

The windowed signal ($s_w(n)$) is obtained as

$$s_w(n) = s_f(n)w(n)$$

where $s_f(n)$ is the speech frame

B. Parameter Extraction

For each frame a set of parameters are extracted. These parameters, which are representative of the speech signal, are used to build HMMs. In all we build $\mathcal{W} \times \mathcal{G}$ models (all words and all categories). The extracted parameters used can be categorized into two groups. The articulatory capability is characterized by the parameter set $I\!\!D$ while the intonation capability is captured by the parameter

[4] in this paper we do not describe the actual process of training because it is captured very well in the literature [3]

[5] The overall effect is to emphasize the high frequency content and deemphasizing the low frequency content. This is done to compensate for the attenuation caused by the radiation from the lips.

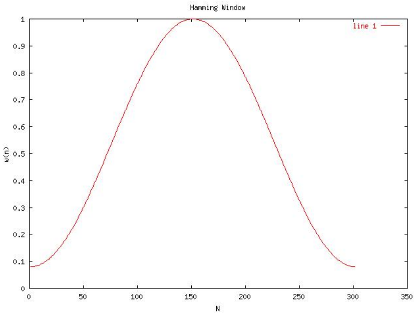

Fig. 2. Hamming Window

set $\mathbb{I}$. Both together, characterize the speaking style of the spoken speech. While $\mathbb{D}$ captures the closeness of the content of the two spoken words or sentences, $\mathbb{I}$ capture the closeness in terms of *intonation* or the characteristic style of spoken speech. Note that $\mathbb{D}$ captures the spectral properties of the speech signal (MFCC), $\mathbb{I}$ depends on the parameter *pitch* and the *stress* of the spoken speech [4].

In our experiments, $\mathbb{D}$ consists of 12 MFCC[6], 8 Δ MFCC, 4 Δ^2 MFCC, while $\mathbb{I}$ consists of variation in pitch and variation in amplitude. In all for every frame of speech 26 features are extracted.

C. Classification

Given a test speech sample t and the number of groups $\mathcal{G}$, the problem is one of tagging the given test speech sample, t of the word x into one of the $|\mathcal{G}|$ groups based on the *closeness*[7] of the test speech sample to the $|\mathcal{G}|$ HMMs corresponding to the word x. We compare the test speech sample, t, with the reference HMMs $\mathcal{H}_{xj}$ for $j = i = 1, \cdots, \mathcal{G}$ and calculate the scores $\mathcal{T}_j = d(\mathrm{t}, \mathcal{H}_{xj})$. The test speech sample, t, is classified as belonging to the group g if

$$\mathcal{T}_g < T_j$$

$\forall j = 1, \cdots, \mathcal{G}$ and $j \neq g$. Note that the categorization is associated with a scalar number p which gives the degree of closeness of the test sample to the HMM.

The classification of a user is based on a total of 20 words and sentences. Each of these 20 words spoken by a user is classified into one of the $|\mathcal{G}|$ groups with an associated closeness match, say p_i, where $i = 1, \cdots, 20$. A final score is obtained by

$$S = \sum_{i=1}^{20} p_i W_i$$

where W_i is the weight associated (see Table I) with categorization of word i. The over all category of the person was determined by the value of S using Table II.

Category	Weights
Very Good	5
Good	4
Average	3
Bad	2
Very Bad	1

TABLE I
Weights associated with category.

Person Category	Overall Score
Very Good	81 - 100
Good	61 - 80
Average	41 - 60
Bad	21 - 40
Very Bad	0 - 20

TABLE II
Person categorization based on overall score.

III. Experimental Results

A set of 20 words and sentences were selected in consultation with phoneticians and accent training experts. The set consisted of words and sentences which were very commonly prone to pronunciation error and in some cases the words were tongue twisters. The choice of the set is deemed to be capable of assessing the development of articulation of a person. Data was collected from a set of 30 people in each category (Very Good, Good, Average, Bad and Very Bad speaking style). All person were asked to speak the predetermined set of 20 words and sentences on the telephone using an IVR application custom built for collecting data. The speech collected was sampled at 8 kHz, 8 bit. The speech data was tagged separately by two accent experts into one of the five (Very Good, Good, Average, Bad, Very Bad) categories. Of the 30 speakers data, data collected from 20 speakers was used to construct the HMMs, while the other 10 speakers[8] were used to test the speaking style classification. This ensured that there was no person based characteristics induced into the constructed HMMs.

Table III gives the agreement between two human accent experts on the 10 speaker test data. Total agreement is when both the human experts categorized the same speech sample as belonging to the same category (example, both the experts say that the speech sample is good) and 1-step agreement corresponds to the the human experts differing on their categorization by a distance of 1 category (example, one expert say that the speech sample is good while the other says that the speech

[6] Mel Frequency Cepstral Coefficient

[7] in terms of log probability score

[8] not part of the training set

	Expert 1 - Expert 2
Total Agreement	26 %
1-step Agreement	45 %

TABLE III
Agreement between two human experts.

Agreement	Exprt 1 - Sys	Exprt 2 - Sys
Total	56 %	47 %
1-step	100 %	90 %

TABLE IV
Agreement between the system and the two human experts.

sample is very good or average). The overall performance of the system (on the 10 person test data[9]) for classifying spoken speech is tabulated in Table IV. As observed, the performance of the automated system is *much better* than the performance between two different human experts. Notice that the performance of the human expert - system (see Table IV) is better than the expert-expert (see Table III) performance.

IV. Conclusions

With increase in BPO's there is a need for automatic speaking style analyzer. Speaking style analysis by human experts is bound to be biased by cues that might not necessarily be associated with the speaking style and the judgment of the speaking style is dependent on the human expert. To over come this bias that may be associated with human expert in analyzing a person for his speaking style we have developed a system to automatically analyze the speaking style of a person. We proposed a set of parameters which captures both the articulatory capability and the intonation of the speaker, both of which jointly characterize the speaking style of the person. Experimental results show that the performance of the system far exceeds the performance between two independent human experts. This system was successfully piloted in a BPO to act as a first level screening agent in their recruitment process.

V. Acknowledgments

The authors would like to thank several members of the Cognitive Systems Research Laboratory, Tata Consultancy Services, Mumbai for their active contributions to various phases of this work.

References

[1] P. V. S. Rao and Sunil Kopparapu, "An approach towards automatic evaluation of accent and style," in *COCSDA*, CDAC, Noida, India, November 2004.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications n speech recognition," *Proc. of IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[3] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of speech recognition*, Prentice Hall, New Jersey, 1993.

[4] Fujisaki H, "Prosody, information, and modeling with emphasis on tonal features of speech," in *Proc. Workshop on Spoken Language Processing*, TIFR, Mumbai, India, January 2003, pp. 3–12.

[9] three round robins