International Journal of Engineering Sciences Paradigms and Researches, Vol. 05, Issue 01, June 2013
ISSN (Online): 2319-6564
www.ijesonline.com

Modeling Speaker Specific Features for Automatic Text Independent Speaker Tracking System using Support Vector Machines (SVMs)

V. Subba Ramaiah[1] and R. Rajeswara Rao[2]

[1]Department of Computer Science and Engineering, MGIT, Hyderabad, AP, India
subbubdl@gmail.com

[2]Department of Computer Science and Engineering, JNTU Kakinada, AP, India
Raob4u@yahoo.com

Abstract

Speaker indexing (tracking) is the process of following who says something in a given speech signal. In this paper, we propose a new set of robust prosodic features for automatic text-independent speaker indexing system. LP analysis is used to extract the prosodic information from the source speech signal. This prosodic information is speaker specific. In this approach, instead of capturing the distribution of feature vectors correspond to vocal tract system of the speakers, the time varying speaker-specific prosodic characteristics are captured using Linear Prediction (LP) residual signal of the given speech signal. MFCC features are extracted from the source speech signal, which contains prosody and speaker specific information. In this paper, we propose effective modeling of prosodic features using support vector machine.

Keywords: *Speaker Indexing, Prosodic Feature, LPC, MFCC, SVMs.*

1. Introduction

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. Speaker Recognition is basically divided into speaker identification and speaker verification [1]. A speaker identification system gets a test utterance as input. The task of the system is to find out which of the training speakers made the test utterance. So, the output of the system is the name of the training speaker, or possibly a rejection if the utterance has been made by an unknown person. Verification is the task of automatically determining if a person really is the person he or she claims to be. This technology can be used as a biometric feature for verifying the identity of a person in applications like banking by telephone, voice dialing telephone shopping, information services, voice mail and security control for secret information areas. Speaker recognition technology is the most potential technology to create new services that will make our everyday lives more secured. Another important application of speaker recognition technology is for forensic purposes. Speaker recognition has been seen as an appealing research field for the last decades which still yields a number of unsolved problems.

In these existing speaker recognition systems, it is supposed that the input speech belongs to one of the known speakers. However, in many applications, such as in a real-time conversation or news broadcasting, the speech stream is continuous and there is no information about the beginning and end of the speech segment of a speaker. Therefore, if we need to index speech streams based on speaker or to perform video content analysis based on audio track, it is necessary to find speaker change points first in such applications before the speaker can be identified. This procedure is called speaker segmentation, or speaker change detection. Furthermore, speaker indexing (tracking), which clusters a speech stream by speaker identities, can be performed based on the results of speaker segmentation. Speaker tracking is also essential in many applications, such as conference and meeting indexing [2], audio/video retrieval or browsing [3, 4], speaker adaptation [5] for speech recognition, and video content analysis.

Traditionally, the speaker recognition task supposes that training and testing are composed of mono-speaker records. Then, to handle this kind of multi-speaker recordings, some extensions of the speaker recognition task are needed, such as

- The N-speaker detection which is similar to speaker verification. It consists in determining whether a set of target speakers are speaking in a conversation.
- Speaker tracking that consists in determining if and when a target speaker speaks in a multi-speaker record.
- Speaker segmentation that is close to speaker tracking but there is no information about the identity and number of speakers present. The aim of this task is to determine the number of speakers and also when they speak. This problem corresponds to a blind classification of the data, and the result is a partition in which every class is composed of segments of one speaker.

In this paper, we focus on the problem of speaker segmentation, detection and tracking in multi-speaker audio recordings using speaker biometrics. With the work presented here, our aim is to explore a new set of robust prosodic features for speaker segmentation and speaker diarization system.

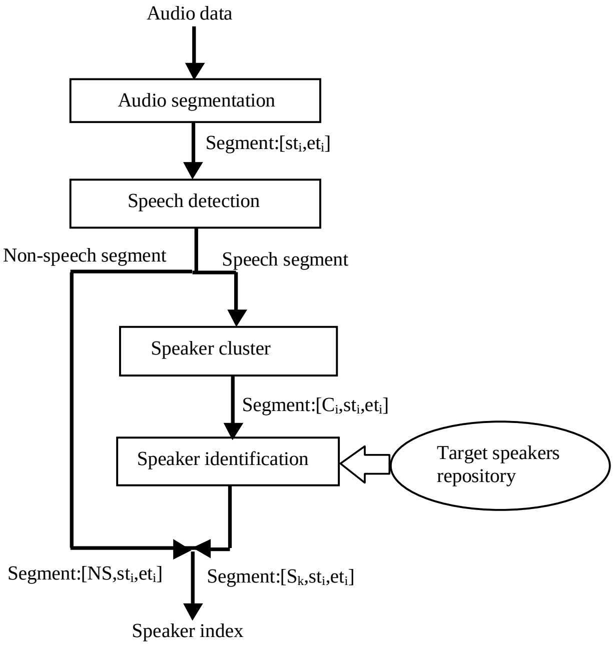

Fig. 1 Block Diagram of a typical speaker-indexing system.

A baseline speaker-indexing system architecture, which was followed in this work, is shown in Fig. 1. First, the audio signal is processed in an audio segmentation module, where time-stamps are produced at the locations of detected acoustic changes. Audio data are thus partitioned into small homogeneous segments labeled by starting and ending time of each segment (segments: [sti, eti] in Fig. 1). It is expected that each such segment should contain data from just one acoustic prosodic, i.e. speech from one speaker or non-speech data corresponding to music, silence or other non-speech prosodic. Therefore, the obtained segments should be additionally divided to those, which contain speech or non-speech data. This is done in a speech detection module. Non-speech segments are marked as [NS, sti, eti] in Fig. 1, and are discarded from further processing. Only speech segments are then passed through a speaker clustering module. The aim of a speaker clustering is to merge speech segments from each speaker together, a major issue being that the information of speakers and the actual number of speakers are unknown a priori and need to be automatically determined. At this stage, just relative speaker labels are produced and segments are marked with automatically derived cluster names (segments [Ci, sti, eti] in Fig. 1). The true identities of the speakers are obtained in a speaker identification module in the next stage. Here, a multiple speaker verification of each cluster is performed. Speaker identification module is capable to recognize just those speakers, who are presented in the repository of the target speakers and are previously enrolled into the system. Speech data from clusters, which do not correspond to any of the speakers from target group, should be marked as unknown speaker data and are discarded from further processing.

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. Speaker Recognition is basically divided into speaker identification and speaker verification [1]. A speaker identification system gets a test utterance as input. The task of the system is to find out which of the training speakers made the test utterance. So, the output of the system is the name of the training speaker, or possibly a rejection if the utterance has been made by an unknown person.

2. Prosodic Features

Prosody refers to non-segmental aspects of speech, including for instance syllable stress, intonation patterns, speaking rate and rhythm. One important aspect of prosody is that, unlike the traditional short-term spectral features, it spans over long segments like syllables, words, and utterances and reflects differences in speaking style, language background, sentence type, and emotions to mention a few. A challenge in text-independent speaker recognition is modeling the different levels of prosodic information (instantaneous, long-term) to capture speaker differences; at the same time, the features should be free of effects that the speaker can voluntarily control.

The most important prosodic parameter is the fundamental frequency (or F0). Combining F0-related features with spectral features has been shown to be effective, especially in noisy conditions. Other prosodic features for speaker recognition have included duration (e.g. pause statistics, phone duration), speaking rate, and energy distribution/modulations among others [27, 29, 44, 46]. Interested reader may refer to [46] for further details. In that study, it was found out, among a number of other observations, that F0-related features yielded the best accuracy, followed by energy and duration features in this order. Since F0 is the predominant prosodic feature, we will now discuss it in more detail.

Reliable F0 determination itself is a challenging task. For instance, in telephone quality speech, F0 is often outside of the narrowband telephone network passband (0.3.3.4 kHz) and the algorithms can only rely on the information in the upper harmonics for F0 detection. For a detailed discussion of classical F0 estimation approaches, refer to [36]. More recent comparison of F0 trackers can be found in [33]. For practical use, we recommend the YIN method [34] and the autocorrelation method as implemented in Praat software [30].

For speaker recognition, F0 conveys both physiological and learned characteristics. For instance, the mean value of F0 can be considered as an acoustic correlate of the larynx size [45], whereas the temporal variations of pitch are related to the manner of speaking. In text-dependent recognition, temporal alignment of pitch contours have been used [28]. In text-independent studies, long-term F0 statistics - especially the mean value have been extensively studied [31, 37, 40, 43, 47, 48]. The mean value combined with other statistics such as variance and kurtosis can be used as speaker model [29, 31, 37], even though histograms [37], latent semantic analysis [32] and support vector machines [46] perform better. It has also been found through a number of experiments that log(F0) is a better feature than F0 itself [37, 48].

F0 is a one-dimensional feature, therefore mathematically, not expected to be very discriminative. Multidimensional pitch- and voicing-related features can be extracted from the auto-correlation function without actual F0 extraction as done in [38, 39, 49] for example. Another way to improve accuracy is modeling both the local and long-term temporal variations of F0.

Capturing local F0 dynamics can be achieved by appending the delta features with the instantaneous F0 value. For longer-term modeling, F0 contour can be segmented and presented by a set of parameters associated with each segment [26, 27, 41, 46, 47]. The segments may be syllables obtained using automatic speech recognition (ASR) system [46]. An alternative, ASR-free approach, is to divide the utterance into syllable-like units using, for instance, vowel onsets [42] or F0/energy inflection points [26, 35] as the segment boundaries.

For parameterization of the segments, prosodic feature statistics and their local temporal slopes (tilt) within each segment are often used. In [27, 47], each voiced segment was parameterized by a piece-wise linear model whose parameters formed the features. In [46], the authors used N-gram counts of discretized feature values as features to an SVM classifier with promising results. In [35], prosodic features were extracted using polynomial basis functions.

2.1 Prosodic Features in the LP Residual

Speech signals, as any other real world signals, are produced by exciting a system with source. A simple block diagram representation of the speech production mechanism is shown in the Fig. 2. Vibrations of the vocal folds, powered by air coming from the lungs during exhalation, are the sound prosodic for speech. Hence, as can be from Fig. 2, the glottal excitation forms the prosodic, and the vocal tract forms the system. One of the most powerful speech analysis techniques is the method of linear predictive analysis. The philosophy of linear prediction is intimately related to the basic speech production model. The Linear Predictive Coding (LPC) analysis approach performs spectral analysis on short segments of speech with an all-pole modeling constraint [18]. Since speech can be modeled as the output of linear, time-varying system excited by a prosodic, LPC analysis captures the vocal tract system information in terms of coefficients of the filter representing the vocal tract mechanism. Hence, analysis of speech signal by LP results in two components, namely the synthesis filter on one hand and the residual on the other hand. In brief, the LP residual signal is generated as a by product of the LPC analysis, and the computation of the residual signal is given below.

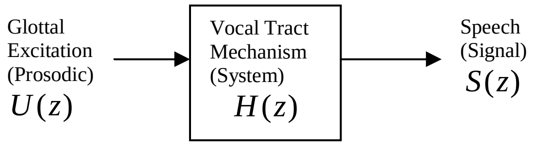

Fig. 2 Prosodic and System representation of speech production mechanism

If the input signal is represented by u_n and the output signal by s_n, then the transfer of the system can be expressed as,

$$H(z) = \frac{S(z)}{U(z)} \qquad (1)$$

Where $s(z)$ and $u(z)$ are z-transforms of s_n and u_n respectively.

Consider the case where we have output signal and the system and have to compute the input signal. The above equation can be expressed as

$$S(z) = H(z)U(z)$$

$$U(z) = \frac{S(z)}{H(z)} \qquad (2)$$

$$U(z) = \frac{1}{H(z)} S(z) \qquad (3)$$

$$U(z) = A(z)S(z) \qquad (4)$$

Where $A(z) = 1/H(z)$ is the inverse filter representation of the vocal tract system?

Linear prediction models the output s_n as the linear function of past outputs and present and past inputs. Since prediction is done by a linear function, the name linear prediction. Assuming an all-pole for the vocal tract, the signal s_n can be expressed as linear combination of past values and some input u_n as shown below.

$$Sn = -\sum_{k=1}^{p} a_k S_{n-k} + GU_n \qquad (5)$$

Where G is a gain factor.

Now assuming that the input u_n is unknown, the signal s_n can be predicted only approximately from a linear weighted sum of past samples. Let this approximation of s_n be, where

$$\tilde{S}_n = -\sum_{k=1}^{p} a_k s_{n-k} \qquad (6)$$

Then the error between the actual value Sn and predicted value is given by $e_n = S_n - \tilde{S}_n$ [20]. This error is nothing but LP residual of signal is shown in Fig 3.

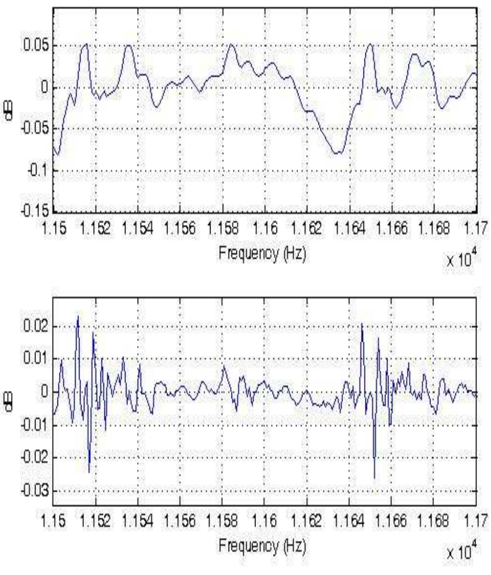

Fig. 3 Actual signal and its LP residual

The significance of the prosodic feature is illustrated in the Fig. 4. The speech utterances sampled at 8 kHz were collected from five male speakers over a microphone. All the speakers uttered the sound unit /aa/. The significant instants of the glottal excitation are computed for the five speakers. The instants corresponding to the steady section of the utterances were displayed in the Fig. 4. It is clearly seen from the Fig. 4 that the periodicity of the instants of glottal excitation for each of the five speakers is different from that of the other's. As shown in the Fig.4, it is clearly evident that the prosodic features for different speaker are different.

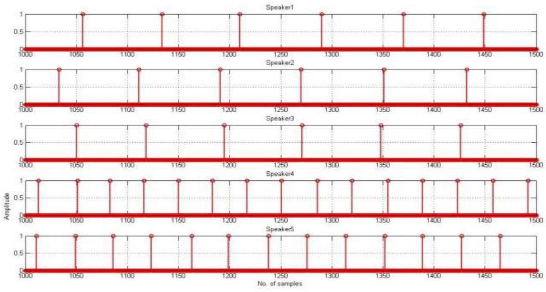

Fig. 4 Instants of significant excitations for five male speakers for the sound unit

3. Feature Extraction of LP Residual Signal

MFCC is the best known and most popular, and this feature has been used for gender identification. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Rather than the speech waveforms themselves, MFFCs are less susceptible to the said variations [12].

3.1 Motivation to use Melfrequency Cepstral coefficients (MFCC)

Since our interest is in capturing global features which correspond to source excitation, the low frequency or pitch components are to be emphasized. To fulfill this requirement it is felt that MFCC are most suitable as they emphasize low frequency and de-emphasize high frequencies.

3.2 MFCC

In this phase the digital speech signal is partitioning into segments (frames) with fixed length 10-30 ms from which the features are extracted due to their spectral qualities. Spectrum is achieved with fast Fourier transformation [20]. Then an arrangement of frequency range to mel scale follows according to relation

$$f_{mel} = 2595 \log\left(1 + \frac{f_{Hz}}{700}\right) \qquad (7)$$

By logarithm of amplitude of mel spectrum and applying reverse Fourier transformation we achieve frame cepstrum:

$$mel - cepstrum(frame) = FFT^{-1}\left[mel(\log|FFT(frame)|)\right] \qquad (8)$$

The FFT-base cepstral coefficients are computed by taking IFFT of the log magnitude spectrum of the Speech signal. The mel-warped cepstrum is obtained by inserting a intermediate step of transforming the frequency scale to place less emphasis on higher frequencies before taking the IFFT [13][21][22].

3.3 Support Vector Machine

Support vector machine (SVM) is a powerful discriminative classifier that has been recently adopted in speaker recognition. It has been applied both with spectral [50, 51], prosodic [46, 53], and high-level features [51]. Currently SVM is one of the most robust classifiers in speaker verification, and it has also been successfully combined with GMM to increase accuracy [50, 52]. One reason for the popularity of SVM is its good generalization performance to classify unseen data.

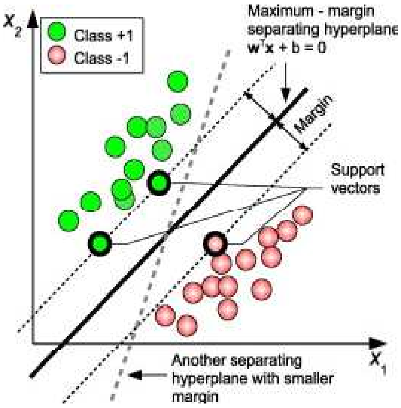
x_2
Class +1
Class -1
Maximum - margin separating hyperplane
$w^Tx + b = 0$
Margin
Support vectors
Another separating hyperplane with smaller margin
x_1

Fig. 5: Principle of support vector machine (SVM). A maximum-margin hyperplane that separates the positive (+1) and negative (-1) training examples is found by an optimization process.

The SVM, as illustrated in Fig. 5, is a binary classifier which models the decision boundary between two classes as a separating hyperplane. In speaker verification, one class consists of the target speaker training vectors (labeled as +1), and the other class consists of the training vectors from an “impostor” (background) population (labeled as -1). Using the labeled training vectors, SVM optimizer finds a separating hyperplane that maximizes the margin of separation between these two classes.

Formally, the discriminant function of SVM is given by [50],

$$f\left(x\right)=\sum_{i=1}^{N}\alpha_i t_i K\left(x,x_i\right)+d \qquad (9)$$

Here $t_i \varepsilon$ {+1,-1} are the ideal output values, $\sum_{i=1}^{N}\alpha_i t_i = 0$ and $\alpha_i > 0$. The support vectors xi, their corresponding weights αi and the bias term d, are determined from a training set using an optimization process. The kernel function K (. , .) is designed so that it can be expressed as $K\left(x,y\right)=\phi\left(x\right)^T\phi\left(y\right)$, where $\phi(x)$ is a mapping from the input space to kernel feature space of high dimensionality. The kernel function allows computing inner products of two vectors in the kernel feature space. In a high-dimensional space, the two classes are easier to separate with a hyperplane. Intuitively, linear hyperplane in the high-dimensional kernel feature space corresponds to a nonlinear decision boundary in the original input space (e.g. the MFCC space). For more information about SVM and kernels, refer to [54, 55].

4. Conclusions

The objective in this paper was mainly to demonstrate the significance of the speaker-specific prosodic information (source) present in the linear prediction residual for speaker segmentation. We propose speaker specific prosodic features for speaker indexing system. This LP residual (source), which is generated by LP analysis, is usually ignored in all the major applications of speech analysis like speaker recognition. Only LPC coefficients are used to compute the feature vectors. But the residual signal is rich with prosodic characteristics, which are also speaker-specific. Hence, the information present in the residual signal can be used for speaker indexing task and it can effectively modeled by Support Vector Machines (SVMs).

References

[1] Campbell JP Jr., “Speaker Recognition: a tutorial”. Proc. IEEE 85(9):1437–1462, 1997.

[2] Bonastre JF, Delacourt P, Fredouille C, Merlin T, and Wellekens C, “A speaker tracking system based on speaker turn detection for NIST evaluation”, in Proc. IEEE international conference on acoustics, speech and signal processing, 2000 , pp. 1177–1180.

[3] Roy D, and Malamud C, “Speaker identification based text to audio alignment for an audio retrieval system”, in Proc. IEEE international conference on acoustics, speech and signal processing, 1997, pp. 1099–1102.

[4] Kimber DG, Wilcox LD, Chen FR, and Moran TP, “Speaker segmentation for browsing recorded audio”, in ACM CHI’95 Mosaic of Creativity, 1995, pp. 212–213.

[5] Mori K, and Nakagawa S, “Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition”, in Proc. IEEE international conference on acoustics, speech and signal processing, 2002.

[6] Alex Acero and Xuedong Huang, “Speaker and gender normalization for continuous-density hidden markov models”, in Proc. of the Int. Conf. on Acoustics, Speech and Signal , IEEE, May 1996.

[7] C. Neti and Salim Roukos, “Phone-specific gender-dependent models for continuous speech recognition”, Automatic Speech Recognition and Understanding Workshop (ASRU97), Santa Barbara, CA, 1997.

[8] R. Vergin, A. Farhat and D. O’Shaughnessy, “Robust gender-dependent acoustic-phonetic modeling in continuous speech recognition based on a new automatic male/female classification”, in Proc. of IEEE Int. Conf. on Spoken Language (ICSLP), Oct. 1996, pp. 1081,.

[9] Gish H., and M. Schmidt (1994), “Text-independent speaker identification”, IEEE Signal Process Mag. 11(4):18–32.

[10] D.O'Shaughnessy, "Speech communication: Human and Machine", Addison-Wesley, New York, 1987.

[11] Rabiner L.R., Juang B.H., "Fundamentals of speech recognition", Prentice-Hall, Englewood Cliffs, NJ, 1993.

[12] Makhoul J., "Linear prediction: a tutorial review", Proc. IEEE 63, 1975, pp. 561–580.

[13] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J. Acoust. Soc. Ameri., vol. 55, pp. 1304-1312, Jun. 1974. K. Elissa, "Title of paper if known," unpublished.

[14] A.E. Rosenberg and M. Sambur, "New techniques for automatic speaker verification", 1975, vol. 23, no. 2, pp. 169-175.

[15] M. R. Sambur, "Speaker recognition using orthogonal linear prediction", IEEE Trans. Acoust. Speech and Signal Processing, Aug. 1976, vol. 24, pp. 283-289.

[16] J. Naik and G. R. Doddington, "High performance speaker verification using principal spectral components", in Proc. IEEE Int. Conf. Acoust. Speech, Singal Processing, 1986, pp. 881-884.

[17] Furui S., "Recent advances in speaker recognition", Pattern Recognition Lett. 18, 1997, pp. 859–872.

[18] S.R.Mahadeva Prassana, Cheedella S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech", Speech Communications, 2006, Vol. 48, pp. 1243-1261.

[19] Dempster A., Laird N., and Rubin D., "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, 1977, vol. 39, pp. 1-38.

[20] Molau S., Pitz M., Schluter R., and Ney H., "Computing mel-frequency cepstral coefficients on the power spectrum", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2001, vol. 1, pp. 73-76.

[21] Picone J. W., "Signal modeling techniques in speech recognition", Proceedings of IEEE, Sep. 1993, vol. 81, no. 9, pp. 1215-1247.

[22] Gish H., Krasner M., Russell W., and Wolf, J., "Methods and experiments for text-independent speaker recognition over telephone channels", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 1986, vol. 11, pp. 865-868.

[23] Reynolds D. A., and Rose R. C., " Robust text-independent speaker identification using gaussian mixture models", IEEE-Transactions on Speech and Audio Processing, 1995, vol. 3, no. 1, pp. 72-83.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. Royal Statist. Soc. Ser. B. (methodological), 1977, vol. 39, pp. 1-38.

[25] K.N. Stevens, Acoustic Phonetics. Cambridge, England: The MIT Press, 1999.

[26] Adami A, "Modeling prosodic differences for speaker recognition", Speech Communication 49, 4 (April 2007), pp. 277-291.

[27] Adami A., Mihaescu R., Reynolds D., and Godfrey J., "Modeling prosodic dynamics for speaker recognition," in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003) (Hong Kong, China, April 2003), pp. 788-791.

[28] Atal B., "Automatic speaker recognition based on pitch contours", Journal of the Acoustic Society of America 52, 6 (1972), pp. 1687-1697.

[29] Bartkova K., D.L.Gac, Charlet, D., and Jouvet D., "Prosodic parameter for speaker identification", in Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002) (Denver, Colorado, USA, September 2002), pp. 1197-1200.

[30] Boersma P., and Weenink D., Praat: doing phonetics by computer [computer program]. Available: http: //www.praat.org/.

[31] Carey M., Parris E., Lloyd-Thomas H., and Bennett S., "Robust prosodic features for speaker identification", in Proc. Int. Conf. on Spoken Language Processing (ICSLP 1996) (Philadelphia, Pennsylvania, USA, 1996), pp. 1800-1803.

[32] Chen Z.H., Liao Y.F., and Juang Y.T, "Eigen-prosody analysis for robust speaker recognition under mismatch handset environment", in Proc. Int. Conf. on Spoken

Language Processing (ICSLP 2004) (Jeju, South Korea, October 2004), pp. 1421-1424.

[33] Cheveigne A., and Kawahara H., "Comparative evaluation of f0 estimation algorithms", in Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) (Aalborg, Denmark, September 2001), pp. 2451-2454.

[34] DeCheveigne, A., and Kawahara, H. YIN, "A fundamental frequency estimator for speech and music", The Journal of Acoustical Society of America 111, 4 (April 2002), 1917-1930.

[35] Dehak N., Kenny P., and Dumouchel P., "Modeling prosodic features with joint factor analysis for speaker verification", IEEE Trans. Audio, Speech and Language Processing 15, 7 Sept. 2007, pp. 2095-2103.

[36] Hess W., "Pitch determination of speech signals: algorithms and devices," Springer Verlag, Berlin, 1983.

[37] Kinnunen T., and Gonzalez-Hautamaki R., "Long-term f0 modeling for text-independent speaker recognition", in Proc. 10th International Conference Speech and Computer (SPECOM'2005) (Patras, Greece, October 2005), pp. 567-570.

[38] Laskowski K., and Jin Q., "Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum", in Proc. Int. conference on acoustics, speech and signal processing (ICASSP 2009) (Taipei, Taiwan, April 2009), pp. 4541-4544.

[39] Ma B., Zhu D., and Tong R., "Chinese dialect identification using tone features based on pitch flux", in Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006) (Toulouse, France, May 2006), vol. 1, pp. 1029-1032.

[40] Markel J., Oshika B., and A.H. Gray, "Long-term feature averaging for speaker recognition", IEEE Trans. Acoustics, Speech, and Signal Processing 25, 4 August 1977, pp. 330-337.

[41] Mary L., and Yegnanarayana B., "Prosodic features for speaker verification", in Proc. Interspeech 2006 (ICSLP), Pittsburgh, Pennsylvania, USA, September 2006, pp. 917-920.

[42] Mary L., and Yegnanarayana B., "Extraction and representation of prosodic features for language and speaker recognition", Speech Communication 50, 10 (2008), pp. 782-796.

[43] Nolan F., "The Phonetic Bases of Speaker Recognition," Cambridge University Press, Cambridge, 1983.

[44] Reynolds D., Andrews W., Campbell J., Navratil J., Peskin B., Adami A., Jin Q., Klusacek D., Abramson J., Mihaescu R., Godfrey J., Jones D., and Xiang B., "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition", in Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong, China, April 2003, pp. 784-787.

[45] Rose P., "Forensic Speaker Identification", Taylor and Francis, London, 2002.

[46] Shriberg E., Ferrer L., Kajarekar S., Venkataraman A., and Stolcke A., "Modeling prosodic feature sequences for speaker recognition", Speech Communication 46, July 2005, pp. 455-472.

[47] Sonmez K., Shriberg E., Heck L., and Weintraub M., "Modeling dynamic prosodic variation for speaker verification", in Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998), Sydney, Australia, November 1998, pp. 3189-3192.

[48] Sonmez M., Heck L., Weintraub M., and Shriberg E., "A lognormal tied mixture model of pitch for prosody-based speaker recognition", in Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997), (Rhodos, Greece, September 1997), pp. 1391-1394.

[49] Wildermoth B., and Paliwal K., "Use of voicing and pitch information for speaker recognition", in Proc. 8th Australian Intern. Conf. Speech Science and Technology (Canberra, December 2000), pp. 324-328.

[50] Campbell W., Campbell J., Reynolds D., Singer E., and Torres-Carrasquillo P., "Support vector machines for speaker and language recognition", Computer Speech and Language 20, 2-3 (April 2006), pp. 210-229.

[51] Campbell W., Sturim D., and Reynolds D., "SVM based speaker veri_cation using a GMMsupervector kernel and NAP variability compensation", in Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005) (Philadelphia, USA, March 2005), pp. 637-640.

[52] Campbell W., Sturim D., and Reynolds D., "Support vector machines using GMMsupervectors for speaker verification".

[53] Ferrer L., Shriberg E., Kajarekar S., and Sonmez K., "Parameterization of prosodic feature distributions for SVM modeling in speaker recognition", in Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007) (Honolulu, Hawaii, USA, April 2007), vol. 4, pp. 233-236.

[54] Bishop C., "Pattern Recognition and Machine Learning", Springer Science, Business Media, LLC, New York, 2006.

[55] Muller K.-R., Mika S., Ratsch G., Tsuda K., and Scholkopf B., "An introduction to kernel-based learning algorithms", IEEE Trans. on Neural Networks 12, 2 May 2001, pp. 181-201.

Authors

Dr. R. Rajeswara Rao was born in India in 1976. He received his B.Tech. degree in Computer Science and engineering from Siddhartha Engineering College, Vijayawada, India, in 1999 and the M.Tech. degree in Computer Science and Engineering from College of Engineering, JNT University, Hyderabad, India, in 2003. He has completed his Ph.D degree in computer science and engineering from JNT University, Hyderabad, India, in 2010. He is currently Associate Professor in the Department of Computer Science and engineering, JNT Uinversity, Kakinada. His research interests are speech processing and pattern recognition.

Mr. V. Subba Ramaiah received his B.Tech. degree in Computer Science and engineering from SITAMS, JNT University, Chittoor, India, in 2002 and the M.Tech. degree in Computer Science from SIT, JNT University, Hyderabad, India, in 2007. He has been working as Senior Assistant Professor in the department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad. His research interests are speech and pattern recognition.