Cepstral Domain Teager Energy for Identifying Perceptually Similar Languages

Hemant A. Patil[1] and T.K. Basu[2]

[1] Dhirubhai Ambani Institute of Information and Communication Technology, DA-IICT , Gandhinagar, Gujarat, India
hemant_patil@daiict.ac.in
[2] Department of Electrical Engineering, Indian Institute of Technology, IIT Kharagpur, West Bengal, India
tkb@ee.iitkgp.ernet.in

Abstract. Language Identification (LID) refers to the task of identifying an unknown language from the test utterances. In this paper, a new feature set, *viz.,*T-MFCC by amalgamating Teager Energy Operator (TEO) and well-known Mel frequency cepstral coefficients (MFCC) is developed. The effectiveness of the newly derived feature set is demonstrated for identifying perceptually similar Indian languages such as Hindi and Urdu. The modified structure of polynomial classifier of 2^{nd} and 3^{rd} order approximation has been used for the LID problem. The results have been compared with state-of-the art feature set, *viz.,*MFCC and found to be effective (an average jump 21.66%) in majority of the cases. This may be due to the fact that the T-MFCC represents the combined effect of airflow properties in the vocal tract (which are known to be language and speaker dependent) and human perception process for hearing.

1 Introduction

Language Identification (LID) refers to the task of identifying an unknown language from the test utterances. LID applications fall into two main categories: pre-processing for machine understanding systems and preprocessing for human listeners. Alternatively, an LID system could be run in advance of the speech recognizer. Alternatively, LID might be used to route an incoming telephone call to a human switchboard operator fluent in the corresponding language [6]. Several techniques such spectral, prosody, phoneme, word-level, etc. have been proposed in the literature for LID problem. In this paper, we adopt spectral-based approach [5] and show the effectiveness of the newly derived feature set,*viz.,*Teager Energy based Mel Frequency Cepstral Coefficients (T-MFCC) for identification of perceptually similar Indian languages, *viz.,*Hindi and Urdu.

2 Data Collection and Corpus Design

Database of 180 speakers (60 in each of Marathi, Hindi and Urdu) is created from the different states of India, *viz.,*Maharashtra, Uttar Pradesh and West Bengal

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 455–462, 2007.

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). During recording of the contextual speech, the interviewer asked some questions to speaker in order to motivate him or her to speak on his or her chosen topic. Other details of the experimental setup and data collection are given in [7].

Table 1. Database Description for LID system

Item	Details
No. of speakers	180 (60 in each of Marathi, Hindi and Urdu)
No. of 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 considerable duration
Application	Text-independent language identification (LID) system
Training language	Marathi, Hindi, Urdu.
Testing language	Marathi, Hindi, Urdu.
No. of 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), Aiwa, 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

3 The Teager Energy Operator (TEO)

Features derived from a linear speech production models assume that airflow propagates in the vocal tract as a linear plane wave. This pulsatile flow is considered the source of sound production [9]. According to Teager [8], this assumption may not hold since the flow is actually separate and concomitant vortices are distributed throughout the vocal tract. 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.1 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)[8].

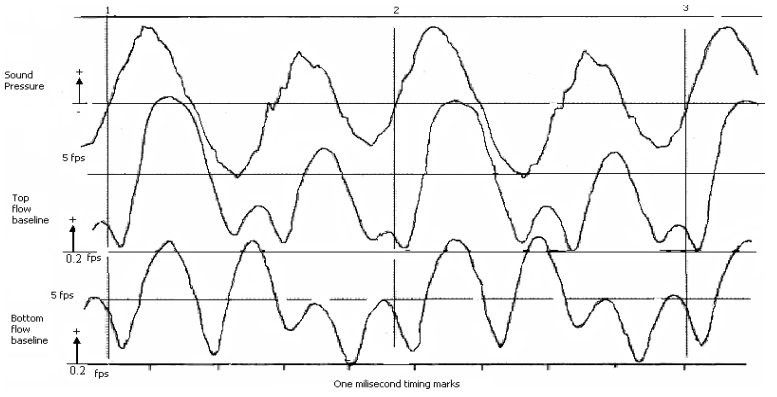

Fig. 1. Representative simultaneous normalized sound and air flow for the vowel *'i'*. Top trace: sound pressure. Middle trace: airflow velocity measured by anemometers at the top rear of the front oral cavity. Bottom trace: air flow velocity measured at the bottom rear of the front oral cavity. (After Teager [8]).

There are two broad ways to model the human speech production process. 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 [4]. 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 [4]), it is the resulting airflow properties which serve to excite those models which a listener will perceive for a particular speaker's voice [8],[9]. 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 observation that hearing is the process of detecting energy. The concept was further extended to continuous-domain by Kaiser [3]. According to Kaiser, energy in a speech frame is a function of amplitude and frequency as well. Let us now discuss this point in brief.

The dynamics and solution (which is a S.H.M.) of mass-spring system are described as

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0 \Rightarrow x(t) = A\cos(\Omega t + \phi)$$

and the energy is given by

$$E = \frac{1}{2}m\Omega^2 A^2 \Rightarrow E \propto (A\Omega)^2 \tag{1}$$

From (1), 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. Kaiser and Teager proposed the algorithm to calculate the running estimate of the energy content in the signal. (1) can be expressed in discrete-time domain as

$$x(n) = A\cos(\omega n + \phi)$$

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 and discrete-time, TEO of a signal *x(t)* is defined by

$$\Psi_c[x(t)] = \left[\frac{dx}{dt}\right]^2 - x(t)\frac{d^2x}{dt^2} \mapsto \Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1) \qquad (2)$$

It is a well known fact that the speech can be modeled as a linear combination of AM-FM signals in some cases [7],[9]. Each resonance or formant is represented by an AM-FM signal of the form

$$x(t) = a(t)\cos(\phi(t)) = a(t)\cos[\int_0^t \omega_i(\tau)d\tau + \phi_0] \Rightarrow \Psi_c[x(t)] \approx \left(a\frac{d\phi}{dt}\right)^2 \qquad (3)$$

where $a(t)$ is a 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. It is known that TEO can track the modulation energy and identify the instantaneous amplitude and frequency. Motivated by this fact, in this paper a new feature set based on nonlinear model of (3) is developed using the TEO. 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* (which is speaker and possibly language specific) in the feature vector than MFCC [7]. In the next section, we will discuss the details of T-MFCC.

4 Teager Energy Based MFCC (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 fre-quency resolution in high frequency range with the help of human ear. 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. Davis and Mermelstein proposed one such

filterbank to simulate this in 1980 for speech recognition application [2]. *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 of human perception process for hearing).* Traditional MFCC-based feature extraction involves preprocessing; Mel-spectrum of preprocessed speech, followed by log-compression of subband energies and finally DCT is taken to get MFCC per frame [2]. In our approach, we employ 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, (2) for pre-processed speech $x_p(n)$ implies,

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

$$F\left\{\Psi_c[x_p(t)]\right\} = F\left\{x_p^2(n)\right\} - F\left\{x_p(n+1)x_p(n-1)\right\} \tag{4}$$

Using shifting and multiplication property of Fourier transform, we have

$$F\left\{x_p(n+1)x_p(n-1)\right\} = \frac{1}{2\pi}\int_0^{2\pi} X_{1p}(\theta)X_{2p}(\omega-\theta)d\theta$$

where $X_{1p}(\omega) = e^{-j\omega}X_p(\omega)$ and$X_{2p}(\omega) = e^{j\omega}X_p(\omega)$. Hence (4) becomes

$$F\left\{\Psi_c[x_p(t)]\right\} = \frac{1}{2\pi}\left\{\int_0^{2\pi}\left(1 - e^{j\omega}e^{-2\theta}\right)X_p(\theta)X_p\left(\omega-\theta\right)d\theta\right\} \tag{5}$$

Thus (5) is difficult to implement in discrete-time and also time-consuming. So we have applied TEO in the time-domain. Let us now see the computational details of T-MFCC.

Speech signal $x(n)$ is first passed through pre-processing stage to give pre-processed speech signal $x_p(n)$. Next we calculate the Teager energy of $x_p(n)$:

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

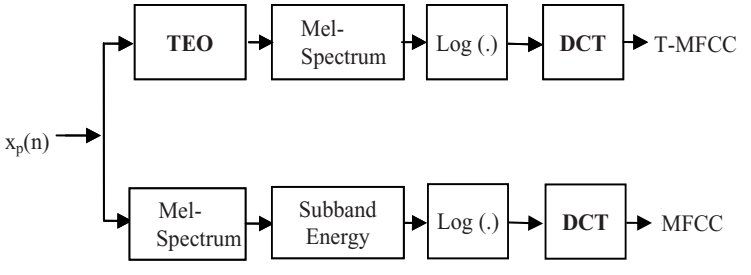

Fig. 2. Block diagram for T-MFCC and MFCC

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 $\log[\Psi(l)]$ is the log-filterbank output and T-MFCC(k) is the k^{th} feature. 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. Fig. 2 shows the functional block diagram of MFCC and T-MFCC.

5 Experimental Results

In this paper, modified polynomial classifier of 2^{nd} and 3^{rd} order approximations is used as the basis for all the experiments [1]. The detailed discussion on modified classifier structure is beyond the scope of the paper and is given in [7]. Feature analysis was performed using 23.2 ms frame with an overlap of 50% and feature dimension is kept as 12. Each frame was pre-emphasized with the filter $1 - 0.97z^{-}1$, followed by Hamming windowing and then. We have taken 2 samples more to com-pute T-MFCC than that for MFCC because of TEO processing. The experiments are performed for different testing speech durations (i.e., 1 s, 3 s, 5 s, 7 s, 10 s, 12 s and 15 s) and training speech durations (i.e., 30 s, 60 s, 90 s, and 120 s). The results are shown as average success rates (over testing speech durations) in Table 2 (for Hindi and Urdu) and Table 3 (for Marathi and Hindi). In addition to this, the results are shown as overall success rates (computed as average over testing speech durations followed by average over training speech durations) in Tables 4 and 5 for polynomial classifiers of 2^{nd} and 3^{rd} order polynomial approximation. Finally, Tables 6-7 show confusion matrices (diagonal elements indicate % correct identification in a particular linguistic group and off-diagonal elements show the misidentification) for Hindi and Urdu with MFCC and T-MFCC, respectively.

Some of the observations from the results are as follows:

- Average success rates increase with the increase in training speech durations.
- For both 2^{nd} order and 3^{rd} order polynomial approximation and identification of perceptually similar languages (i.e., Hindi and Urdu), 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 and Urdu), 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

Table 2. Average Success Rates for Hindi & Urdu with 2^{nd} Order Approximation

TRFS	30s	60s	90s	120s
MFCC	21.42	22.97	23.57	23.69
T-MFCC	41.42	42.26	42.73	42.14

Table 3. Average Success Rates for Marathi & Hindi with 2^{nd} 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 4. Overall Average Success Rates for Hindi and Urdu

OrderFS	2	3
MFCC	22.91	19.46
T-MFCC	42.14	43.56

Table 5. Overall Average Success Rates for Marathi and Hindi

Order FS	2	3
MFCC	66.51	62.22
T-MFCC	57.07	56.09

Table 6. Confusion Matrix with 2^{nd} order approximation for MFCC (TR=120 s and TE=15 s)with Hindi(H) & Urdu(U)

Ident. Act.	H	U
H	85.55	14.44
U	76.66	23.33

Table 7. Confusion Matrix with 2^{nd} order approximation for T-MFCC (TR=120 s and TE=15 s) with Hindi(H) & Urdu(U)

Ident. Act.	H	U
H	71.11	28.88
U	5.55	94.44

be language and speaker dependent [7]) and human perception process. So, T-MFCC is able to capture the speaker and language -specific information better than MFCC.
- On the other hand, for both 2^{nd} order and 3^{rd} order polynomial approximation and identification of perceptually *distinct* languages (i.e., Marathi and Hindi), MFCC outperformed T-MFCC.
- There is a significant improvement in the performance of T-MFCC for 3^{rd} order approximation as compared to the 2^{nd} order approximation. This is quite expected for a classifier of higher order polynomial approximation.
- Confusion matrix for T-MFCC performed better than MFCC. This shows that T-MFCC has better *class discrimination* power than MFCC for distinguishing perceptually similar languages.

6 Conclusion

In this paper, Teager Energy based MFCC (T-MFCC) features are proposed for identifying perceptually similar Indian languages, *viz.*, Hindi and Urdu. The performance of newly proposed feature set was compared with MFCC and found to be effective. This research work can be readily extended to identifying other perceptually similar Asian or European languages.

References

1. Campbell, W.M., Assaleh, K.T., Broun, C.C.: Speaker recognition with polynomial classifiers. IEEE Trans. on Speech and Audio Processing 10, 205–212 (2002)
2. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech and Signal Processing 28, 357–366 (1980)
3. Kaiser, J.F.: On a simple algorithm to calculate the 'energy' of a signal. Proc. of Int. Conf. on Acoustic, Speech and Signal Processing 1, 381–384 (1990)
4. Kersta, L.G.: Voiceprint Identification. Nature 196, 1253–1257 (1962)
5. Mary, L., Yegnanarayana, B.: Autoassociative neural network models for language identification. In: Int. Conf. on Intelligent Sensing and Information Processing, ICISIP, pp. 317–320 (2004)
6. Muthusamy, Y.K., Barnard, E., Cole, R.A.: Reviewing automatic language identification. IEEE Signal Processing Mag. 11, 3341 (1994)
7. Patil, H.A.: Speaker Recognition in Indian languages: A feature based approach. Ph.D. Thesis, Department of Electrical Engineering, IIT Kharagpur, India (July 2005)
8. Teager, H.M.: Some observations on oral air flow during phonation. IEEE Trans. Acoust., Speech, Signal Process. 28, 599–601 (1980)
9. Zhau, G., Hansen, J.H.L., Kaiser, J.F.: Non-linear feature based classification of speech under stress. IEEE Trans. on Speech and Audio Processing 9, 201–216 (2001)