Identifying Perceptually Similar Languages Using Teager Energy Based Cepstrum

Hemant A. Patil[1] *and T.K.Basu*[2]
[1]Dhirubahi Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India.
[2]Department of Electrical Engineering, IIT Kharagpur, West Bengal, India.

Abstract— **Language Identification (LID) refers to the task of identifying an unknown language from the test utterances. In this paper, a new method of feature extraction, *viz.*, Teager Energy Based Mel Frequency Cepstral Coefficients (T-MFCC) is developed for identification of perceptually *similar* languages. Finally, an LID system is presented for Hindi and Urdu (perceptually similar Indian languages) to demonstrate effectiveness of newly proposed feature set with short discussion on experimental results.**

***Keywords*- Language identification, Teager Energy Operator (TEO), Mel cepstrum, polynomial classifier, discriminative training.**

I. INTRODUCTION

LANGUAGE Identification (LID) refers to the task of identifying an unknown language from the test utterance. LID applications fall into two main categories: pre-processing for machine understanding systems and preprocessing for human listeners. Alternatively, LID might be used to route an incoming telephone call to a human switchboard operator fluent in the corresponding language. Such scenarios are already occurring today: for example, AT&T offers the *Language Line* interpreter service to, among others, police departments handling emergency calls. When a caller to *Language Line* does not speak any English, a human operator must attempt to route the call to an appropriate interpreter. Much of the process is trial and error (for example, recordings of meetings in various languages may be used) and can require connections to several human interpreters before the appropriate person is found. As reported by Muthusamy [19], when callers to Language Line do not speak any English, the delay in finding a suitable interpreter can be of the order of minutes, which could prove devastating in an emergency situation. Thus, an LID system that could quickly determine the most likely languages of the incoming speech might cut the time required to find an appropriate interpreter by one or two orders of magnitude [20],[29]. In addition to this, in the *multilingual* countries like India, automatic LID systems have an important significance because multi-lingual interoperability is an important issue for many applications of modern speech technology. The need for development of multi-lingual speech recognizers and spoken dialogue systems are very important in Indian scenario. An LID system can be connected as an excellent front-end device for multi-lingual speech recognizers or language translation systems [18].

Human beings and machines use different perceptual cues (such as phonology, morphology, syntax and prosody) to distinguish one language from the other. Based on this, to solve LID problem, following approaches are used [18]:

- spectral similarity approaches;
- prosody-based approaches;
- phoneme-based approaches;
- word level based approaches;
- continuous speech recognition approaches.

It has been observed that human beings often can identify the language of an utterance even when they have no strong linguistic knowledge of that language. This suggests that they are able to learn and recognize language-specific patterns directly from the signal [12], [18]. In the absence of higher level knowledge of a language, a listener presumably relies on lower level constraints such as acoustic-phonetic, syntactic and prosody. In this paper, spectral similarity based approach for language identification is used which concentrates on the differences in spectral content among languages. This is for exploiting the fact that speech spoken in different languages contains different phonemes and phones. The training and testing spectra could be used directly as feature vectors or they could be used instead to compute cepstral feature vectors [18]. In addition to this, a new method of feature extraction, *viz.*, Teager Energy Based Mel Frequency Cepstral Coefficients (T-MFCC) is developed for identification of perceptually *similar* languages.

The organization of the paper is as follows. Section II discusses details of Teager Energy Operator (TEO). Development of T-MFCC is given in section III. Section IV discusses details of experimental setup used in this study. Finally, section V discusses experimental results while section VI concludes the paper with short discussion.

II. TEAGER ENERGY OPERATOR

Speech features such as LPC, LPCC and MFCC are derived from a linear speech production models which assumes that airflow propagates in the vocal tract as a linear plane wave. This may be due to the fact that the theory of linear prediction (LP) is based on a speech model which considers speech production process as a convolution of source signal (e.g., periodic impulse train for voiced speech or random noise for unvoiced speech) with the time-varying vocal tract filter. In case of MFCC, we warp the speech spectrum into Mel frequency scale to mimic the human perception process. This Mel frequency warping is done by *multiplying* the magnitude speech spectrum for a preprocessed frame by magnitude of triangular filters in Mel filterbank. Since the multiplication in frequency domain corresponds to the *convolution* in time, so this process of Mel frequency warping is based on linear

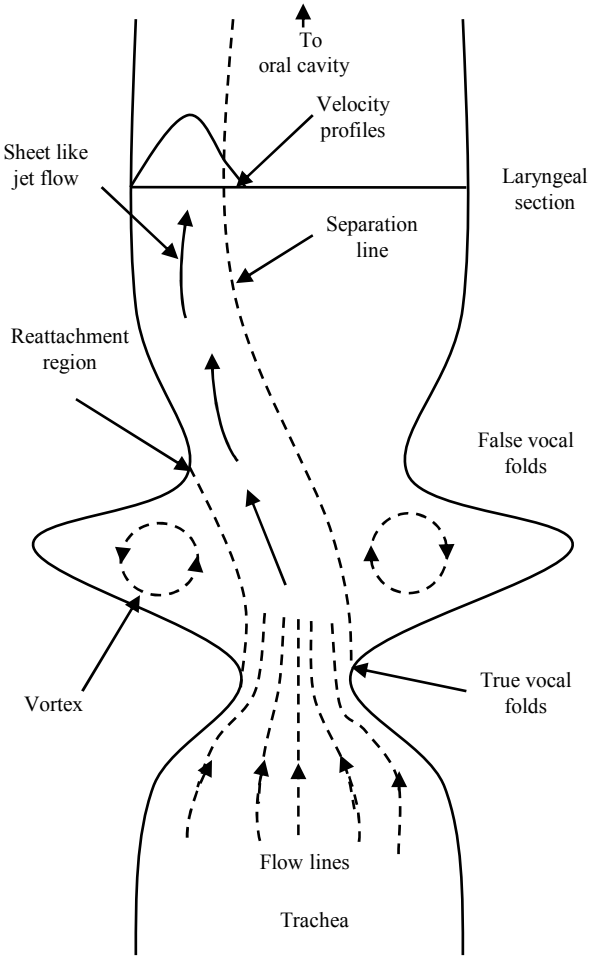


Fig.1. Nonlinear model of sound production along the vocal tract [28].

filtering process to warp the vocal tract spectrum into Mel scale and hence MFCC extraction assumes linear speech production model. This pulsatile flow is considered the source of sound production [28]. According to Teager [28], this assumption may not hold since the flow is actually separate and concomitant vortices are distributed throughout the vocal tract (as shown in Fig. 1). He suggested that the true source of sound production is actually the vortex-flow interactions, which are non-linear and a non-linear model has been suggested based on the *energy* of airflow. Fig. 2 shows Teager's original investigations about distinct flow pattern of vowel *'i'* at top and bottom rear of the front oral cavity (due to the non-linear airflow) [26].

The human speech production process can be modeled by two broad ways. One approach is to model the vocal tract structure using a source-filter model. This approach assumes that the underlying source of speaker's identity is coming from the vocal tract configuration of the articulators (i.e., size and shape of the vocal tract) and the manner in which speaker uses his articulators in sound production [1] and [15]. An alternative way to characterize speech production is to model the airflow pattern in the vocal tract. The underlying concept here, is that while the vocal tract articulators do move to configure the vocal tract shape (making cues for speaker's identity [11]), it is the resulting airflow properties which serve to excite those models which a listener will perceive for a particular speaker's voice [26] and [28]. Modeling the time-varying vortex flow is a formidable task and Teager devised a simple algorithm which uses a non-linear energy-tracking operator called as Teager Energy Operator (TEO) (in discrete-time) for signal analysis with the supporting observa-

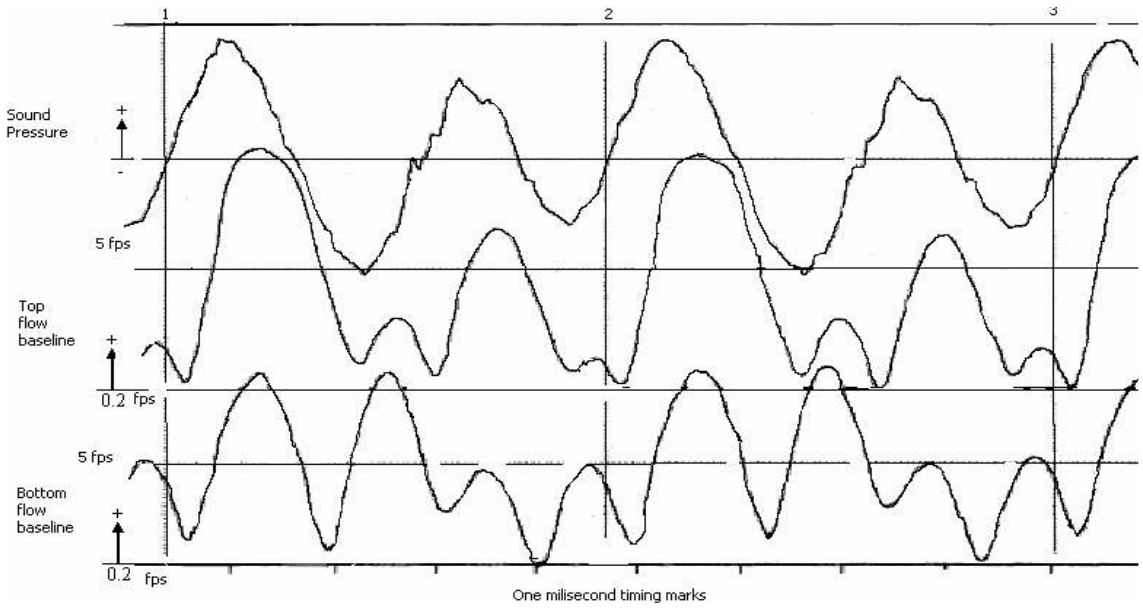


Fig. 2. Representative simultaneous normalized sound and air flow for the vowel *"i"*. Top trace: sound pressure. Middle trace: airflow velocity measured by anemometers at the top rear of the front oral cavity. Bottom trace: air flow velocity measured at the bottom rear of the front oral cavity. (After Teager [26]).

tion that hearing is the process of detecting energy. The concept was further extended to continuous-domain by Kaiser [9].

The reasoning given by Kaiser is as follows: When one speaks about the 'energy' in a signal the usual tendency is to talk about the average of the sum of the squares of the magnitude of that signal as energy required to generate that signal. This is often the case in speech processing literature where the energy in a speech frame is calculated in this way and hence by this approach of calculating the energy, a unit 40 Hz signal is said to have a *same energy* as a unit 4000 Hz. However the energy required to generate the acoustic signal of 40 Hz is much less than that for the 4000 Hz. In order to understand the difference between these energies, let us focus on the signal generation process and then the energy required to generate it.

A. Signal Generation Process and Teager's Algorithm

For the mass-spring system shown in Fig. 3, the dynamics are described by

$$\frac{d^2x}{dt^2}+\left(\frac{k}{m}\right)x=0$$

whose solution is a simple harmonic motion (S.H.M.) given by

$$x(t)=A\cos(\Omega t+\phi) \tag{1}$$

The above solution can also be justified in the following way. Any periodic function can be decomposed into Fourier series which is an aggregate of an infinite set of sinusoids. The general solution of a second order linear differential equation of this type with positive values of k and m is $\exp\left(\pm j\sqrt{k/m}\right)$ which leads finally to the form $x(t)=A\cos(\Omega t+\phi)$ and the energy is given by

$$\begin{aligned} E&=\frac{1}{2}kx^2+\frac{1}{2}m\left(\frac{dx}{dt}\right)^2=\frac{1}{2}m\Omega^2A^2 \\ &\Rightarrow E\propto\left(A\Omega\right)^2 \end{aligned} \tag{2}$$

From eq. (2), it is clear that the energy of the S.H.M. of displacement signal $x(t)$ is directly proportional not only to the square of the amplitude of the signal but also to the square of the frequency of the signal. Another motivation for considering the S.H.M. for understanding TEO for speech processing, is due to the recent proposal of Maragos *et al.* [16,17] that speech can be modeled as a linear combination of AM-FM signals (within one pitch period) which will be discussed very shortly. Kaiser and Teager proposed the algorithm to calculate the running estimate of the energy content in the signal. Eq. (1) can be expressed in discrete-time domain as

$$x(n)=A\cos(\omega n+\phi) \tag{3}$$

There are three unknown parameters in eq. (3) and therefore the solution will require three samples of signal $x(n)$. Let us consider three adjacent samples of $x(n)$ given by,

$$\begin{aligned} x(n)&=A\cos(\omega n+\phi) \\ x(n+1)&=A\cos(\omega(n+1)+\phi) \\ x(n-1)&=A\cos(\omega(n-1)+\phi) \end{aligned}$$

By trigonometry,

$$x^2(n)-x(n+1)x(n-1)=A^2\sin^2\omega\approx A^2\omega^2\approx E_n$$

where E_n gives the running estimate of signal's energy. In continuous-time, TEO of a signal $x(t)$ is defined by

$$\Psi_c[x(t)]=\left[\frac{dx}{dt}\right]^2-x(t)\frac{d^2x}{dt^2}$$

and can be approximately discretized with the normalized sampling time as

$$\Rightarrow\Psi_c[x(t)]\mapsto\Psi_d[x(n)]=x^2(n)-x(n+1)x(n-1) \tag{4}$$

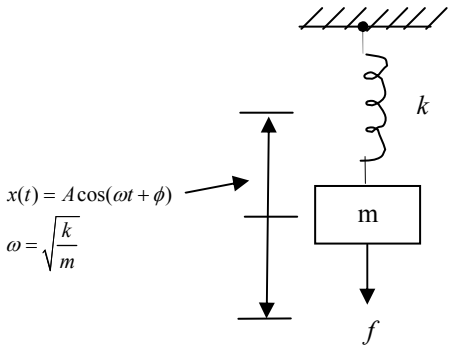


Fig. 3. S.H.M. of an undamped oscillator

TEO has the following properties [10]

1) E_n is independent of the initial phase in the signal.
2) E_n is symmetric.
3) Algorithm is robust even if the signal passes through zero, i.e, $x(n)=0$ as there is no division by $x(n)$ involved.
4) The algorithm is capable of responding very rapidly (in two sampling instants) to changes in both A and ω.

It is shown in [16] that the speech can be modeled as a linear combination of AM-FM signals in some cases. Each resonance or formant is represented by an AM-FM signal of the form

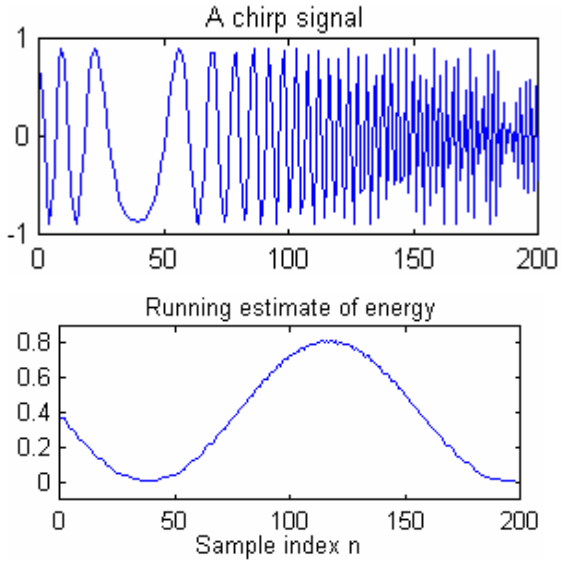
$$\begin{aligned} x(t)&=a(t)\cos(\phi(t))=a(t)\cos[\int_0^t\omega_i(\tau)d\tau+\phi(0)] \\ &\Rightarrow\Psi_c[x(t)]\approx\left(a\frac{d\phi}{dt}\right)^2 \end{aligned} \tag{5}$$

where $a(t)$ is a time varying amplitude signal and $\omega_i(t)$ is the instantaneous frequency given by $\omega_i(t)=d\phi/dt$. This model allows the amplitude and formant frequency (resonance) to vary instantaneously within one pitch period. In [17], it is shown that TEO can track the modulation energy and identify the instantaneous amplitude and frequency. For example, Fig. 4 shows the application of TEO algorithm for a chirp signal. It is evident from the figure that, the algorithm is able to track very efficiently the *instantaneous frequency* of the signal. Thus, on the whole,the motivation for using S.H.M. model was first to understand a concrete process of signal generation in a physical sense and then to find the expression for the energy required to generate that signal. Moreover, this SHM model is considered as an AM-FM model for one speech resonance and we can get information about the instantaneous amplitude and the formant frequency in a speech frame [28].

Motivated by this fact, in this paper a new feature set based on nonlinear model of (5) is developed using the TEO for LID problem. The idea of using TEO instead of the commonly used instantaneous energy is to take advantage of the modulation energy tracking capability of the TEO. This leads to a better representation of formant information in the feature vector than MFCC [7] and [8]. Recent applications of TEO in speech literature can be found in [7], [8], [14], and [28]. In the next section, we will discuss the details of T-MFCC.

III. MFCC vs. T-MFCC

For a particular speech sound in a *language*, the *human perception process* responds with better frequency resolution to lower frequency range and relatively low frequency resolution in high frequency range. To mimic this process MFCC is developed. For computing MFCC, we warp the speech spectrum into Mel frequency scale. This Mel frequency warping is done by multiplying the magnitude of speech spectrum for a preprocessed frame by magnitude of triangular filters in Mel filterbank followed by log-compression of sub-band *energies* and finally DCT.



(a) (b)

Fig. 4. Chirp signal and its running estimate of energy by using TEO.

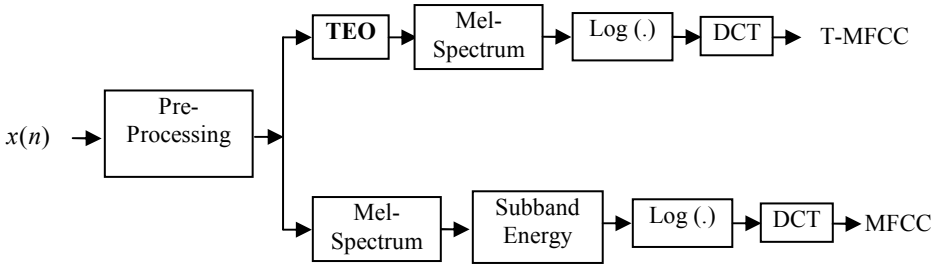


Fig. 5. Functional block diagram for T-MFCC and MFCC implementation.

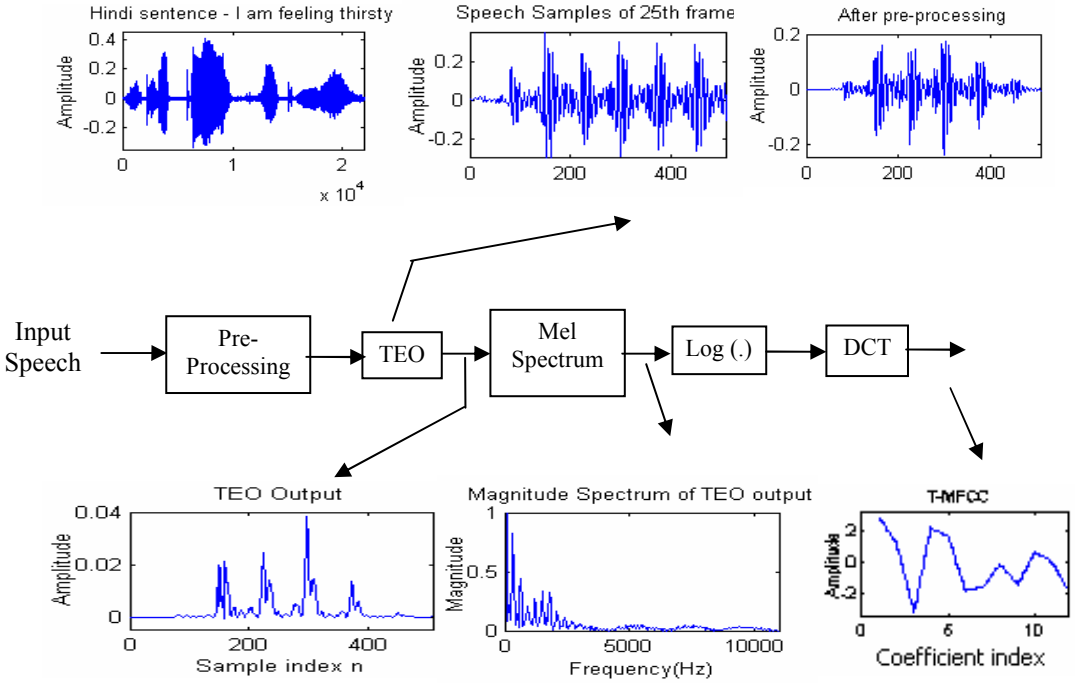


Fig. 6. Computation of T-MFCC for a voiced speech frame of Hindi sentence, *viz.,* 'Mujhe *pyas lagi hai'*.

Davis and Mermelstein proposed one such filterbank to simulate this in 1980 for speech recognition application [6]. The frequency spacing of the filters used in Mel filterbank is kept as linear up to 1 kHz and logarithmic after 1 kHz. The frequency spacing is designed to simulate the *subjective spectrum* from *physical spectrum* to emphasize the human perception process. *Thus, MFCC can be a potential feature to identify perceptually distinct languages (because for perceptually similar languages there will be confusion in MFCC due to its dependence on human perception process for hearing).*

In our approach, we employ Teager Energy Operator (TEO) for calculating the energy of speech signal. Now, one may apply TEO in frequency domain, i.e., TEO of each subband at the output of Mel-filterbank, but there is difficulty from implementation point of view. Let us discuss this point in detail. In frequency-domain, eq. (4) for pre-processed speech $x_p(n)$ implies,

$$F\left\{\Psi_c[x_p(t)]\right\} \mapsto F\left\{{x_p}^2(n) - x_p(n+1)x_p(n-1)\right\} = F\left\{{x_p}^2(n)\right\} - F\left\{x_p(n+1)x_p(n-1)\right\} \tag{6}$$

Using shifting and multiplication property of Fourier transform, we have

$$F\left\{{x_p}^2(n)\right\} = \frac{1}{2\pi}\int_{2\pi} X_p(\theta)X_p(\omega-\theta)d\theta$$

$$F\left\{x_p(n+1)x_p(n-1)\right\} = \frac{1}{2\pi}\int_{2\pi} X_{1p}(\theta)X_{2p}(\omega-\theta)d\theta$$

where $X_{1p}(\omega) = e^{-j\omega}X_p(\omega)$ and $X_{2p}(\omega) = e^{j\omega}X_p(\omega)$. Hence eq. (6) becomes

$$F\left\{\Psi_c[x_p(t)]\right\} \mapsto \frac{1}{2\pi}\int_{2\pi}\left(1 - e^{j\omega}e^{-2\theta}\right)X_p(\theta)X_p(\omega-\theta)d\theta \tag{7}$$

It is difficult to implement eq. (7) in discrete-time and it is also time-consuming. So we have applied TEO in the time-domain. Let us now see the computational details of T-MFCC. The speech signal $x(n)$ is first passed through pre-processing stage (which includes frame blocking, hamming windowing and pre-emphasis) to give pre-processed speech signal $x_p(n)$. Next we calculate the Teager energy of $x_p(n)$:

$$\Psi_d[x_p(n)] = {x_p}^2(n) - x_p(n+1)x_p(n-1) = \psi_1(n)(say)$$

The magnitude spectrum of the TEO output is computed and warped to Mel frequency scale followed by usual log and DCT computation (of MFCC) to obtain T-MFCC.

$$T-MFCC = \sum_{l=1}^{L}\log\left[\Psi_1(l)\right]\cos\left(\frac{k(l-0.5)}{L}\pi\right), k = 1,2,....,Nc.$$

where $\Psi_1(l)$ is the filterbank output of $F\{\psi_1(n)\}$ and $\log\left[\Psi_1(l)\right]$ is the log of filterbank output and $T-MFCC(k)$ is the k^{th} T-MFCC. Fig. 5 shows functional block diagram of T-MFCC and MFCC implementation. Fig. 6 shows output at different stages for computing T-MFCC for a voiced speech frame of a Hindi sentence, *viz., 'Mujhe Pyas Lagi Hai'* (i.e., I am feeling thirsty). It is evident from the TEO output of pre-processed speech frame that TEO algorithm tracks running estimates of time-varying energy in speech frame.

T-MFCC differs from the traditional MFCC in the definition of *energy measure*, i.e., MFCC employs L^2 energy in frequency domain (due to *Parseval's equivalence*) at each subband whereas T-MFCC employs Teager energy in time domain and determines the spectrum.

IV. SETUP

In this section, brief details of the setup used for this study, *viz.*, data collection and corpus design and different speech features is given. Earlier studies for corpus development in speaker recognition are given in [2].

A. Data Collection and Corpus Design

Database of 120 subjects (60 in each of Hindi and Urdu) is prepared from the different states of India, *viz*., Maharashtra, Uttar Pradesh and West Bengal with the help of a voice activated tape recorder (*Sanyo* model no. M-1110C & *Aiwa* model no. JS299) with microphone input, a close talking microphone (*viz*., *Frontech* and *Intex*). The data is recorded on the Sony high fidelity voice and music recording cassettes (C-90HFB). A list consisting of five questions, isolated words, digits, combination-lock phrases, read sentences and a contextual speech of considerable duration was prepared. The contextual speech consisted of description of nature or memorable events etc. of community or family life of the subject. The data was recorded with 10 repetitions except for the contextual speech. During recording of the contextual speech, the interviewer asked some questions to subject in order to motivate him or her to speak on his or her chosen topic. This also helps the subject to overcome the initial nervousness and come to his or her natural mode so that the acoustic characteristics of his or her language are tracked precisely. During recording, both the interviewer and the subject were able to see and interact with each other. The subject's voice and interviewer's voice were recorded on the same track. Once the magnetic tape was played into the computer, the subject's voice was played again to check the wrong editing. The interviewer's voice was deleted from the speech file so that there would not be any biasing for a particular language. The automatic silence detector was employed to remove the silence periods in the speech recordings to get only the language model for the speaker's voice and not the background noise and silence interval. Also, each subject's voice was normalized by the peak value so that the speech amplitude level could be normalized for all the subjects in a particular language. Finally, the corpus was designed into training segments of 30 s, 60 s, 90 s and 120 s durations and testing segments of 1 s, 3 s, 5 s, 7 s, 10 s, 12 s and 15 s durations in order to find the performance of the LID system for various training and testing durations. Table I shows the details of the corpus. Other details of the experimental setup and data collection are given in [22]. Following are the salient features of our corpus:

1) wide varieties of acoustic environments were considered during recording ranging from office-roads-train, to noisy workstations, etc. which added *realistic acoustic noise*

TABLE I
DATABASE DESCRIPTION FOR LID SYSTEM

Item	Details
# subjects	120 (60 in each of Hindi and Urdu)
# sessions	1
Data type	Speech
Sampling rate	22,050 Hz
Sampling format	1-channel, 16 bit resolution
Type of speech	Read sentences, isolated words and digits, combination-lock phrases, questions, contextual speech of approximately 30 sec duration
Application	Text-independent language identification system
Training language	Hindi, Urdu.
Testing language	Hindi, Urdu.
# repetitions	10 except for contextual speech.
Training segments	30 s, 60 s, 90 s, 120 s.
Test segments	1 s, 3 s, 5 s, 7 s, 10 s, 12 s, 15 s.
Microphone	Close talking microphone
Recording Equipment	Sanyo Voice Activated System (VAS: M-1110C), Aiwa (JS299),Panasonic magnetic tape recorders
Magnetic tape	Sony High-Fidelity (HF) voice and music recording cassettes
Channels	EP to EP Wire
Acoustic environment	Home/slums/college/remote villages/roads

(e.g., crosstalk, burst, spurious noise activity, traffic noise, etc.) to the data. This is the most important feature of our corpus;

2) speech units including specific vocabularies, e.g., isolated words, digits, combination-lock phrases, read sentences, question-answer session, and contextual speech/spontaneous speech of considerable duration with varying topics were considered in the recording which added realistic situations in the speech data;
3) data was not recorded in closed booth (research lab/office), where the speaker might not have felt free to give his/her speech data;
4) speakers of wide ranging ages (15-80 years) and a variety of educational backgrounds (from illiterate to university post graduates) have been considered in the corpus;
5) database is prepared from voluntary *subjects* and hence their natural mode of excitement was not altered.

Following are some of our practical experiences during recording:

1) the presence of interviewer, recording equipment or any other tool of measurement affects the natural acoustical characteristics for a particular language (i.e., *Lobov's observer's paradox* [5]);
2) subjects unconsciously talk louder in front of microphone (i.e., *Lombard effect* [13]);
3) some initial resistance was experienced from a small section of native speakers in Marathi for recording of Hindi language;
4) speakers occasionally became bored or distracted, and lowered their voice intensity or turned their heads away from the microphone;
5) there was stammering, laughter, throat clearing, tittering and poor articulation. All these cases were recorded in normal fashion.

B. Polynomial Classifier

Till recently, Gaussian Mixture Model (GMM) [27], Support Vector Machine (SVM) [4] and Autoassociative Neural Network (AANN) [18]-based techniques were applied to the LID problem. Although the GMM-based approach has been successfully employed for speaker recognition, its performance in the task of identification of language has been inferior compared to that of phone-based approaches [25]. In this paper, the problem of LID is viewed from the standpoint of speaker recognition and a new method for classier design is suggested for LID problem by *modifying* the structure of the polynomial networks for preparing language models for different Indian languages, *viz.,* Hindi and Urdu this method of classifier design as *Modified Polynomial Networks* (MPN). The detailed discussion of MPN technique is beyond the scope of the paper and it is reported in [23]. In this sub-section, we will briefly discuss computational details of this technique.

Due to Weierstrass-Stone approximation theorem, polynomial classifiers are universal approximators to the optimal Bayes classifier [3]. Feature vectors are processed by the polynomial discriminant function. Every speaker j has $\mathbf{w}_j$ as his model, and the output of a discriminant function is averaged over time resulting in a score for every $\mathbf{w}_j$ [3]. The score for j^{th} speaker is then given by,

$$S_j = \frac{1}{M}\sum_{i=1}^{M} \mathbf{w}_j^T \mathbf{p}(\mathbf{x_i})$$

where $\mathbf{x_i} = i^{th}$ input test feature vector
$\mathbf{w}$ = speaker model
$\mathbf{p(x)}$ = vector of polynomial basis terms of the input test feature vector.
M = total number of testing feature vectors,

Training polynomial classifier is accomplished by obtaining the optimum speaker model for each speaker using discriminatively trained classifier with mean-squared error (MSE) criterion, i.e., for speaker's feature vector, an output of one is desired, whereas for impostor data an output of zero is desired. For the two-class problem, let $\mathbf{w_{spk}}$ be the optimum speaker model, ω class label, and $y(\omega)$ the ideal output, i.e., $y(spk) = 1$ and $y(imp) = 0$.
The resulting problem using MSE is

$$\mathbf{w_{spk}} = \arg\min_{\mathbf{w}} E\left\{\left(\mathbf{w^T p(x)} - \mathbf{y(\omega)}\right)^2\right\} \qquad (8a)$$

where $E\{.\}$ means expectation over $\mathbf{x}$ and ω. This can be approximated using training feature set as

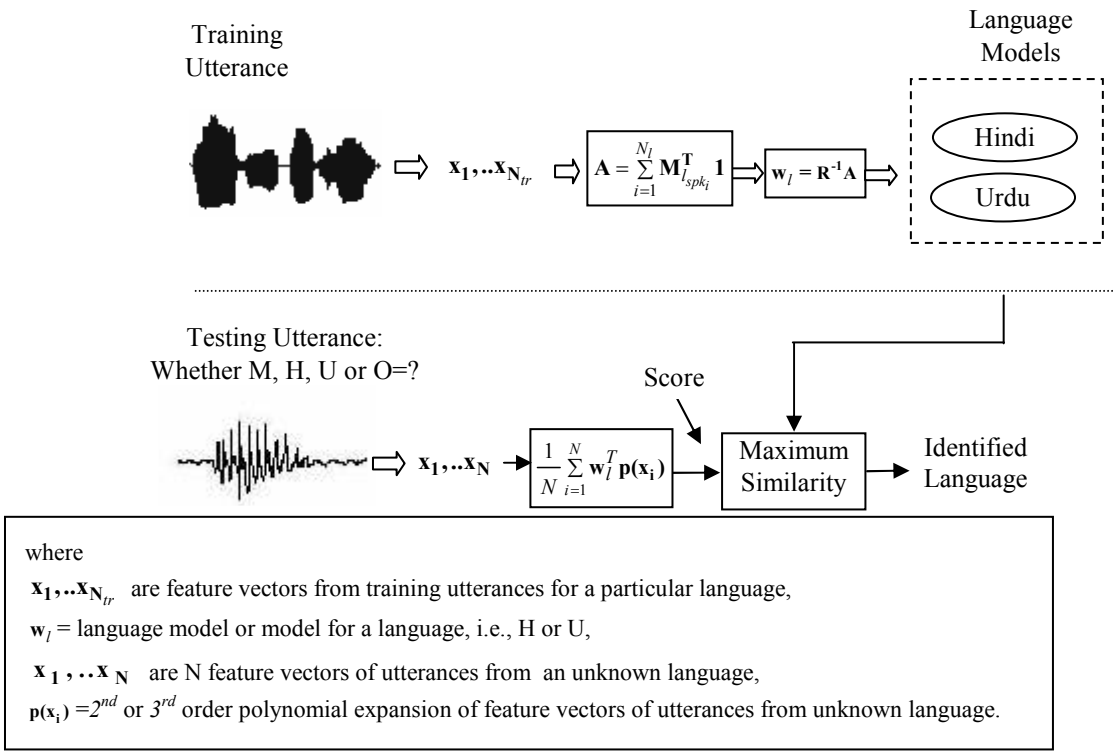


Fig. 7. LID system architecture

$$\mathbf{w_{spk}}=\arg\min_{\mathbf{w}}\left[\sum_{i=1}^{N_{spk}}\left|\mathbf{w^T p(x_i)}-1\right|^2+\sum_{i=1}^{N_{imp}}\left|\mathbf{w^T p(y_i)}\right|^2\right] \quad (8b)$$

where $\mathbf{x_1,..x_{N_{spk}}}$ are speaker's training data and $\mathbf{y_{1,.....,y_{N_{imp}}}}$ is the impostor data. This training algorithm can be expressed in matrix form. Let $\mathbf{M_{spk}}=[\mathbf{p(x_1)}\ \ \mathbf{p(x_2)}\ \ ....\ \ \mathbf{p(x_{N_{spk}})}]$ and similar matrix for $\mathbf{M_{imp}}$. Also let $\mathbf{M}=[\mathbf{M_{spk}}\ \ \mathbf{M_{imp}}]$ and thus the training problem in eq. (8) is reduced to the well-known linear approximation problem in normed space as

$$\mathbf{w}_{spk}=\arg\min_{w}\left\|\mathbf{Mw}-\mathbf{o}\right\|_2,$$

where $\mathbf{o}$ consisting of N_{spk} ones followed by N_{imp} zeros. This problem can be solved using the method of normal equations

$$\mathbf{M^T M w_{spk}}=\mathbf{M^T o}$$

which after rearranging gives

$$(\mathbf{M_{spk}^T M_{spk}}+\mathbf{M_{imp}^T M_{imp}})\mathbf{w_{spk}}=\mathbf{M_{spk}^T 1} \quad (9)$$

where $\mathbf{1}$ is the vector of all ones. Now we define $\mathbf{R_{spk}}=\mathbf{M_{spk}^T M_{spk}}$ and define $\mathbf{R_{imp}}$ similarly, then eq. (9) can be written as

$$(\mathbf{R_{spk}}+\mathbf{R_{imp}})\mathbf{w_{spk}}=\mathbf{M_{spk}^T 1}$$

Also define $\mathbf{R}=\mathbf{R_{spk}}+\mathbf{R_{imp}}$ and $\mathbf{A_i}=\mathbf{M_{spk_i}^T 1}\Rightarrow\mathbf{A}=\sum_{i=1}^{Lspk}\mathbf{M_{spk_i}^T 1}$

$$\mathbf{w}_{Lspk}=\mathbf{R}^{-1}\mathbf{A} \quad (10)$$

where $\mathbf{w}_{Lspk}$ is the optimum language model and $Lspk$ is the number of speakers in each language class (60 in present problem). Fig. 7 shows the structure of polynomial classifier for LID problem. One of the advantages of training algorithm, i.e., eq. (10), is that optimum language model, *viz.*, $\mathbf{w}_{Lspk}$ does not depend upon the duration of the training speech but it is the *length* of the feature vector which predominantly determines the computational load on the machine. The details of training algorithm for multi-class problem, polynomial basis determination and mapping algorithm based semi-group isomorphism property of monomials for computing unique terms in $\mathbf{R_{spk}}$ (and hence $\mathbf{R}$) are given in [3] and [23].

V. Experimental Results

In this paper, polynomial classifier of *2nd* and *3rd* order approximation is used as the basis for all the experiments. Feature analysis was performed using a 23.2 ms speech frame with an overlap of 50%. Each frame was pre-emphasized with the filter $1\text{-}0.97z^{-1}$, followed by Hamming window. After this pre-processing, different speech features, *viz.*, MFCC and T-MFCC discussed in section III were extracted. The results are shown in Tables II-V as average success rates (average is computed over testing segments of 1 s, 3 s, 5 s, 7 s, 10 s, 12 s, and 15 s durations) with various training durations for LID experiments. In this work, the success rates are defined as

$$SR=\frac{N_c}{N_t}\times 100,$$

where N_c is the number of correctly identified language class/testing segments for a particular language class and N_t is the total number of speakers used for machine learning for a particular language class .

Tables V and VI show confusion matrices (diagonal elements indicate % correct identification in a particular linguistic group and off-diagonal elements show the

misidentification) for 2 languages with MFCC and T-MFCC, respectively. In Tables IV and V, ACT represents the actual language of the speaker and IDENT represents the identified language of an unknown speaker. The performance of a confusion matrix is evaluated based on its diagonal and off-diagonal entries meaning a confusion matrix will be ideal, i.e., all the testing samples are correctly identified to their respective classes; if all the off-diagonal elements are zero and diagonal elements are 100. *Any deviation from this will judge the relative performance of the confusion matrix*). *Thus the confusion matrix indicates the effectiveness of the proposed MPN model in capturing language-specific information with the help spectral features*.

TABLE II
AVERAGE SUCCESS RATES FOR 2 LANG. (H& U) WITH 2ND ORDER APPROXIMATION

TR / FS	30S	60S	90S	120S
MFCC	21.42	22.97	23.57	23.691
T-MFCC	**41.42**	**42.262**	**42.738**	**42.143**

TABLE III
AVERAGE SUCCESS RATES FOR 2 LANG. (H& U) WITH 3RD ORDER APPROXIMATION

TR / FS	30S	60S	90S	120S
MFCC	12.73	20.95	21.66	22.5
T-MFCC	**44.76**	**42.97**	**43.21**	**43.33**

TABLE IV
AVERAGE SUCCESS RATES FOR 2 LANG. (M & H) WITH 2ND ORDER APPROXIMATION

TR / FS	30S	60S	90S	120S
MFCC	**62.97**	**67.02**	**68.09**	**67.97**
T-MFCC	55.83	57.85	58.21	56.42

TABLE V
AVERAGE SUCCESS RATES FOR 2 LANG. (M& H) WITH 3RD ORDER APPROXIMATION

TR / FS	30S	60S	90S	120S
MFCC	55.47	**64.04**	**64.64**	**64.76**
T-MFCC	53.92	56.42	57.14	56.90

Some of the observations from the results are as under:

1) For both *2nd* order and *3rd* order polynomial approximation, T-MFCC outperformed MFCC in all the cases of training speech durations. This may be due to the fact that MFCC is known to be developed to mimic human perception process and since the present problem deals with identification of perceptually similar languages (i.e., confusion in perception of phonemes of two languages, *viz.*, Hindi (H) and Urdu (U)), MFCC gets confused in discriminating the language-specific features. On the other hand, T-MFCC represents the combined effect of airflow properties in the vocal tract (which are known to be language and speaker dependent [21]) and human perception process. So, T-MFCC is able to capture the speaker and language -specific information better than MFCC.
2) On the other hand, for both *2nd* order and *3rd* order polynomial approximation and identification of perceptually *distinct* languages (i.e., Marathi (M) and Hindi), MFCC outperformed T-MFCC.
3) There is a significant improvement in the performance of T-MFCC for *3rd* order approximation as compared to the *2nd* order approximation. This is quite expected for a classifier of higher order polynomial approximation.
4) Average success rates increase with the increase in training speech durations.
5) Confusion matrix for T-MFCC performed better than MFCC. This shows that T-MFCC has better *class discrimination* power than MFCC.

TABLE VI
CONFUSION MATRIX WITH 2ND ORDER APPROXIMATION FOR MFCC (TR=120S AND TE=15S) WITH 2 LANG.

ACT. / IDENT.	H	U
H	**85.556**	14.444
U	76.667	**23.333**

TABLE VII
CONFUSION MATRIX WITH 2ND ORDER APPROXIMATION FOR T-MFCC (TR=120S AND TE=15S) WITH 2 LANG.

ACT. / IDENT.	H	U
H	**71.111**	28.889
U	5.5556	**94.444**

VI. SUMMARY AND CONCLUSIONS

In this paper, an LID system is presented to demonstrate the effectiveness of newly proposed feature sets, *viz.*, T-MFCC for identifying perceptually similar languages with the database prepared in real-world scenario. The novelty of the proposed method consists of building up language models by capturing airflow properties of the vocal tract and human perception process to recognize phonemes of a particular language.

ACKNOWLEDGEMENT

The authors would like to thank authorities of DA-IICT Gandhinagar and IIT Kharagpur for their support for this work. They also thank to those people of India who have given their kind support and cooperation during data collection phase.

REFERENCES

[1] B. S. Atal, "Effectiveness of linear prediction of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.* vol. 55, pp. 1304-1312, 1974.

[2] J. P. Campbell, Jr. and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing, ICASSP'99,* vol. 2, pp. 829-832, March 15-19, 1999.

[3] W. M. Campbell, K. T. Assaleh and C. C. Broun, "Speaker recognition with polynomial classifiers," *IEEE Trans. on Speech and Audio Processing,* vol. 10, no.4, pp. 205-212, May 2002.

[4] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo and D. A. Reynolds, "Language recognition with support vector machines," *Proc. Odyssey: The Speaker and Language Recognition Workshop,* Toledo, Spain, ISCA, pp. 41-44, 2004.

[5] C. G. Clopper *et al*., "A multi-talker dialect corpus of spoken American English: An initial report," *Research on Spoken Language Processing, Progress Report,* Bloomington, Speech Research Laboratory, Indiana University, vol. 24, pp. 409-413, 2000.

[6] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech and Signal Processing,* vol. ASSP-28, no.4, August 1980.

[7] F. Jabloun and A. E. Cetin, "The Teager energy based feature parameters for robust parameters in car noise," in *Proc. of the Acoustics, Speech, and Signal Processing,* vol. 1, pp. 273-276, 1999.

[8] F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," *IEEE Signal Processing Lett.,* vol. 6, pp. 259-261, 1999.

[9] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. of Int Conf. on Acoustic, Speech and Signal Processing*, vol. 1, pp. 381-384, 1990.

[10] J.F. Kaiser, "Some useful properties of Teager's energy operator.," i*n Proc. of the Int. Conf. on Acoust., Speech, and Signal Processing*, vol. 3, pp. 149-152, 1993.

[11] L.G. Kersta, "Voiceprint Identification," *Nature,* vol. 196, pp. 1253-1257, 1962.

[12] K-P Li, "Automatic language identification using syllabic spectral features", *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, ICASSP,* vol. 1, pp. 297-300, 1994.

[13] E. Lombard, Le Signe de l'Elevation de la Voix. Ann. Maladies Oreille, Larynx, Nez, *Pharynx,* vol. 37, pp. 101–119, 1911.

[14] C-T. Lu and H-C Wang, "Enhancement of single channel speech based on masking property and wavelet transform," *Speech Communication,* vol. 41, pp. 409-427, 2003.

[15] R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust speaker recognition-A feature based approach*," IEEE Signal Processing Mag.* vol. 13, pp. 58-71, 1996.

[16] P. Maragos, T. Quatieri and J. F. Kaiser, "Speech non-linearities, modulation and energy operators," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* Toronto, ON, Canada, pp. 421-424, 1991.

[17] P. Maragos, T. Quatieri and J. F. Kaiser, "On amplitude and frequency demodulation using energy operators," *IEEE Trans. on Signal Processing,* vol. 41, pp. 1532-1550, 1993.

[18] L. Mary and B. Yegnanarayana, "Autoassociative neural network models for language identification", *Int. Conf. on Intelligent Sensing and Information Processing, ICISIP*, pp. 317-320, 2004.

[19] Y. K. Muthusamy, E. Barnard and R. A. Cole, "Reviewing automatic language identification", *IEEE Signal Processing Mag.,* vol. 11, pp. 3341, 1994.

[20] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd edition, Universities Press, 2001.

[21] Hemant A. Patil and T. K. Basu, "Detection of bilingual twins by Teager energy based features," in *Proc. Int. Conf. Signal Processing and Commun., SPCOM'04,* IISc, Bangalore, India, pp. 32-36, Dec. 11-14, 2004.

[22] Hemant A. Patil, *Speaker recognition in Indian languages: A feature based approach*, Ph.D. Thesis, Department of Electrical Engineering, IIT Kharagpur, India, July 2005.

[23] Hemant A. Patil and T. K. Basu, "A novel approach to language identification using modified polynomial networks," to appear in *Speech, Audio, Image and Biomedical Signal Processing Using Neural Networks, B. Prasad and S R M Prasanna (eds.)*, Springer-Verlag, Heidelberg, Germany, 2007.

[24] A. E. Rosenberg, "Automatic Speaker Verification: A review," *Proc. of* IEEE*,* vol. 64, pp. 475-487, 1976.

[25] E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell and D.A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification", *Proc. Eurospeech*, Geneva, Switzerland, ISCA, pp. 1345-1348, 2003.

[26] H.M. Teager, "Some observations on oral air flow during phonation," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 28, pp. 599-601, 1980.

[27] P. A. Torres-Carrasquillo, D. A. Reynolds and J. R. Deller Jr., "Language identification using Gaussian mixture model tokenization", *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, ICASSP*, Orlando, FL, vol. I, pp. 757-760, 2002.

[28] G. Zhau, J.H.L. Hansen and J. F. Kaiser, "Non-linear feature based classification of speech under stress," *IEEE Trans. on Speech and Audio Processing,* vol. 9, pp. 201-216, 2001.

[29] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech", *IEEE Trans. Speech and Audio Processing,* vol. 4, pp. 31-44, 1986.