Role of neural network models for developing speech systems

K SREENIVASA RAO

School of Information Technology, Indian Institute of Technology Kharagpur,
Kharagpur 721302, West Bengal, India
e-mail: ksrao@iitkgp.ac.in

Abstract. This paper discusses the application of neural networks for developing different speech systems. Prosodic parameters of speech at syllable level depend on positional, contextual and phonological features of the syllables. In this paper, neural networks are explored to model the prosodic parameters of the syllables from their positional, contextual and phonological features. The prosodic parameters considered in this work are duration and sequence of pitch (F_0) values of the syllables. These prosody models are further examined for applications such as text to speech synthesis, speech recognition, speaker recognition and language identification. Neural network models in voice conversion system are explored for capturing the mapping functions between source and target speakers at source, system and prosodic levels. We have also used neural network models for characterizing the emotions present in speech. For identification of dialects in Hindi, neural network models are used to capture the dialect specific information from spectral and prosodic features of speech.

Keywords. Feedforward neural network (FFNN); autoassociative neural network (AANN); prosody models; speaker recognition; language identification; voice conversion; Emotion recognition; Dialect identification.

1. Introduction

Neural networks are well known for capturing the complex non-linear relations present in data. Neural networks can be broadly classified into three categories, namely, feedforward neural networks (FFNN), feedback neural networks (FBNN) and the combination of both feedforward and feedback neural networks. The combined feedforward and feedback neural network is also known as competitive learning neural network (CLNN) (Yegnanarayana 1999).

Feedforward neural networks can be used for classification and function approximation tasks. Feedback neural networks can be used for pattern storage and pattern association tasks. Competitive learning neural networks are used for pattern storage, pattern clustering and feature mapping (Yegnanarayana 1999). In this study, we focus mainly on the application of FFNNs for different speech tasks. A simple multilayer feedforward neural network (MLFFNN) can perform the classification task by yielding the feature vectors as input, and their respective class labels as

output. The performance of the model depends on number of hidden layers, number of nodes in hidden layers, nature of processing units and number of examples used for training the network. MLFFNN with symmetric structure will perform the pattern association task. Here, the input patterns are given as input to the network, and the corresponding output patterns are given as output. The network is expected to capture the hidden associations between the input and output patterns. If the input and output patterns are the same, then the network will perform the autoassociation, otherwise it performs heteroassociation. An autoassociative neural network (AANN) captures the distribution of the data points depending on the constraints imposed by the structure of the network (Yegnanarayana & Kishore 2002). The probability distribution of the feature vectors can be viewed as an error surface generated by the AANN. A heteroassociative neural network (HANN) can be used for capturing the mapping function between the set of input and output patterns. In general for FFNN, the weights are adjusted to realize the global minimum of the total error for the training data in the weight space. Irrespective of the topology of the FFNN, minimization of the training error leads to optimization of the performance of their respective tasks. The designated tasks may be either classification or association or function approximation. In the following paragraphs, various FFNN topologies are briefly described with respect to different speech tasks.

We discuss different topologies of MLFFNNs used for (i) predicting the prosodic parameters such as durations and average F_0 values of the syllables, (ii) speech, speaker and language recognition using prosody models, (iii) developing the text-to-speech (TTS) synthesis system using the prosody models, (iv) developing the mapping functions between source and target speakers using source, system and prosodic features, (v) capturing the emotion specific information from source, system and prosodic features and (vi) capturing the dialect specific information using spectral and prosodic features. MLFFNNs are explored to capture the duration and intonation patterns of the sequence of syllables from their linguistic and production constraints. For this task, a four-layer FFNN is trained using positional, contextual and phonological features of the syllables as input and the duration or F_0 values of the syllables as output of the network. The details of this task are given in section 3. Neural network models also captures the prosodic knowledge specific to speaker, language and sound units. This is demonstrated by developing the prosody models for different speakers, languages and categories of sound units. The prosody models developed using the FFNNs can be used for building the TTS synthesis system. For synthesizing the speech from the given text, the prosodic models are used for generating the desired prosodic parameters associated to the sequence of sound units. Based on these parameters the appropriate units will be selected in the case of unit selection-based TTS synthesis, and the units will be further modified according to the required prosody.

The basic goal of the voice conversion system is to transform the speaker characteristics, while maintaining the message intact. FFNN models are explored to capture the non-linear mapping of speaker specific characteristics between source and target speakers. A five-layer symmetric structure FFNN is used to capture the association between the speaker specific features of source and target speakers. The proposed five-layer symmetric FFNN used for capturing the non-linear mapping function between the features of source and target speakers is also known as heteroassociative neural network. This network is trained with the time aligned features corresponding to source and target speakers. The details of the network models used for the voice conversion are given in section 5.

AANN models are explored for capturing the emotion and dialect specific information from the spectral and prosodic features of speech. Here, AANNs are used to capture the distribution of feature vectors specific to each emotion or dialect. Therefore, for developing the emotion or dialect recognition system, we need to develop AANN models for each of the emotions or

dialects. The details of the emotion and dialect models, and their recognition systems are given in sections 6 and 7, respectively.

This paper is organized as follows: The details of two basic topologies of FFNNs are discussed in section 2. Prediction of duration and F_0 values of the sequence of syllables using MLFFNN models is given in section 3. Applications of the prosody models are briefly demonstrated in section 4. In section 5, functionality of HANN for voice conversion application is discussed. The use of AANNs for developing the emotion and dialect recognition systems are given in sections 6 and 7, respectively. The summary and conclusions of the paper are given in the last section.

2. Feedforward neural network

A feedforward neural network model is expected to capture the functional relationship between the input and output feature vectors of the given training data. The general structure of the FFNN is shown in figure 1. The number of layers considered for this network is four. It is known that a neural network with two hidden layers can realize any continuous vector-valued function (Sontag 1992). The first layer is the input layer with linear units. The second and third layers are hidden layers. The second layer (first hidden layer) of the network has more units than the input layer, and it can be interpreted as capturing some local features in the input space. The third layer (second hidden layer) has fewer units than the first layer, and can be interpreted as capturing some global features (Haykin 1999; Yegnanarayana 1999). The fourth layer is the output layer, and the number of units represent the dimension of output feature vector. The activation function for the units at the input layer is linear, and for the units at the hidden layers, it is non-linear. Generalization by the network is influenced by three factors: the size of the training set, the architecture of the neural network, and the complexity of the problem. We have no control over the first and

Figure 1. A four-layer feedforward neural network.

last factors. Various network structures will be explored for optimizing the performance. The structure of the network is represented by $PL\ QN\ RN\ SN$, where L denotes a linear unit, and N denotes a non-linear unit. P , Q , R and S are the integer values indicate the number of units used in different layers. It should be noted that the structure in terms of number of units in the hidden layer is not critical. The number of units in the two hidden layers is guided by the heuristic arguments given above. The non-linear units use $\tanh(s)$ as the activation function, where s is the activation value of that unit. All the input and output features are normalized to the range $[-1, +1]$ before presenting to the neural network. The backpropagation learning algorithm is used for adjusting the weights of the network to minimize the mean squared error at the output (Yegnanarayana 1999). The above network can be used for classification, by providing the class labels of the input feature vectors as output of the network.

It is well-known that feedforward neural network with symmetric structure shown in figure 2 is used specifically for capturing the associations present in the input and output features (Haykin 1999). The network is supposed to provide the mapping function between source and target patterns. The number of input parameters (dimension of the input vector) and the number of output parameters (dimension of the output vector) fed to the network are the same. For capturing the complex non-linear relations between source and target patterns, more than one hidden layer is essential. In general, two hidden layers will be chosen, one for capturing global relations and the other for capturing local relations (Rao & Yegnanarayana 2007). To exploit the advantage of symmetric network structure, the minimum number of hidden layers needed is three. Hence, the overall network structure consists of five layers, extreme layers being input and output layers used to feed the feature vectors to the network, the middle layer will be the compression layer used to capture the global behaviour and the remaining two layers positioned between input and compression layers, and between compression and output layers are used to capture the local behaviour between input and output patterns. The structure of the symmetric network can be represented as $PL\ QN\ RN\ QN\ PL$; where P , Q and R indicate the number of units present in different layers, and the nature of functionality of the units is denoted by L and N which indicates linearity and non-linearity.

Feedforward neural network with the desired output being the same as the input vector can be explored for capturing the distribution of input feature vectors. These neural network models

Figure 2. A five-layer symmetric FFNN model.

try to map an input vector onto itself, and hence they are known as autoassociation or identity mapping networks (Yegnanarayana 1999; Haykin 1999). An autoassociation network consists of an input layer, an output layer and one or more hidden layers. The number of units in the input and output layers are equal to the size of the input vectors. The number of nodes in the middle hidden layer is less than the number of units in the input or output layers. The middle layer is also the dimension compression hidden layer. The activation function of the units in the input and output layers are linear, whereas the activation function of the units in the hidden layer can be either linear or non-linear. The performance of AANN models can be interpreted in different ways, depending on the problem and the input data. If the data are a set of feature vectors in the feature space, then the performance of AANN models can be interpreted either as linear and non-linear principal component analysis or distribution capturing of the input data (Diamantaras & Kung 1996; Ikbal *et al* 1999; Oppenheim *et al* 1999; Kishore & Yegnanarayana 2001). On the other hand, if the AANN is presented directly with signal samples, such as speech signal, the network captures the implicit linear and non-linear relations among the samples (Anjani 2000; Yegnanarayana *et al* 2001a; Gupta 2003). The use of the above mentioned topologies of FFNNs are described in the following sections.

3. Prosody models

In this section, we briefly describe the features (representing the linguistic context and production constraints) influencing the duration and intonation patterns of the sequence of syllables. Feedforward neural networks are used to capture the duration and intonation knowledge associated to the sequence of syllables. In this study, separate neural networks are used for capturing the duration and intonation patterns. The prediction performance of the neural network models is analysed using objective measures.

3.1 Features for modelling prosody

In this study, we use 25 features (which form a feature vector) for representing the linguistic context and production constraints of each syllable. These features represent positional, contextual and phonological information of each syllable. Features representing the positional information are further classified based on the position of a word in the phrase and the position of the syllable in a word and phrase.

Syllable position in the phrase: A phrase is delimited by the orthographic punctuation. The syllable position in a phrase is characterized by three features. The first one represents the distance of the syllable from the starting position of the phrase. It is measured in number of syllables, i.e., the number of syllables ahead of the present syllable in the phrase. The second feature indicates the distance of the syllable from the terminating position of the phrase. The third feature represents the total number of syllables in the phrase.

Syllable position in the word: In Indian languages, words are identified by the spacing between them. The syllable position in a word is characterized by three features similar to the phrase. The first two features are the positions of the syllable with respect to the word boundaries. The third feature is the number of syllables in a word.

Position of the word: The F_0 and duration of a syllable may depend on the position of the word in an utterance. Therefore, the word position is used for developing the intonation and duration

models. The word position in an utterance is represented by three features. They are the positions of the word with respect to the phrase boundaries, and the number of words in the phrase.

Syllable identity: A syllable is a combination of segments of consonants (C) and vowels (V). In this study, syllables with more than four segments (Cs or Vs) are ignored since the number of such syllables present in the database is less than 1%. Each segment of a syllable is encoded separately, so that each syllable identity is represented by a four-dimensional feature vector. Each of the C and V segments are uniquely coded based on their identity.

Context of the syllable: The F_0 and duration of a syllable may be influenced by its adjacent syllables. Hence, for modelling the F_0 and duration of a syllable, the contextual information is represented by the previous and following syllables. Each of these syllables is represented by a four-dimensional feature vector, representing the identity of the syllable.

Syllable nucleus: Another important feature is the vowel position in a syllable, and the number of segments before and after the vowel in a syllable. This feature is represented with a three-dimensional feature vector specifying the consonant-vowel structure present in the syllable.

Pitch: The F_0 value of a syllable may be influenced by the pitch value of the preceding syllable. Therefore, this information is used in the feature vector of the syllable for predicting the F_0 value of the present syllable.

Gender identity: The database contains speech from both male and female speakers. The gender information is represented by a single feature. This feature is used for modelling the duration of a syllable.

The list of features and the number of input (feature) nodes for a neural network are given in table 1. These features are coded and normalized before presenting to the neural network model.

Table 1. List of the factors affecting the duration/ F_0 of a syllable, features representing the factors and the number of nodes needed for neural network to represent the features.

Factors	Features	No. of nodes
Syllable position in the phrase	Position of syllable from beginning of the phrase	3
	Position of syllable from end of the phrase	
	Number of syllables in the phrase	
Syllable position in the word	Position of syllable from beginning of the word	3
	Position of syllable from end of the word	
	Number of syllables in the word	
Word position in the phrase	Position of word from beginning of the phrase	3
	Position of word from end of the phrase	
	Number of words in the phrase	
Syllable identity	Segments of the syllable (consonants and vowels)	4
Context of the syllable	Previous syllable	4
	Following syllable	4
Syllable nucleus	Position of the nucleus	3
	Number of segments before the nucleus	
	Number of segments after the nucleus	
Pitch	F_0 of the previous syllable	1
Gender	Gender of the speaker	1

3.2 Speech database

The database for this study consists of 19 Hindi, 20 Telugu, 33 Tamil and 20 Kannada broadcast news bulletins (Khan *et al* 2003). In each language, these news bulletins were read by male and female speakers. Total durations of speech in Hindi, Telugu, Tamil and Kannada are 3.5 h, 4.5 h, 5 h and 4.5 h, respectively. The speech signal was sampled at 16 kHz and represented as 16 bit numbers. The speech utterances are manually transcribed into text using common transliteration code (ITRANS) for Indian languages (Chopde 2001). The speech utterances are segmented and labelled manually into syllable-like units. Each bulletin is organized in the form of syllables, words and orthographic text representations of the utterances. Each syllable and word file contains the text transcriptions and timing information in number of samples. The fundamental frequencies of the syllables are computed using the autocorrelation of the Hilbert envelope of the linear prediction residual (Prasanna & Yegnanarayana 2004a). The average pitch (F_0) for male speakers and female speakers in the database was found to be 129 and 231 Hz, respectively.

3.3 Prosody modelling with feedforward neural networks

A four-layer feedforward neural network is used for modelling the duration/ F_0 patterns of syllables. The structure of the FFNN used in this study is shown in figure 3. Here, the FFNN model is expected to capture the functional relationship between the input and output feature vectors of the given training data. The mapping function is between the 25-dimensional input vector and the one-dimensional output. Several network structures are explored in this study. The (empirically optimized) final structure of the network is 25L 50N 12N 1N, where L denotes a linear unit, and N denotes a non-linear unit. The integer value indicates the number of units used in that layer.

Figure 3. A four-layer feedforward neural network model for capturing the duration/ F_0 patterns of syllables.

For modelling the durations of syllables, neural network models are developed with respect to language. In the case of intonation modelling, a separate model is developed for each of the speakers in four languages. For each syllable, a 25-dimension input vector is formed, representing the positional, contextual and phonological features. The fundamental frequency or duration of each syllable is obtained from the database.

The prosody models are evaluated using syllables in the test set. For each syllable in the test set, the F_0 /duration value is predicted using the FFNN by presenting the feature vector of each syllable as input to the neural network. The deviation of the predicted F_0 /duration from the actual F_0 /duration is obtained. The prediction performance of the intonation models is illustrated in table 2 for one male (M) and one female (F) speaker for each language. The percentages of syllables predicted for different deviations from their actual F_0 values are given in table 2. For each syllable, the deviation (D_i) is computed as follows:

$$D_i = \frac{|x_i - y_i|}{x_i} \times 100,$$

where x_i and y_i are the actual and predicted F_0 /duration values, respectively. To evaluate the prediction accuracy, the average prediction error (μ), the standard deviation (σ), and the correlation coefficient ($\gamma_{X,Y}$) are computed using actual and predicted F_0 /duration values. These results are given in table 2.

The first column indicates different languages used in the analysis. The second column indicates the number of syllables specific to a speaker used in testing. Columns 3–7 indicate the percentage of syllables predicted within different deviations from their actual F_0 values. Columns 8–10 indicate the objective measures (μ , σ and $\gamma_{X,Y}$). The definitions of average prediction error (μ), standard deviation (σ) and linear correlation coefficient ($\gamma_{X,Y}$) are given below.

$$\mu = \frac{\sum_i |x_i - y_i|}{N},$$

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, \quad d_i = e_i - \mu, \quad e_i = x_i - y_i,$$

Table 2. Performance of the FFNN models for predicting the F_0 values of syllables for different languages. The results for female (F) and male (M) speakers are given separately.

Language	Gender No. of syllables	% Predicted syllables within deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
Hindi	F(660)	19	42	70	84	97	19.27	17.82	0.79
	M(1143)	21	45	76	93	98	11.72	9.42	0.81
Telugu	F(1276)	20	41	74	92	99	16.02	13.04	0.80
	M(984)	18	39	68	85	97	9.93	9.59	0.82
Tamil	F(741)	23	46	78	92	99	17.26	13.73	0.87
	M(1267)	24	43	79	91	97	12.41	11.54	0.81
Kannada	F(741)	23	46	78	92	99	17.26	13.73	0.87
	M(1267)	24	43	79	91	97	12.41	11.54	0.81

Table 3. Performance of the FFNN models for predicting the durations of syllables for different languages.

Language	% Predicted syllables within deviation			Objective measures		
	10%	25%	50%	μ (ms)	σ (ms)	γ
Hindi	29	68	84	32	26	.75
Telugu	29	66	86	29	23	.78
Tamil	34	75	96	26	22	.82
Kannada	34	75	96	26	22	.82

where x_i , y_i are the actual and predicted F_0 values, respectively, and e_i is the error between the actual and predicted F_0 values. The deviation in error is d_i , and N is the number of observed F_0 values of the syllables. The correlation coefficient is given by

$$\gamma_{X,Y} = \frac{V_{X,Y}}{\sigma_X \cdot \sigma_Y}, \text{ where } V_{X,Y} = \frac{\sum_i |(x_i - \bar{x})| \cdot |(y_i - \bar{y})|}{N}.$$

The quantities σ_X , σ_Y are the standard deviations for the actual and predicted F_0 values, respectively, and $V_{X,Y}$ is the correlation between the actual and predicted F_0 values.

The prediction performance of the duration models is given in table 3. The number of epochs needed for training depends on the behaviour of the training error. It was found that about 500 epochs are adequate for this study. The learning ability of the network from training data can be observed from the trend of the training error with the number of epochs. The training errors for neural network models for one speaker in each of the three languages are shown in figure 4. The decreasing trend in the training error indicate that the network is capturing the implicit relation between the input and output.

Figure 4. Training errors for the neural network models for predicting the intonation patterns in three languages Hindi, Telugu and Tamil.

The prediction performance of the intonation model for a male speaker in Hindi is shown in figure 5. The thick solid line in figure 5 represents the average predicted F_0 vs. the average F_0 of the syllables. These F_0 values are derived as follows: The range of the available F_0 values is uniformly divided into a finite number of bins. In this study, the range of F_0 for each bin is 10 Hz. The average F_0 corresponds to the mean of the F_0 values that fall into a particular bin. The average predicted F_0 corresponds to the mean of the predicted F_0 values for the syllables whose actual F_0 values fall into a particular bin. The average performance plot (thick solid line) in figure 5 is slightly deviated from the dashed line (diagonal), indicating that there is some prediction error at lower and higher F_0 values. For illustrating the actual prediction performance, the actual and the predicted F_0 values are jointly plotted (as 'X') on the same figure. The figure shows that the prediction is better in the range of 140 to 170 Hz, the same can be observed in the average performance plot (thick solid line) in figure 5.

The prediction performance of the duration models for the three languages is shown in figure 6. Each plot represents average predicted duration vs. average duration of a syllable. These duration values are derived as follows: The range of the available syllable durations is uniformly divided into a finite number of bins. In this study, the range of duration for each bin is 10 ms. The average duration corresponds to the mean of the syllable durations that fall into a particular bin. The average predicted duration corresponds to the mean of the predicted durations for the syllables whose actual durations fall into a particular bin. Figure 6 gives the prediction performance only in the average sense. For illustrating the actual prediction performance, a set of 1000 syllables were chosen randomly from the test set of Tamil language. Figure 7 shows the points corresponding to actual vs. predicted syllable durations. The plot shows that the prediction is better in the range of 80 to 170 ms. A similar performance can be observed in figure 6.

Apart from neural networks, other popular models used for predicting the durations of sound units are sums-of-products (SOP) models and classification and regression tree (CART) models. In the SOP model, the duration of a segment is represented as a sum of factors that affect the

Figure 5. Prediction performance of the neural network model for a male speaker in Hindi.

Figure 6. Prediction performance of the neural network models for the languages Hindi, Telugu and Tamil.

duration, and their interactions (product terms) (Santen 1994). Van Santen used SOP models for predicting the segment durations, and reported a correlation coefficient of 0.93 between predicted and observed segment durations for English, and 0.90 for German (Mixdorff & Jokisch 2001;

Figure 7. Actual vs. predicted duration values of the syllables in Tamil language.

Santen 1994). Chung used SOP models for predicting the durations of vowels and consonants in Korean language. For vowels, the root mean square error (RMSE) and the correlation coefficient values were observed to be 32.13 ms and 0.68 ms, respectively, and for consonants they were 28.86 ms and 0.54 ms, respectively (Chung 2002a,b). On TIMIT database, RMSE of 9 ms and correlation coefficient of 0.94 were observed using SOP models (Goubanova & Taylor 2000).

Riley (1992) used CART model to predict the phoneme durations using their context. Recently, CART models were used for predicting the phoneme durations and F_0 (pitch) values in the Festival framework for developing TTS systems in different languages (Black *et al* 2000). On TIMIT database, the performance of CART model was observed to be of RMSE 20 ms and correlation coefficient of 0.82 (Goubanova & Taylor 2000). For predicting the durations of vowels and consonants in Korean language, the CART model gave the RMSE of 27.51 ms for vowels and 24.2 ms for consonants, and correlation coefficient of 0.78 for vowels and 0.71 for consonants (Chung 2002a,b). Krishna & Murthy (2004) used CART models for predicting the segment durations in two Indian languages namely, Hindi and Telugu. The RMSE and correlation coefficient values were observed to be 22.86 ms, and 0.80 ms for Telugu, and 27.14 ms and 0.75 ms for Hindi, respectively.

Several models based on neural network principles are described in the literature for predicting the durations of syllables in continuous speech (Campbell 1990; Campbell & Isard 1991; Campbell 1992, 1993; Barbosa & Bailly 1994, 1992; Cordoba *et al* 1999; Hifny & Rashwan 2002; Sonntag *et al* 1997; Teixeira & Freitas 2003). Campbell (1992) used a feedforward neural network trained with feature vectors, each representing six features of a syllable. Barbosa & Bailly (1994) used a neural network model to capture the perception of rhythm in speech. Campbell (1993) reported a correlation coefficient of 0.89 between observed and predicted syllable durations. Teixeira & Freitas (2003) used neural network for predicting the segment durations, and they reported RMSE of 19.5 ms and correlation coefficient of 0.839.

Some of the prominent intonation models reported in the literature are: Tone sequence model, Fujisaki model, IPO (Institute of Perception Research) model and tilt model (Pierrehumbert 1980; Fujisaki 1983, 1988; t'Hart *et al* 1990; Taylor 2000). Among several database-driven methods, CART and neural network models are more popular (Breiman *et al* 1984; Dusterhoff *et al* 1999; Krishna *et al* 2004; Goubanova & King 2008; Cosi *et al* 2001; Tesser *et al* 2004; Vainio & Altosaar 1998; Vegnaduzzo 2003). Several models based on neural network principles are described in the literature for predicting the intonation patterns of syllables in continuous speech (Scordilis & Gowdy 1989; Vainio & Altosaar 1998; Vainio 2001; Sonntag *et al* 1997; Buhmann *et al* 2000; Hwang & Chen 1994). Scordilis & Gowdy (1989) used neural networks in a parallel and distributed manner to predict the average F_0 value for each phoneme, and the temporal variations of F_0 within a phoneme. The network consists of two levels: macroscopic and microscopic levels. At the macroscopic level, a FFNN is used to predict the average F_0 value for each phoneme. The input to the FFNN consists of the set of phonemic symbols, which represents the contextual information. At the microphonemic level, a recurrent neural network is used to predict the temporal variations of F_0 within a phoneme. Vainio & Altosaar (1998) used a three-layer FFNN to predict F_0 values for a sequence of phonemes in Finnish language. The features used for developing the models are: phoneme class, length of a phoneme, identity of a phoneme, identities of previous and following syllables (context), length of a word and position of a word. Buhmann *et al* (2000) used a RNN for developing multilingual intonation models. The features used in this study are the universal (language independent) linguistic features such as part-of-speech and type of punctuation, along with prosodic features such as word boundary, prominence of the word and duration of the phoneme.

4. Applications of prosody models

The prosody models (duration models and intonation models) can be used for different applications. In particular, prosody models are exploited for recognizing some specific categories of speech sounds in speech recognition (Shriberg & Stolcke 2001). Prosody models can be used in other applications such as speaker and language identification, TTS synthesis, expressive speech processing and voice conversion. In this section, we discuss the applicability of prosody models in speech recognition, TTS synthesis, speaker recognition and language identification. TTS synthesis application is illustrated to show the accuracy of the prosodic parameters (duration and intonation) predicted from the models. Speech, speaker and language recognition tasks are meant for demonstrating the ability of prosody models to capture some speaker, language and sound unit-specific information. The intention is not to build complete systems using only the prosodic features, but rather to use this information to enhance the performance of the conventional systems based on spectral parameters.

4.1 Text-to-speech synthesis

For TTS synthesis, prosody models (duration and intonation models) provide the specific duration and intonation information associated with the sequence of sound units present in the given text. In this study, we employ waveform concatenation for synthesizing the speech. The basis of concatenative synthesis is to join short segments of speech, usually taken from a pre-recorded database, and then impose the associated prosody by appropriate signal processing methods. We considered the syllable as a basic unit for synthesis. The database used for concatenative synthesis consists of syllables, which are extracted from carrier words spoken in isolation. We have selected nonsense words as carrier words, because they offer minimum bias towards any of the factors that affect on the basic characteristics of the syllable. These syllables are considered as neutral syllables. Using some guidelines, the carrier words can be formed for different categories of syllables. Some of the guidelines are prepared while developing the TTS system for Hindi at IIT Madras (Srikanth *et al* 1989).

In TTS synthesis, we need to synthesize the speech for the given text. First, the given text is analysed using a text analysis module, and then the positional, contextual and phonological features (linguistic and production constraints) for each of the syllables present in the given text are derived. These features are presented to the duration and intonation models (section 3.3), which will generate the appropriate duration and intonation information corresponding to the syllables.

At the synthesis stage, first, the pre-recorded (neutral) syllables are concatenated according to the sequence present in the text. The derived duration and intonation knowledge, corresponding to the sequence of syllables is incorporated in the sequence of concatenated syllables using prosody modification methods. The prosody parameters (duration and pitch) are incorporated by manipulating the instants of significant excitation of the vocal tract system during the production of speech (Rao & Yegnanarayana 2003). Instants of significant excitation are computed from the linear prediction (LP) residual of the speech signals by using the average group-delay of minimum phase signals (Smits & Yegnanarayana 1995; Murthy & Yegnanarayana 1999).

The quality (intelligibility and naturalness) of synthesized speech is evaluated by perceptual studies. Perceptual evaluation is performed by conducting subjective tests with 25 research scholars in the age group of 25–35 years. The subjects have sufficient speech knowledge for proper assessment of the speech signals. Five sentences are synthesized from text for each of the

Table 4. Ranking used for judging the quality of the speech signal.

Rating	Speech quality
1	Unsatisfactory
2	Poor
3	Fair
4	Good
5	Excellent

languages Hindi, Telugu and Tamil. Each of the subjects were given a pilot test about perception of speech signals with respect to intelligibility and naturalness. Once they were comfortable with judging, they were asked to take the tests. The tests were conducted in a laboratory environment by playing the speech signals through headphones. In the test, the subjects were asked to judge the intelligibility and naturalness of the speech. Subjects have to assess the quality on a 5-point scale for each of the sentences. The 5-point scale for representing the quality of speech is given in table 4 (Deller *et al* 1993).

The mean opinion scores (MOS) for assessing the intelligibility and naturalness of the synthesized speech in each of the languages Hindi, Telugu, Tamil and Kannada are given in columns 2 and 3 of table 5. The scores indicate that the intelligibility of the synthesized speech is fairly acceptable for all the languages, whereas the naturalness seems to be poor. Naturalness is mainly attributed to individual perception. Naturalness can be improved to some extent by incorporating the coarticulation and stress information along with duration and intonation.

For analysing the accuracy of the prediction of prosody models, we also conducted the listening tests for assessing the intelligibility and naturalness on the speech without incorporating the prosody. In this case, speech samples are derived by concatenating the neutral syllables without incorporating the prosody. The MOS of these listening tests are given in columns 4 and 5 of table 5. The MOS of the quality of the speech without incorporating the prosody have been observed to be low compared to the speech synthesized by incorporating the prosody. The significance of the differences in the pairs of the MOS for intelligibility and naturalness is tested using hypothesis testing (Hogg & Ledolter 1987). The level of confidence is high ($> 99.5\%$) in all cases (shown in columns 6 and 7 of table 5). This indicates that the differences in the pairs of MOS in each case is significant.

Table 5. Mean opinion scores for the quality of synthesized speech in the languages Hindi, Telugu, Tamil and Kannada (Intl: Intelligibility, Nat: Naturalness).

Language	Mean opinion score				Level of confidence (%)	
	TTS with prosody		TTS without prosody		Intl	Nat
	Intl	Nat	Intl	Nat		
Hindi	3.87	2.93	2.91	2.03	> 99.5	> 99.5
Telugu	4.15	3.24	3.27	2.32	> 99.5	> 99.5
Tamil	4.19	3.18	3.13	2.29	> 99.5	> 99.5
Kannada	4.19	3.18	3.13	2.29	> 99.5	> 99.5

Table 6. Performance of the speaker-specific intonation models (SD: Standard deviation, CC: Correlation coefficient).

	% syllables within 15% deviation				Mean error (Hz)				SD (Hz)				CC			
	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
S_1	92	65	83	79	16	28	25	22	14	20	19	19	0.79	0.72	0.74	0.74
S_2	74	92	67	69	27	18	31	27	24	14	24	23	0.76	0.83	0.73	0.73
S_3	68	70	85	74	26	25	19	24	21	27	15	20	0.72	0.75	0.82	0.75
S_4	72	78	78	88	25	24	25	18	21	19	19	15	0.75	0.77	0.77	0.82

4.2 Speaker specific prosody for speaker recognition

For demonstrating the evidence of speaker-specific information captured by the prosodic models, speech data of four different female speakers of the Hindi language are considered. Separate models are developed to capture the speaker-specific intonation and duration knowledge from the given labelled speech data. The speaker models are denoted by S_1 , S_2 , S_3 and S_4 . The models were tested with four sets of test data, where each set belongs to one of the speakers. The outcome of the models for each set of the test data is given in tables 6 and 7. Row 1 in table 6 indicates the performance of the intonation models (S_1 – S_4) tested with S_1 test data. Similarly, the numbers in rows 2–4 of table 6 indicate the performance of the models tested with the test data of the speakers 2–4 (S_2 – S_4), respectively.

The results show that the percentage of syllables predicted within 15% deviation from the actual F_0 value is highest for the intended speaker compared to others. Likewise, the average prediction error (μ), standard deviation (σ) (SD) values are low, and the correlation coefficient (γ) (CC) values are high for the intended speaker. This can be observed in the elements of the principle diagonal for each of the performance measures. It indicates that the intonation models (S_1 – S_4) capture the knowledge of speaker-specific F_0 values. This information can be exploited in enhancing the performance of speaker recognition system using the conventional spectral features.

A similar study is performed using duration models separately for each of the speakers. The performance of the speaker-specific duration models is given in table 7. Similar performance is observed as that of the speaker-specific intonation models, although the intonation models seem to capture speaker-specific information better than the duration models (Rao & Yegnanarayana 2009).

Table 7. Performance of the speaker-specific duration models.

	% syllables within 15% deviation				Mean error (ms)				SD (ms)				CC			
	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
S_1	58	51	53	53	33	38	37	38	27	30	29	31	0.76	0.69	0.69	0.70
S_2	53	58	52	51	35	31	37	37	29	24	27	27	0.70	0.74	0.70	0.69
S_3	53	52	61	52	36	35	29	34	29	27	23	26	0.71	0.70	0.75	0.71
S_4	54	54	56	60	36	35	35	29	27	28	28	24	0.72	0.71	0.68	0.76

Table 8. Performance of the language-specific intonation models (Hi: Hindi, Te: Telugu, Ta: Tamil and Ka: Kannada).

	% syllables within 15% deviation				Mean error (Hz)				SD (Hz)				CC			
	Hi	Te	Ta	Ka	Hi	Te	Ta	Ka	Hi	Te	Ta	Ka	Hi	Te	Ta	Ka
Hi	91	57	62	58	11	23	20	22	9	14	13	14	0.79	0.73	0.68	0.63
Te	44	88	64	69	19	10	17	16	13	9	13	12	0.73	0.79	0.72	0.73
Ta	58	63	84	65	19	18	12	20	14	14	10	13	0.74	0.75	0.81	0.73
Ka	55	66	64	89	20	17	18	11	15	14	13	9	0.68	0.71	0.73	0.81

4.3 Language specific prosody for language recognition

For demonstrating the evidence of language-specific information captured by the prosodic models, speech data of male speakers in Hindi (Hi), Telugu (Te), Tamil (Ta) and Kannada (Ka) was chosen. Separate models were developed to capture the language-specific intonation knowledge from the given labelled speech data. The models were tested with four sets of the test data, where each set belongs to one of the languages. The performance of the models on each set of test data is given in table 8. The results show that the intonation models also capture the language-specific knowledge.

For demonstrating the ability of the duration models to capture the language-specific information, separate models are developed for each of the languages. The performance of the language-specific duration models is given in table 9. The results show that the performance of the models is superior, when the test data are from the same language.

4.4 Syllable-specific prosody for speech recognition

For demonstrating the evidence of syllable-specific information captured by the prosodic models, speech data of syllables in Hindi (Hi) were chosen. For this study, syllables were categorized into five groups based on the vowel category (/a/, /e/, /i/, /o/ and /u/). Separate models were developed to capture the syllable-specific intonation and duration knowledge from the given labelled speech data. The models were tested with five sets of the test data, where each set belongs to one of the above-mentioned syllable categories. The syllable models are denoted by the vowel identity. The performance of the models to each set of test data is given in table 10. The results show that the intonation models also captures the syllable-specific knowledge.

Table 9. Performance of the language-specific duration models (Hi: Hindi, Te: Telugu, Ta: Tamil and Ka: Kannada).

	% syllables within 15% deviation				Mean error (Hz)				SD (Hz)				CC			
	Hi	Te	Ta	Ka	Hi	Te	Ta	Ka	Hi	Te	Ta	Ka	Hi	Te	Ta	Ka
Hi	70	59	58	56	31	38	37	39	24	29	30	30	.77	.71	.69	.68
Te	50	68	59	57	37	28	32	34	27	23	26	27	.74	.79	.75	.73
Ta	56	63	76	58	36	33	25	31	27	26	21	26	.74	.75	.82	.74
Ka	54	59	61	69	38	34	35	28	27	26	27	22	.73	.74	.72	.80

Table 10. Performance of the syllable-specific intonation models.

	% syllables within 15% deviation					Mean error (Hz)					SD (Hz)					CC				
	<i>a</i>	<i>e</i>	<i>i</i>	<i>o</i>	<i>u</i>	<i>a</i>	<i>e</i>	<i>i</i>	<i>o</i>	<i>u</i>	<i>a</i>	<i>e</i>	<i>i</i>	<i>o</i>	<i>u</i>	<i>a</i>	<i>e</i>	<i>i</i>	<i>o</i>	<i>u</i>
<i>a</i>	86	75	79	81	80	14	17	17	18	16	12	14	15	15	16	.84	.79	.81	.80	.79
<i>e</i>	77	89	83	84	82	16	13	15	16	17	15	11	14	16	15	.78	.89	.83	.82	.80
<i>i</i>	79	80	86	81	82	17	16	13	18	17	14	15	11	15	15	.80	.78	.84	.79	.77
<i>o</i>	80	81	82	88	83	17	16	16	10	14	15	14	15	10	13	.77	.79	.80	.89	.84
<i>u</i>	80	79	81	85	90	16	17	16	14	11	15	16	16	14	11	.77	.79	.79	.83	.85

For demonstrating the ability of the duration models to capture the syllable-specific information, separate models are developed for each category of the syllables. The performance of the syllable-specific duration models is given in table 11. From this study, it is observed that duration models capture more syllable specific information compared to intonation models (see tables 10 and 11).

Results from tables 6 to 11 indicate that the prosody models capture some speaker, language and syllable-specific information. This information can be used in enhancing the performance of the conventional speech recognition, speaker recognition and language identification systems.

Prosody models are used extensively in literature for developing speech systems (Rao 2011). Batliner *et al* (2001) have explored the prosodic models in speech processing for automatic speech understanding and speech synthesis. They have used the Tones and Break Indices intonation model for representing the positions of accents and phrase boundaries in both automatic speech understanding and TTS. Ostendorf *et al* (2003) have designed and developed a formal model for incorporating prosody for conversational speech recognition for improving the word recognition rate. Hwang & Chen (1995) have proposed neural network-based prosodic model to explore the hidden pronunciation states embedded in the input text for assisting the prosody synthesis in Mandarin TTS. Shriberg & Stolcke (2004) have used the computational prosody modelling for a variety of tasks in automatic speech recognition and understanding. Vainio has proposed artificial neural network-based prosody models for predicting the duration patterns, pitch and energy contours, and further the predicted prosody information is used for developing TTS system for Finnish language (Vainio 2001; Vainio & Altosaar 1998). Krishna *et al* (2002) have used CART-based prosody models for developing TTS system in Indian languages. Kishore *et al* (2002) have used the CART-based prosody models to select the units satisfying the local phonetic constraints and prosodic constraints for developing TTS system for Hindi. Yin1 *et al* (2006) have used prosodic features along with cepstral features for language identification (LID).

Table 11. Performance of the syllable-specific duration models.

	% syllables within 15% deviation					Mean error (Hz)					SD (Hz)					CC				
	<i>a</i>	<i>e</i>	<i>i</i>	<i>o</i>	<i>u</i>	<i>a</i>	<i>e</i>	<i>i</i>	<i>o</i>	<i>u</i>	<i>a</i>	<i>e</i>	<i>i</i>	<i>o</i>	<i>u</i>	<i>a</i>	<i>e</i>	<i>i</i>	<i>o</i>	<i>u</i>
<i>a</i>	52	44	35	29	26	28	32	33	34	34	24	27	28	27	29	.79	.71	.72	.69	.67
<i>e</i>	42	59	39	41	46	33	26	31	33	35	29	22	27	28	29	.72	.79	.69	.68	.73
<i>i</i>	39	43	55	41	36	33	31	27	33	32	28	27	21	26	28	.70	.68	.77	.71	.69
<i>o</i>	39	42	37	51	36	35	33	31	24	30	28	26	27	22	29	.68	.67	.70	.79	.71
<i>u</i>	42	38	37	39	49	32	34	31	29	24	28	29	28	27	21	.67	.71	.70	.69	.77

By combining pitch, intensity and Mel frequency cepstral coefficients (MFCC) features properly, the LID system performance is improved by 12% over the features with cepstral features alone. Leung *et al* (2008) have explored prosody models along with the conventional cepstral features for speaker recognition. Fusion of spectral and prosody systems have shown the improvement in the performance of the speaker recognition system. Weber *et al* (2002) have used the prosodic and lexical information for speaker identification task at National Institute of Standards and Technology 2001 speaker identification evaluation. The results show that with sufficient test and training data, suprasegmental information can significantly enhance the performance of traditional speaker identification system. Mary & Yegnanarayana (2008) have proposed prosodic features, which can be used jointly for performing speaker and language identification tasks.

5. Voice conversion

The basic goal of voice conversion is to transform the characteristics of an input (source) speech signal such that the output (transformed) signal is perceived to be produced by another (target) speaker. Voice transformation is generally performed in two steps. In the first step, the training stage, a set of speech feature parameters of both the source and target speakers are extracted, and appropriate mapping rules that transform the parameters of the source speaker onto those of the target speaker are generated. In the second step, the transformation stage, the features of the source signal are transformed using mapping rules developed in the training stage so that the synthesized speech possesses the personality of the target speaker (Lee 2007).

To implement voice transformation, two problems have to be considered: what features are extracted from the underlying speech signals, and how to modify these features in such a way so that the transformed speech signals mimic target speakers' voices. The first problem can be solved by identifying the speaker-specific features from the given speech signals. It is known that the shape of the glottal pulse, vocal tract transfer function and the prosodic features uniquely characterize the speaker (Yegnanarayana *et al* 2001b).

In this study, we used neural network models for mapping the vocal tract characteristics, excitation source characteristics and prosodic characteristics. The main reason for exploring neural network models for developing the mapping functions at different levels (source, system and prosodic levels) is that they are good at capturing the non-linear relations present in the feature patterns of source and target speakers.

In the proposed voice conversion system, vocal tract parameters are extracted using pitch synchronous analysis. In the pitch synchronous approach, the LP analysis is performed for each pitch period. The instants of significant excitation are used as pitch markers in the voiced region, and the speech signal between two successive instants is considered as pitch period. The advantage of this method is that, the end points of the speech segment considered for the analysis are the same for all the segments (i.e., all speech segments begin with the instant of glottal closure, and terminate with its following instant of glottal closure) (Yegnanarayana & Veldhuis 1998), whereas in the case of conventional methods (i.e., using block processing), the end points of the segments are different, because of the fixed block size used in the analysis (Sundermann 2005; Turk 2007; Lee 2007; Narendranadh *et al* 1995; Rao & Yegnanarayana 2006b; Toda *et al* 2001; Abe *et al* 1998; Stylianou *et al* 1998; Percybrooks & Moore-II 2008; Kain 2001; Ye & Young 2004; Turk & Arslan 2005, 2006). With this, the extracted vocal tract parameters using conventional methods are not consistent. As the segment considered for parameter extraction consists of multiple pitch periods, the estimated vocal tract parameters will be difficult to attribute to a specific region of the segment. Although, we assume that the speech signal in the voiced region

is stationary over a period of 20–30 ms, yet there exists a non-stationary behaviour from one pitch cycle to its adjacent pitch cycle. In this case, for accurate estimation of vocal tract parameters, pitch synchronous analysis will be the right choice. Even within a pitch period, there exists a change in the vocal tract characteristics between glottal closure and glottal open phases. However, it is difficult to mark the time instant between the closure and open phases of the glottis. Hence, we cannot extract the vocal tract parameters for the segments whose size is less than the pitch period. Therefore, in this study, we have proposed pitch synchronous analysis for extracting the vocal tract parameters from the speech segments of the source and target speakers.

At the excitation source level, speaker-specific information is present around the instant of glottal closure. Therefore, the LP residual signal around the instant of glottal closure is used as an excitation source feature for deriving the mapping function. As the excitation source signal contains only higher order correlations, neural network models are used for deriving the mapping function at this level.

In this study, we have considered duration, intonation (pitch contour) and energy (intensity) patterns of the sequence of sound units as the prosodic features for characterizing the speaker. For Indian languages, syllables are the most suitable basic units for the analysis and synthesis of speech. Syllables implicitly capture the duration patterns, shape of the vocal tract and coarticulation effects (Rao & Yegnanarayana 2007). Although the global characteristics of the prosody patterns depend on the nature of the utterance, the local variations (rise–falls) depend on the nature of the individual syllables. Therefore, for mapping the prosody patterns from source speaker to target speaker, we use the mapping functions for capturing both global and local variations in the prosodic patterns. For capturing the prosodic patterns at global level, we use the prosodic features extracted from the phrase (utterance) level. For deriving the mapping functions at local level, the prosodic features extracted from syllable level segments are used. Separate mapping functions are employed for mapping each of the prosodic features. Neural network models are explored for deriving the mapping functions at prosodic level.

5.1 *Mapping of the vocal tract characteristics*

In this work, line spectral frequencies (LSFs) are used for representing the vocal tract information. These LSFs are derived from the LP analysis through LP coefficients (LPCs) (Mei & Sun 2000). LSFs are known to be more stable compared to LPCs for deriving the mapping function between source and target speakers. At synthesis stage, again LPCs are used for synthesizing the target speaker speech (i.e., LPCs are derived from the target speaker LSFs).

For deriving the mapping function using neural network, the network has to be trained with the LSFs derived from the spoken utterances of source and target speakers. For this study, we prepared the text transcription for 500 Hindi sentences. These 500 sentences are recorded by 2 male and 2 female speakers. The durations of these sentences are varying between 3 and 5 s. Each sentence has roughly 15–20 syllables.

In the pitch synchronous analysis, the vocal tract parameters are estimated by performing the LP analysis on each pitch cycle. The instants of significant excitation are used to mark the pitch cycles of the speech signal. In this paper, we have used 10th-order LP analysis for extracting the LPCs from each pitch cycle. The corresponding LSFs are derived from the LPCs (Mei & Sun 2000). For extracting the vocal tract parameters in the nonvoiced regions, the LP analysis is performed with a fixed frame size of 20 ms.

To capture the relationship between the vocal tract shapes of source and target speakers, we need to feed the time-aligned vocal tract features of source and target speakers to the input and

Table 12. Steps for deriving the time-aligned LPCs.

-
1. Derive the syllable pair from the utterance pair spoken by the pair of speakers using VOPs.
 2. Preemphasize the speech segments corresponding to the syllable pair.
 3. Obtain the epoch sequence for the syllable pair
 4. Compute LPCs with 10th-order LP analysis for the speech signal within a pitch cycle using the knowledge of the epoch sequence associated to the syllable pair.
 5. Apply DTW on the syllable pair (at epoch level) to derive the matching epochs.
 6. Select the LPCs corresponding to the matching epochs of the syllable pair and they become pairs of the time-aligned LPCs.
 7. Same process (steps 1–6) is repeated for other syllable pairs.
-

output of the neural network, respectively. In this study, we used dynamic time warping (DTW) to derive the time-aligned vocal tract features (Itakura 1975). The procedure for deriving the time-aligned LPCs is given in table 12.

Syllable boundaries are automatically determined using vowel onset points (VOP). Here, VOPs for the given speech signal are determined from its LP residual (Prasanna & Zachariah 2002; Prasanna & Yegnanarayana 2004b). The procedure for detecting the VOPs in the speech signals is summarized in table 13.

The above procedure is illustrated for the Hindi sentence '*antarAshtriyA basseva pichale mahlne shuru huyi thi*' in figure 8. In this sentence, there are 16 VOP events, as marked (manually) in figure 8a. The Hilbert envelope and the VOP evidence plots are shown in figure 8b and c, respectively. The output of the peak picking algorithm is given in figure 8d. The hypothesized VOPs after eliminating the spurious ones are shown in figure 8e. We can interpret the vowel onset point as the junction point between consonant and vowel of a CV unit. The region to the left of the VOP is considered as the consonant region, and to the right of the VOP as the vowel region. In the vowel portion, a small region following the VOP is treated as transition region (Gangashetty *et al* 2004). After determining the vowel onset point, 30 ms to the left of VOP is marked as consonant region. With this knowledge, we can determine the syllable boundaries as the time instants, which are 30 ms ahead of the VOPs.

The performance of the proposed voice conversion system depends on the accuracy of the VOP detection process. For clean speech, the accuracy of VOP detection is about 95% (Prasanna 2004). For developing the mapping function accurately, we have considered the syllables derived using VOPs with high confidence. The syllables of this category is about 78%. The impact of the accuracy of VOP detection is also observed in perceptual analysis (using subjective

Table 13. Steps for detection of the VOP events.

-
1. Preemphasize input speech.
 2. Compute LP residual with 10th-order LP analysis, with a frame size of 20 ms and shift of 5 ms.
 3. Compute Hilbert envelope of the LP residual.
 4. Obtain the VOP event evidence plot from the Hilbert envelope by convolving it with the Gabor filter.
 5. Identify the peaks in the VOP event evidence plot using peak picking algorithm.
 6. For each peak, if there is no negative region before the next peak, then eliminate such a peak as spurious.
 7. Eliminate peaks which are at a distance less than 50 ms with respect to their next peak.
 8. Hypothesize remaining peaks as the VOP events.
-

Figure 8. Hindi sentence ‘*antarAshtriyA bassevA pichale mahIne shuru huyi thi*’ (a) Speech waveform with manual marked VOP events, (b) Hilbert envelope of LP residual, (c) VOP evidences, (d) output of peak picking algorithm, and (e) hypothesized VOP events after eliminating some spurious peaks.

listening tests), as well as objective analysis. The objective evaluation of the predicted vocal tract characteristics is described at the end of this section. The subjective evaluation is performed in section 5.6.

The neural network model used in this work is a five-layer symmetric feedforward neural network, and its basic structure is shown in figure 2. The basic reason for choosing the five layers is to support the symmetric structure of the network. It is well known that feedforward neural network with symmetric structure is used specifically for capturing the associations present in the input and output features (Haykin 1999). Here, the network is supposed to provide the mapping function between source and target speaker parameters at different levels. The mapping function is between the 10-dimensional input and the output vectors. The 10-dimensional input and output vectors correspond to the time-aligned frame LSFs. Several network structures are explored in this study. The (empirically optimized) final structure of the network is 10L 20N 5N 20N 10L.

A separate model is developed for each pair of the speakers. Ten-dimensional input and output feature vectors consist of 10 LSFs derived from the speech frames of source and target speakers. As the number of training samples are significantly larger than the number of weights,

the possibility of over-fitting of the training data is rare. Also, the training is stopped after about 500 epochs in each case, as there is no significant reduction in the error for further increase in the number of epochs. After the training phase, the weights in the network represents the mapping function between input and output. The performance of the model can be evaluated by both subjective and objective measures. Subjective evaluation consists of synthesizing the speech for the desired speaker using LSFs derived from the neural network model, and conducting listening tests to assess the desired speaker characteristics present in the synthesized speech. Objective measures consist of analysing the spectral envelopes derived from the predicted and actual LPCs. Figure 9 shows the predicted and actual spectral envelopes of the target speaker. The spectral envelopes in the figure indicate that the neural network model used for mapping the vocal tract features succeeded in transforming the source speaker's vocal tract features close to the target speaker's actual vocal tract features.

The spectral envelopes are mainly characterized by significant frequencies, which have dominant energy over other frequencies. These frequencies are also known as formant frequencies. In general, for speech segment the spectral envelope is characterized by four formant frequencies (f_1 , f_2 , f_3 and f_4). The prediction accuracy of the spectral envelope can be evaluated by computing the objective measures such as deviation, RMSE and CC using actual and predicted formant frequency values.

For evaluating the performance of the mapping function, 1500 syllables (15,000 speech frames) are considered. For each speech frame, the desired speaker's LSFs are predicted, and from that the corresponding LPCs are derived. Formant frequencies are derived from the roots of the LPC polynomial. The objective measures are tabulated separately for each of the mapping

Figure 9. Predicted and actual spectral envelopes of the target speaker.

function associated to each pair of speakers. Tables 14, 15, 16 and 17 carry the objective measures for M1–F1, F2–M2, M1–M2 and F1–F2 mapping functions, respectively. In table 14, first column shows the formant frequencies. Columns 2–5 indicate the percentage of speech frames, predict the formant frequencies are predicted within the specified deviation. Columns 6 and 7 indicate the RMSE (μ_{RMSE}) and CC ($\gamma_{X,Y}$), respectively.

Table 14. Performance of the FFNN model for predicting the formant values for M1–F1 (M1: male speaker-1 and F1: female speaker-1) transformation.

Formants	% Predicted frames within deviation				μ_{RMSE} (%)	$\gamma_{X,Y}$
	2%	5%	10%	25%		
f1	24	58	82	97	7.27	0.72
f2	29	69	89	98	5.73	0.81
f3	46	83	95	100	3.13	0.89
f4	63	94	98	100	2.36	0.92

Table 15. Performance of the FFNN model for predicting the formant values for F2–M2 (F2: female speaker-2 and M2: male speaker-2) transformation.

Formants	% Predicted frames within deviation				μ_{RMSE} (%)	$\gamma_{X,Y}$
	2%	5%	10%	25%		
f1	19	49	77	96	8.13	0.73
f2	27	62	87	98	6.37	0.79
f3	39	78	96	99	4.05	0.84
f4	57	91	97	100	3.22	0.89

Table 16. Performance of the FFNN model for predicting the formant values for M1–M2 (M1: male speaker-1 and M2: male speaker-2) transformation.

Formants	% Predicted frames within deviation				μ_{RMSE} (%)	$\gamma_{X,Y}$
	2%	5%	10%	25%		
f1	28	68	87	99	6.88	0.77
f2	34	76	93	99	5.19	0.83
f3	51	87	96	99	3.44	0.87
f4	69	92	98	100	2.12	0.91

Table 17. Performance of the FFNN model for predicting the formant values for F1–F2 (F1: female speaker-1 and F2: female speaker-2) transformation.

Formants	% Predicted frames within deviation				μ_{RMSE} (%)	γ
	2%	5%	10%	25%		
f1	31	72	91	98	6.48	0.79
f2	39	81	95	99	4.93	0.85
f3	53	86	96	100	3.26	0.89
f4	67	93	97	100	2.52	0.90

The prediction performance of the neural network model for transforming the vocal tract characteristics is illustrated using scatter plots. For generating these plots, 10 sentences (1000 speech frames) are chosen randomly from the test data. These plots are generated by plotting the predicted and actual formant frequencies derived from the chosen speech frames jointly. Figure 10 illustrates the prediction performance of vocal tract characteristics in terms of formant frequencies for M1–F1 transformation. The orientation of the data points in the subplots of the figure indicate the closeness of the prediction towards the desired (target) values. In the ideal case (for zero prediction error), all the data points are expected to lie on the diagonal.

5.2 Mapping of excitation source characteristics

The LP residual signal derived from the LP analysis can be viewed as excitation source signal for the given speech segment. As signal autocorrelation at lags of 1–10 samples is zeroed out by an LPC inverse filter, the resultant excitation signal contains the remaining higher order dependencies, as well as correlations at lags higher than 10th order. In this section, we will develop the mapping function between these higher order relations associated to source and target speakers.

In the LP residual signal, the region around the instant of glottal closure corresponds to the significant part of the excitation, in the sense that the strength of excitation is maximum in that region of the pitch period. Hence, by retaining those samples, the naturalness of the original

Figure 10. Scatter plots for the formant frequencies f1, f2, f3 and f4 between target speaker's actual and predicted values for M1–F1 transformation.

voice can be preserved. The glottal closure region within the pitch period is very crucial in view of deriving the stable vocal tract information. During the closed phase of the glottal cycle, the vocal tract will be completely isolated, and the resonances during this phase will be the true resonances of the vocal tract system. By comparison in the open phase, the vocal tract system is coupled to the trachea, lungs and chest cavity, therefore the resonances during this portion will not be accurate. In the LP analysis the filter coefficients can predict the speech samples well during the open phase, but the prediction will be poor near the instant of glottal closure due to the sudden closure of the vocal folds which generates an impulse of higher strength. Therefore, during the LP analysis, the prediction error is high in the closure region, i.e., characteristics of the vocal tract during the glottal closure are not properly captured through LPCs. Hence, the region around the instant of glottal closure in the LP residual signal is crucial in view of preserving the remains of the vocal tract information after the LP analysis. In the glottal open region, the strength of the residual samples is very low and they have random behaviour. Hence, they are not significant in view of perception.

In this study, we have explored two neural network models for mapping the excitation source characteristics between the speakers. The first model is developed by using the LP residual from the glottal closure region. The second model is developed by using the LP residual from the entire pitch cycle (both open and closure regions). The performances of the models are evaluated using synthesis experiments. The evaluation results indicate that the performance of both models is similar. It is observed that the difference in the mean opinion scores of the two models is not

Figure 11. (a) A segment of LP residual of source speaker, (b) the associated segment of LP residual of target speaker, and (c) the predicted LP residual from the source LP residual shown in (a), using the mapping function (neural network model).

significant. As the computational complexity of the first model is low compared to the second model, we have considered the first model (i.e., the LP residual samples within the glottal closure region are used for mapping) for developing the mapping function for transforming the excitation source characteristics. Several model structures are examined in this study. The final network structure (empirically optimized) used for mapping the excitation source information is 24L 48N 12N 48N 24L.

Mapping of the LP residual (excitation source information) between source and target speakers using a neural network is illustrated in figure 11. In the figure, the plot(a) indicates a segment of LP residual of the source speaker, plot(b) indicates the LP residual of target speaker corresponding to the LP residual segment in plot(a), and plot(c) is the derived (predicted) LP residual using the mapping function. From the figure, it is observed that the predicted LP residual is very close to the desired LP residual of the target speaker (see figures 11b and c). This result indicates that the proposed mapping function using neural network model has captured the higher order relations present in the excitation source (LP residual) of the source and target speakers. The prediction accuracy of the LP residual is further analysed using objective measures such as RMSE (μ_{RMSE}) and CC ($\gamma_{X,Y}$). For deriving these objective measures, we have considered about 1000 syllables, which contains about 17,000 epoch intervals (pitch periods). For each epoch interval, RMSE and CC are computed using predicted and actual LP residual segments. For each mapping function, RMSE and CC are determined by averaging over all the epoch intervals. Table 18 shows the objective measures to indicate the accuracy of prediction for deriving the LP residual of the desired speaker using their associated mapping functions.

5.3 Mapping of prosodic parameters

In this section, we discuss the mapping function for transforming the intonation, duration and energy patterns between source and target speakers.

5.3a Mapping of intonation patterns: Intonation characteristics of a speaker are represented by the dynamics of pitch (pitch contour), and they are characterized by gross level and fine level variations. The gross level intonation pattern depends on the semantics and the linguistic meaning of the utterance. The finer level intonation pattern indicates the prominence of the individual words and syllables. In this study, two neural network models are used for developing the mapping functions for transforming the global and local (fine) level intonation patterns.

The pitch contours of the speech utterances of source and target speakers are derived using instants of significant excitation of speech signal. The epoch interval (i.e., the time interval between the successive epochs) of each epoch represents the pitch period. The pitch contour can be derived from the epoch interval contour, by transforming each epoch interval into its associated pitch. For developing the mapping function using a neural network, the pitch contours of

Table 18. Performance of the FFNN model for predicting the LP residual under different transformations.

Transformation	RMSE (μ_{RMSE})	CC ($\gamma_{X,Y}$)
M1-F1	17.62	0.66
F2-M2	19.14	0.62
M1-M2	15.81	0.71
F1-F2	14.47	0.74

the source and target speakers should have the same length. Therefore by using resampling, the source and target pitch contours are resized to 30-dimensional feature vectors. Here, the size of the feature vector is not critical. Each of these feature vectors represents the gross shape of the pitch contour of the whole utterance. The structure of the neural network model used for developing this mapping function is 30L 60N 15N 60N 30L.

An example of the mapping of intonation patterns at the gross level (utterance level) between source and target speakers is shown in figure 12. From the figure, it is observed that the predicted pitch contour for the target speaker is close to the desired one, and it also follows the rise–fall patterns of the original target speaker’s pitch contour. This shows that the chosen neural network model has successfully captured the hidden relations at gross level in mapping the pitch contour between source and target speakers. The prediction accuracy of gross level intonation contours is further analysed by using RMSE and CC. These measures are computed at utterance level. About 100 utterances (sentences) are used for deriving the objective measures. Second and third columns of table 19 shows the RMSE and CC values derived from each of the transformations. These are the averaged values derived from the 100 utterances.

The other mapping function is used for transforming the local intonation patterns (fine level). The local intonation patterns refer to the pitch contours at the syllable level. Pitch contours at the syllable level are derived by segmenting the pitch contour of the utterance using syllable boundaries. Here, syllable boundaries can be determined using the knowledge of vowel onset points. Then, the mapping function for transforming the pitch contours at syllable level can be developed by feeding the syllable level pitch contours of source and target speakers. The network structure used for developing this model is 10L 20N 5N 20N 10L. The prediction accuracy of the local intonation patterns is analysed similar to gross intonation patterns, by determining the objective measures for 1000 syllables. The objective measures for different transformations are shown in 4th and 5th columns of table 19.

Figure 12. An example mapping of the intonation patterns at gross level (utterance level) between source and target speakers.

Table 19. Performance of the FFNN models for predicting the intonation contours at gross and fine levels under different transformations.

Transformation	Gross level		Fine level	
	RMSE (μ_{RMSE})	CC ($\gamma_{X,Y}$)	RMSE (μ_{RMSE})	CC ($\gamma_{X,Y}$)
M1–F1	13.23	0.76	12.54	0.78
F2–M2	15.46	0.73	13.83	0.79
M1–M2	12.87	0.79	10.12	0.82
F1–F2	10.93	0.82	9.37	0.84

5.4 Mapping of duration patterns

The duration patterns of the sequence of sound units and pauses are specific to the speaker. Therefore, we are developing two mapping functions to transform the duration patterns of syllables as well as pauses (silence regions) between the words or syllables.

Syllable durations can be obtained using the knowledge of VOPs of the speech utterances. Here, nonspeech regions indicate the pauses, and they are determined using energy threshold and speech-specific knowledge such as pitch periodicity and spectral characteristics. Neural network models are used for developing the mapping functions, and they expect the fixed dimensional input and output patterns for training the models. In the database, the maximum number of syllables present in the utterance is 20. For accommodating these utterances, a feature vector size of 20 is fixed for developing the mapping function for transforming the syllable durations of the utterances. Similarly, a 10-dimensional feature vector is used to represent the duration values of the pauses present in the utterance. If the number of syllables and pauses are less than the size of the feature vector, the remaining entries of the feature vector are represented by zeros. The network structures used in developing the mapping functions for transforming the syllable and pause durations are 20L 40N 10N 40N 20L and 10L 20N 5N 20N 10L, respectively.

The prediction accuracy of the duration patterns of the syllables as well as pauses is analysed using objective measures such as percentage of syllables or pauses predicted within the specified deviation, RMSE and CC. In this study, about 1500 syllables and 500 pauses derived from 100 sentences are used for computing the objective measures. Table 20 shows the performance of the different transformations for predicting the durations of the syllables. Table 21 shows the performance of the different transformations for predicting the durations of the pauses.

Table 20. Performance of the FFNN models for predicting the durations of syllables under different transformations.

Transformation	% Predicted syllables within deviation				μ_{RMSE} (%)	$\gamma_{X,Y}$
	2%	5%	10%	25%		
M1–F1	6	32	56	88	15.34	0.82
F2–M2	9	35	61	91	13.76	0.85
M1–M2	8	29	49	90	16.12	0.79
F1–F2	13	39	67	93	11.88	0.83

Table 21. Performance of the FFNN models for predicting the durations of pauses using different transformations.

Transformation	% Predicted pauses within deviation				μ_{RMSE} (%)	$\gamma_{X,Y}$
	2%	5%	10%	25%		
M1–F1	8	38	67	92	11.27	0.79
F2–M2	12	43	71	94	10.83	0.82
M1–M2	9	31	53	90	13.58	0.77
F1–F2	14	46	77	93	10.26	0.84

5.5 Mapping of energy patterns

The energy patterns present in the speech signal indicate the strength of voice specific to the speaker. The energy contours of the speech utterances can be characterized at the utterance and syllable levels to capture the strength of voice at global and local (gross and fine) levels. In this study, two neural network-based mapping functions are developed for transforming the energy contours at gross and fine levels between source and target speakers. The methodology for developing these two mapping functions follows similar procedures to those which we have used for developing the mapping functions for transforming the intonation patterns. The objective measures for different transformations for predicting the gross and fine level energy patterns are given in table 22.

5.6 Synthesis and evaluation of target speaker voice

In the previous sections, we have described methods for mapping the vocal tract, excitation source and prosodic characteristics between source and target speakers. In this section, we discuss about the synthesis of target speaker's speech from source speaker's speech by using mapping functions, and then evaluating the presence of desired speaker characteristics in the synthesized speech. The detailed block diagram for representing the proposed voice conversion process is shown in figure 13. The transformation of source speaker's speech to target speaker's speech is performed as follows:

- (i) The instants of significant excitation and vowel onset points are determined for the utterance of the source speaker.
- (ii) The vocal tract information (i.e., LSFs) and the excitation source information (LP residual) of the source speaker's speech are derived by performing the pitch synchronous LP analysis using the knowledge of the instants of significant excitation.

Table 22. Performance of the FFNN model for predicting the energy contour at gross and fine levels under different transformations.

Transformation	Gross level		Fine level	
	RMSE (μ_{RMSE})	CC ($\gamma_{X,Y}$)	RMSE (μ_{RMSE})	CC ($\gamma_{X,Y}$)
M1–F1	10.85	0.79	9.26	0.80
F2–M2	12.43	0.75	10.17	0.77
M1–M2	9.89	0.81	7.92	0.88
F1–F2	8.56	0.87	8.98	0.81

Figure 13. Block diagram of proposed voice conversion system for transforming the source speaker's speech to target speaker's speech.

- (iii) The vocal tract information and the excitation source information of the target speaker are derived from the mapping functions, by giving the LSFs and LP residual of the source speaker as input to the corresponding mapping functions. In this paper, all the mapping functions are realized by using 5-layer FFNN models.
- (iv) The syllable and pause durations of the utterance spoken by the source speaker are derived using VOPs and speech-specific knowledge.
- (v) The syllable and pause durations of the target speaker are derived from their respective mapping functions, by giving the syllable and pause durations of the source speaker's speech as input to the corresponding mapping functions.
- (vi) The pitch contours at the utterance and syllable levels are derived from the source speaker's speech using the knowledge of instants of significant excitation and vowel onset points.
- (vii) The pitch contours at the utterance and syllable levels are derived for the target speaker using their respective mapping functions, by giving the syllable and utterance level pitch contours of the source speaker's speech as input to the corresponding mapping functions.
- (viii) Using the derived pitch contours and duration patterns of target speaker, the epoch sequence for the target speaker's speech is generated.
- (ix) The target speaker's LP residual (derived in step iii) is modified according to the target speaker's epoch sequence (derived in step viii).
- (x) The LSFs of the target speaker (derived in step iii) are updated according to the target speaker's epoch sequence (derived in step viii), and the corresponding LPCs are derived for synthesizing the target speaker's speech.
- (xi) The target speaker's speech is synthesized by pitch synchronous LP analysis (exciting the time varying filter represented by the LPCs of the target speaker with the modified LP residual according to the target speaker's prosody) using the knowledge of target speaker's epoch sequence (derived in step viii).
- (xii) The energy patterns of the synthesized speech at the utterance and syllable levels are scaled according to the target speaker's energy patterns at the utterance and syllable levels, respectively. The target speaker's energy patterns are derived from their respective mapping functions, similar to the intonation patterns derived for the target speaker.

In this study, the desired prosody (target speaker's intonation and duration patterns) modification is carried out in LP residual domain (Rao & Yegnanarayana 2006a,b). The performance of the mapping functions can be evaluated by using subjective and objective measures. The basic goal is the voice transformation, therefore the mapping functions are evaluated by using perceptual tests (i.e., by conducting listening tests). Four separate mapping functions are developed for transforming the speaker voices: (i) male to female (M1–F1), (ii) female to male (F2–M2), (iii) male to male (M1–M2) and (iv) female to female (F1–F2). Here M1, M2, F1 and F2 represent male and female speaker voices present in the database. For each case, five utterances are synthesized using their associated mapping functions. Listening tests are conducted to assess the desired (target) speaker characteristics, and the level of distortion present in the synthesized speech. The recorded speech utterances of the target speaker corresponding to the synthesized speech utterances are made available to the listeners to judge the relative performance. Twenty students have participated in these tests. Each of the synthesized speech utterance is played to the listener, after playing its original recorded utterance. The subjects were asked to give their opinion score on a 5-point scale. The rating 5 indicates the excellent match between the original target speaker speech and the synthesized speech (i.e., synthesized speech is close to the original speech of the target speaker). The rating 1 indicates a very poor match between the original and synthesized utterances, and the other ratings indicate different levels of deviation between 1

Table 23. Ranking used for judging the quality and distortion of the synthesized speech signal.

Rating	Speech quality	Level of distortion
1	Unsatisfactory	Very annoying and objectionable
2	Poor	Annoying but not objectionable
3	Fair	Perceptible and slightly annoying
4	Good	Just perceptible but not annoying
5	Excellent	Imperceptible

and 5. The level of distortion in the synthesized speech is also evaluated by the listeners using 5-point scale. The 5-point scale for representing the quality of speech and distortion level is given in table 23. Each listener has to give the opinion scores for each of the five utterances in all the four cases (altogether 20 scores) mentioned above. The MOS indicating the synthesized speech quality (perceptual distortion) and the target speaker characteristics are given in table 24. The second and third columns in table 24 shows the MOS for speech quality and target speaker characteristics, respectively.

The MOS indicate that the transformation is effective, if the source and target speakers are from different genders. The basic reason for this variation in MOS with respect to gender is that, the variation in the shapes of the vocal tract and intonation patterns may be large in the case of source and target speakers belong to different genders. As the listener is exposed to source and target speakers voices (original) as well as transformed voice with respect to target speaker, he or she may observe wide transformation in the case of male to female or female to male voice conversions compared to the other cases (male to male or female to female). Hence their feeling is reflected in the judgement.

The performance of the proposed voice conversion system is compared with our earlier voice conversion system. The earlier two versions of the voice conversion systems were developed by us. In the first version, the shape of the vocal tract is characterized by the formants and their bandwidths (Rao & Yegnanarayana 2006b). For transforming the vocal tract characteristics, the gross relationship between the formant frequencies and average pitch is used. Intonation characteristics (pitch contour) are mapped by using fixed-scale factor, which is derived from the average pitch of the source and target speakers. In the second version of the voice conversion system, the vocal tract characteristics are represented by LPCs, and the FFNN is used as the

Table 24. Mean opinion scores for the quality of synthesized speech and target speaker characteristics for different transformations.

Transformation	MOS				Level of confidence (%)	
	Proposed voice conversion system		Earlier voice conversion system (Rao & Yegnanarayana 2006b)			
	Speech quality & distortion	Target speaker characteristics	Speech quality & distortion	Target speaker characteristics	Speech quality & distortion	Target speaker characteristics
M1–F1	4.47	4.28	3.63	3.22	> 99.5	> 99.5
F2–M2	4.56	4.42	3.42	3.37	> 99.5	> 99.5
M1–M2	4.62	3.95	3.79	2.74	> 99.5	> 99.5
F1–F2	4.71	4.03	3.86	3.13	> 99.5	> 99.5

mapping function for deriving the target speaker vocal tract parameters (Rao *et al* 2007). The intonation patterns are represented by the pitch contours at the syllable level. The pitch contour of the utterance is obtained by concatenating the segments of pitch contours at syllable level. The pitch contours at the syllable level are mapped using code books. The MOS of the second version of our earlier voice conversion system are given in 4th and 5th columns of table 24. The significance of the differences in the MOS pairs is tested using hypothesis testing (Hogg & Ledolter 1987). The level of confidence for the observed differences in the sample means was obtained from each model using the sample variances and values of Student-t distribution. The level of confidence is high (> 99.5%) in all cases. This indicates that the differences in the MOS pairs for each case is significant.

The performance of the proposed voice conversion system is also evaluated using ABX tests (Ye & Young 2004; Sundermann *et al* 2005a,b; Blin *et al* 2008). The ABX test is the specific evaluation test for analysing the voice conversion system. Here, the listeners have to mark the test speech utterance (synthesized target speaker speech utterance) as one of the three categories: close to source speaker (A), close to target speaker (B) and different from source and target speakers (X). The result of this test indicates the gross speaker characteristics, whether the voice transformation is effective (in the case of high score for B) or not (in the case of low score for B). These tests conducted by using same set of sentences with same group of students are used in the above MOS studies. The time gap between these two tests (MOS and ABX) is about 15 days. Each listener has to mark 20 sentences as either A, B or X. The scores of the ABX test conducted on the proposed and the earlier voice conversion systems are given in table 25. The first row indicates the scores of the ABX test for the transformed speech utterances derived using the proposed approach. The second row shows the scores of the ABX test for the speech utterances derived from the earlier approach. The results show that about 96% of the speech utterances are marked as B (close to target speaker), while playing the speech utterances transformed using the proposed method. The results of the ABX test on the earlier voice conversion approach shows that about 82% speech utterances are marked as B.

The performance of the MOS and ABX tests indicates that the proposed voice conversion system outperforms the earlier system. This is clearly evident from the level of confidence measures derived using MOS of the two systems (columns 6 and 7 of table 24). The performance of the ABX test also indicates the superiority of the proposed system over the earlier system by showing the lead of 14% in marking B (column 3 of table 25).

For mapping the speaker-specific features between source and target speakers, various models have been explored in the literature. These models are specific to the kind of features used for mapping. For instance, Gaussian mixture model (GMM), neural network models and vector quantization (VQ) are widely used for mapping the vocal tract characteristics (Lee 2007; Toda *et al* 2001; Abe *et al* 1998; Desai *et al* 2010; Rao 2009; Kain & Macon 2001; Toth & Black 2008; Watanabe *et al* 2002; Valbret *et al* 1992; Stylianou *et al* 1995; Narendranadh *et al* 1995). Scatter plots, GMMs and linear models are used for mapping the prosodic features (Inanoglu 2003; Stylianou *et al* 1998).

Table 25. The scores of the ABX test for assessing the characteristics of target speaker from the synthesized speech. The numbers in the braces indicate the percentage of performance.

	Close to source source (A)	Close to target speaker (B)	Different from source and target speakers (X)
Proposed voice conversion approach	5 (1.25)	385 (96.25)	10 (2.5)
Earlier voice conversion approach	26 (6.5)	328 (82)	46 (11.5)

6. Emotion recognition from speech

Human beings exploit emotions extensively to convey the message and its intention. In general, the message will be conveyed in different forms, such as speech, facial expressions, gestures and text. Among these modalities, facial expressions and speech are known to be more effective, for expressing the emotions. From the literature, it is observed that, human perception of the emotions is about 55% from facial expressions, 38% from the speech and 7% from the text (Busso *et al* 2004). There are numerous applications based on emotion recognition (ER). It provides a natural interface between the humans and machines. If the machines could recognize the emotional inputs either through speech or through facial expressions, they could provide appropriate help to users, in particular to handicapped people. Automatic emotion recognition and analysis will be used in call centres to handle the customer queries either by machine or by human operator based on the customer's mood (Yoon & Kyu-SikPark 2007; Lee & Narayanan 2005b). In the case of story telling and e-tutoring applications, the system should automatically analyse the behavioural characteristics of the students' based on their emotions and respond to them accordingly with desired emotions (Schroder *et al* 2001; Pantic & Bartlett 2007; Schrder 2001). An automatic way to analyse the emotions in multimedia documents will be useful for indexing and retrieving the multimedia files based on emotions. This is highly useful in organizing the movies and documentaries based on the emotional contents (Ververidis & Kotropoulos 2006a; Gunes & Pantic 2010). Medical doctors may use the emotional contents of the patient's speech as a diagnosing tool for various ailments (Douglas-Cowie *et al* 2003). Emotion analysis of tapped telephonic conversation, and the surveillance video tapes of criminals or terrorists would help crime investigation departments to predict the activities of the extremists. Conversation with robotic pets and humanoid partners would be more realistic and enjoyable, if they are able to express and understand emotions similar to humans (Douglas-Cowie *et al* 2001; Turk & Schroder 2010; Ververidis & Kotropoulos 2006b). Other similar applications are discussed in the literature (Schroder *et al* 2001; Schrder 2001; Ververidis & Kotropoulos 2006b; Douglas-Cowie *et al* 2001; Schroder & Cowie 2006).

6.1 Database

For developing the ER system, video data was collected using 20 subjects (10 male and 10 female), recorded in a studio environment. In each session, about 3–4 minutes of video data was collected for each emotion from each of the subjects. The total video data was collected in three sessions, and the sessions are separated by 2-week interval. For expressing the emotions, the subjects were asked to speak the emotion-specific text in front of the camera. All the subjects have chosen Hindi text for expressing their emotions. Video data are recorded at 30 frames per second, with a resolution of 640×480 . The distance between the video camera and the subject, lighting conditions and the background are maintained constant for all the recordings. All the subjects who participated in the recording were cooperative. We have used only audio information (speech) for characterizing and recognizing the emotions.

6.2 Extraction of acoustic features

Emotion-specific information is present in speech at different levels. At the segmental level, the emotion specific information can be observed in the form of unique sequence of the shapes of the vocal tract for producing the sound units. The shape of the vocal tract is characterized by the

spectral envelope. Spectral envelope is represented by MFCC (Davis & Mermelstein 1980). At the suprasegmental level, the emotion-specific knowledge is embedded in the duration patterns of the syllable sequences and the dynamics of the pitch and energy contours. At the subsegmental level, the emotion specific information may be present in the shape of the glottal pulse and durations of open and closed phases of the vocal folds.

We have explored segmental and suprasegmental features for the identification of emotions from speech. Usually, segmental features are extracted by analysing the speech segments of duration 20–30 ms. Mostly, these features are extracted from the frequency spectrum of the speech segment, hence these features are known as spectral features. Suprasegmental features, also known as prosodic features, are extracted from the speech segments of duration greater than 100 ms. Subsegmental features are extracted from the speech segments of duration less than 3 ms.

Figure 14 exemplifies the discriminative capability of spectral features for the emotions considered in this study. The plot shows the LP spectrum of vowel /aa/ from the utterance '*maataa aur pitaa kaa aadar karnaa chaahiee*' spoken in five different emotions. From the figure, it is observed that the spectral envelope of /aa/ for each of the emotions is distinct. The formant frequencies, their bandwidths and strengths and spectral roll-off are observed to be different for different emotions. In particular, higher formants and their strengths and bandwidths are highly discriminative.

Figure 15 shows three subplots indicating the duration patterns of the sequence of syllables, energy contours and pitch contours of an utterance '*maataa aur pitaa kaa aadar karnaa chaahie*' in five different emotions. The subplot indicating the duration patterns shows that for some emotions such as fear and happiness, the duration of the initial syllable of the utterance is longer,

Figure 14. LP spectrum of vowel /aa/ in five different emotions.

Figure 15. (a) Duration patterns for the sequence of syllables, (b) Energy contours and (c) Pitch contours in different emotions for the utterance '*maataa aur pitaa kaa aadar karnaa chaahiee*'.

for happy and neutral emotions middle syllables of the utterance seem to be longer, and the final syllables of the utterance seem to be longer for fear and anger (see figure 15a). From the energy plots, it is observed that the utterance with anger emotion has highest energy for the entire duration. Next to the anger emotion, fear and happy shows somewhat more energy than the other two emotions. The dynamics of energy contours can be used to discriminate fear and happiness (see figure 15b). Figure 15c shows that anger, happy and neutral have some what higher pitch values, compared to the other two emotions. Using the dynamics (changes with respect to time) of pitch contours, one can easily discriminate anger, happy and neutral emotions, even though they have higher pitch values on average. Thus figures 14 and 15 provide the basic motivation to pursue the spectral and prosodic features for discriminating the emotions.

MFCC are extracted from a speech frame of 20 ms, with a frame shift of 10 ms. MFCCs are estimated based on human perception of critical bandwidths. The mel-frequency scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. For calculation of MFCCs, the short-term Fourier transform (STFT) analysis is performed on a speech signal using a frame size of 20 ms with a shift of 10 ms. For each frame, an STFT magnitude spectrum is computed, and is further processed by 24 triangular mel-filter banks to find out the filter bank energies. Then discrete cosine transform is taken on the spectral energies to obtain MFCCs. We have used a 13-dimensional MFCC feature vector for representing the log-spectral information.

For deriving the prosodic feature vectors, speech data are segmented into phrases using the knowledge of longer pauses. The average phrase duration is observed to be about 2.5 s, maximum and minimum phrase durations are observed to be 4.2 and 0.9 s respectively. Maximum

number of syllables in a phrase is found to be 23. Therefore, the size of the duration vector is fixed to 23 dimensions indicating 23 duration values. If the number of syllables in a phrase is less than 23, then the tail portion of the duration vector is appended with zeros to maintain the size of the duration vector to be 23. Syllable durations are determined automatically, using vowel onset points (Prasanna & Zachariah 2002; Prasanna 2004). The sequence of fundamental frequency values constitutes pitch contour. Pitch contours are extracted from speech using the autocorrelation of the Hilbert envelope of the linear prediction residual signal (Prasanna & Yegnanarayana 2004a). Energy contour of a speech signal is derived from the sequence of frame energies. Frame energies are computed by summing the squared sample amplitudes. We have chosen the frame size of 20 ms and a frame shift of 10 ms for extracting the pitch and energy contours. The size of pitch and energy contours of the phrases are proportional to the length of the utterance. To obtain the fixed dimensional vector, we have used the resampling technique. The dimension of pitch and energy contours is chosen to be 25. Here, the dimension 25 for pitch and energy contours is not crucial. The reduced size of pitch and energy contours has to be chosen such that the dynamics of the original contours are retained in the resampled versions. The basic reasons for reducing the dimensionality of the original pitch and energy contours are (i) Need for the fixed dimensional input feature vector for developing the AANN models and (ii) The number of feature vectors required for the training is proportional to the size of the feature vector. For capturing the gross behaviour of prosody over a utterance, gross statistical parameters are derived from the frame-level prosodic parameters of the utterance. The pitch and energy parameters at the gross (global) level are represented by maximum, minimum, mean and standard deviation of the frame-level values of the utterance. The duration parameters at the gross level are number of syllables, maximum, minimum and mean durations of the syllables present in the utterance. A 12-dimensional feature vector is used to represent the prosodic information at the gross level.

6.3 *Emotion models using acoustic features*

The audio track is separated from the video data using Adobe Audition software. The spectral and prosodic features are extracted from the speech signal as described in section 6.2. Spectral features are represented by 13-dimensional MFCC feature vectors extracted from the speech frames of 20 ms with a frame shift of 10 ms. Prosodic parameters at the gross (global) level are represented by a 12-dimensional feature vector extracted from the speech utterance. Prosodic features at the fine level are represented by (i) durations of the sequence of syllables present in the utterance, (ii) sequence of fundamental frequency (pitch) values and (iii) sequence of frame energies. The fine level prosodic parameters indicate the dynamics or the time-varying nature of prosodic information.

For each emotion, five AANN models are developed using (i) spectral features (MFCCs), (ii) gross prosodic features, (iii) durations of the sequence of syllables (duration contour), (iv) sequence of pitch values (pitch contour) and (v) sequence of frame energies (energy contour). Models 3–5 represent the emotion models corresponding to dynamic prosodic information. For training the models, features extracted from the speech recorded during sessions 1 and 2 is used. The number of MFCC feature vectors used for training the models is about 5,00,000 (20 speakers $\times$ 2 sessions $\times$ 3 min $\times$ 60 sec $\times$ 70 frames). Spectral feature vectors are extracted from speech frames for every 10-ms shift. On an average, about 70% of the frames are voiced, hence the number of spectral vectors derived from 1 s of speech will be about 70. Prosodic features are extracted at the phrase level (i.e., from each speech phrase, one feature vector will be derived). The terms speech phrase and speech utterance are used interchangeably. The number of prosodic

feature vectors used for training the models is about 2880 (20 speakers $\times$ 2 sessions $\times$ 3 min $\times$ 24 phrases). The structure of the AANN models corresponding to spectral, gross prosodic, duration pattern, pitch contour and energy contour features are 13L 20N 7N 20N 13L, 12L 20N 6N 20N 12L, 23L 40N 10N 40N 23L, 25L 40N 15N 40N 25L and 25L 40N 15N 40N 25L, respectively. It is found that 75 epochs are adequate for training the AANN models using spectral and gross prosodic features, and about 100 epochs are found to be sufficient using dynamic prosodic features.

6.4 Evaluation of the emotion models

The developed emotion-specific models are evaluated using the video data collected in the third session. We have five ER systems developed using acoustic features (namely, spectral (MFCC), gross prosodic, duration pattern of the sequence of syllables, pitch contour and energy contour). Each ER system consists of five AANN models representing the five emotions: Anger (A), Fear (F), Happy (H), Neutral (N) and Sad (S). The block diagram of the emotion recognition system using AANN models is shown in figure 16. For evaluating the performance of the ER system, the feature vectors derived from the test video clips are given as input to five AANN models. The output of the each model is compared with the input to compute the normalized squared error. The normalized squared error (e) for the feature vector y is given by, $e = \frac{\|y-o\|^2}{\|y\|^2}$, where o is the output vector given by the model. The error e is transformed into a confidence score (c) using $c = \exp(-e)$. The average confidence score is calculated for each model. The category of the emotion is decided based on the highest confidence score.

First we have analysed the performance of the five ER systems separately, and then they are combined in three phases. Phase 1: combining the developed ER systems using individual dynamic prosodic features. Phase 2: combining the developed ER system using gross prosodic features with combined ER system developed using dynamic prosodic features in Phase 1. Phase 3: combining the ER system developed using spectral features with the combined ER system obtained from Phase 2. All the three phases of combining the individual ER systems are shown in figure 17.

Figure 16. Emotion recognition system (ERS) using AANN models.

Figure 17. Illustration of overall ER system (ERS) by combining the evidences at different phases.

ER systems of different modalities are combined by summing the weighted confidence scores (evidences). The weighting rule for combining the confidence scores of individual modalities is as follows: $c^m = \frac{1}{m} \sum_{i=1}^m w_i c_i$, where c^m is the multimodal confidence score, w_i and c_i are weighting factor and confidence score of the i^{th} modality, and m indicates number of modalities used for combining the scores. We have combined three modalities in Phase 1, and two modalities in phases 2 and 3. One of the weights (w_i) is varied in steps of 0.1 from 0 to 1, and the other weights are determined using the formula: $w_j = \frac{1-w_i}{m-1}$, where $j = 1$ to m and $j \neq i$, $i = 1$ to m .

Performance of the ER systems developed using dynamic prosodic features extracted from the speech utterances is given in table 26. Dynamic prosodic features are represented by (i) duration pattern of the sequence of syllables (duration contour), (ii) sequence of pitch values (pitch contour) and (iii) sequence of frame energies (energy contour). Columns 2–6, 7–11 and 12–16 show the performance of the ER systems developed by the features derived from duration contour, pitch contour and energy contour, respectively. The recognition performance shows that among the five emotions Anger and Fear form one group and Happy, Neutral and Sad form the other group. The average recognition performance of ER systems developed using the features extracted from duration, pitch and energy contours are observed to be 69%, 65% and 52%, respectively.

Table 27 shows the recognition performance by combining the confidence scores of the three ER systems developed using duration, pitch and energy contours. Weighting factor associated to each system is varied from 0 to 1, with the steps of 0.1. The recognition performance of the combined system for various combinations of the weighting factors is shown in figure 18. It

Table 26. Performance of the ER systems developed using (i) duration contour, (ii) pitch contour and (iii) energy contour. The entries in the table indicate the percentage of recognition. (A: Anger, F: Fear, H: Happy, N: Neutral and S: Sad).

	Duration contour					Pitch contour					Energy contour				
	A	F	H	N	S	A	F	H	N	S	A	F	H	N	S
A	66	16	7	9	2	68	8	14	6	4	45	12	23	7	3
F	13	70	8	5	4	16	61	11	7	5	18	55	12	9	6
H	8	6	62	14	10	9	8	58	16	9	14	4	50	19	9
N	4	2	8	76	10	3	5	9	73	10	6	4	12	58	20
S	7	2	9	10	72	1	2	9	21	67	5	3	16	22	54

Table 27. Performance of the combined ER system by combining the evidences of individual ER systems developed using the features of duration, pitch and energy contours. The weighting factors correspond to the following performance are 0.5, 0.25 and 0.25 for the ER systems based on duration, pitch and energy contour features, respectively.

	Emotion recognition performance (%)				
	Anger	Fear	Happy	Neutral	Sad
Anger	77	6	12	3	2
Fear	21	66	4	6	3
Happy	13	8	64	12	3
Neutral	3	1	5	82	9
Sad	1	2	6	12	79

is observed that the best recognition performance is about 74% for the weighting factors 0.5, 0.25 and 0.25 for the confidence scores of duration, pitch and energy contour-based ER systems, respectively.

Performance of the ER systems developed using spectral and gross prosodic features is given in tables 28 and 29, respectively. The average recognition performance using spectral features is observed to be 74%, whereas with gross prosodic features, it is observed to be about 68%.

The recognition performance of the ER systems developed using gross and dynamic prosodic features is 68% and 74%, respectively. These two systems acquire the emotion-specific knowledge from complementary features. One of the systems used the gross statistics of the prosodic parameters, and the other one used local as well as sequential knowledge of prosodic parameters. Therefore, combining the evidences of these two systems may improve the recognition performance. The evidences of gross and dynamic prosodic feature-based ER systems are combined using the weighting rule. The weighting factor of the ER system based on static prosodic features

Figure 18. Recognition performance for different combinations of weighting factors associated to ER systems based on dynamic prosodic features.

Table 28. Performance of the emotion recognition system developed using spectral features.

	Emotion recognition performance (%)				
	Anger	Fear	Happy	Neutral	Sad
Anger	72	8	4	9	7
Fear	5	77	8	10	0
Happy	6	7	69	9	9
Neutral	5	0	11	76	8
Sad	0	8	8	10	74

is varied from 0 to 1, in steps of 0.1. With this, we obtain a total of 11 combinations of weighting factors. The recognition performance of the combined system for various combinations of the weighting factors is shown in figure 19. It is observed that the best recognition performance is about 79% for the weighting factors 0.3 and 0.7 for the confidence scores of gross and dynamic prosodic feature-based ER systems, respectively. Table 30 shows the confusion matrix for the ER system developed by combining the evidences of gross and dynamic prosodic feature-based ER systems.

The evidences of ER systems based on spectral and prosodic features can be viewed as complementary. This is because, spectral features (i.e., MFCCs) are extracted by processing speech frames of size 20 ms, whereas prosodic features are extracted from the whole speech utterance. Another basic difference between these two features is that spectral features characterize the emotions in the form of distinct shapes of the vocal tract, whereas prosodic features characterize the emotions in terms of distinct duration, intonation and energy patterns. Hence, combining the evidences of these two systems may enhance the recognition performance further. Evidences of the spectral and prosodic systems are combined using weighting rule. Recognition performance for different weighting factors is shown in figure 20. The best performance is observed to be about 85% for the weighting factors of 0.4 and 0.6 to spectral and prosodic feature-based ER systems, respectively. Table 31 shows the recognition performance of the ER system developed by combining the evidences of spectral and prosodic feature-based ER systems.

Several approaches to recognize emotions from speech have been reported (Wua *et al* 2010; Busso *et al* 2009; Giannakopoulos *et al* 2009; Grimm *et al* 2007; Lee *et al* 2007; Lee & Narayanan 2005a; Lugger & Yang 2008; Sun *et al* 2009). A comprehensive review of these approaches can be found in Douglas-Cowie *et al* (2001), Ververidis & Kotropoulos (2006b),

Table 29. Performance of the ER system developed using gross prosodic features.

	Emotion recognition performance (%)				
	Anger	Fear	Happy	Neutral	Sad
Anger	72	8	9	7	4
Fear	20	65	7	8	0
Happy	17	6	65	7	5
Neutral	4	2	8	74	12
Sad	7	0	9	18	66

Figure 19. Recognition performance for different combinations of weighting factors associated to ER systems based on gross and dynamic prosodic features.

Pantic & Rothkrantz (2003), Schroder (1996) and Rao *et al* (2010). Most researchers have used global suprasegmental/prosodic features as their acoustic cues for emotion recognition. Pitch-related features are used in Dellaert *et al* (1996) and Busso *et al* (2009) for classifying four emotions. Peaks and troughs in the profile of fundamental frequency and intensity, durations of pauses are explored in McGilloway *et al* (2000), for identifying the four emotions: fear, anger, sadness and joy. Prosodic and phonetic features are used in Nicholson *et al* (1999) for recognizing eight emotions. Prosodic features used are power and intonation patterns, whereas phonetic features adopted were linear prediction coefficients and their delta features. Recognition performance of 50% was achieved using neural networks (Nicholson *et al* 1999). Short-time suprasegmental features and their statistics are used in Ververidis & Kotropoulos (2006b) for discriminating the emotions. Log frequency power coefficients (LFPC) are used for classifying

Table 30. Performance of the combined ER system by combining the evidences of individual ER systems developed using dynamic and gross prosodic features. The weighting factors corresponding to the following performance are 0.7 and 0.3 for the ER systems based on dynamic and gross prosodic features, respectively.

	Emotion recognition performance (%)				
	Anger	Fear	Happy	Neutral	Sad
Anger	80	3	11	6	2
Fear	10	73	8	4	5
Happy	18	5	74	3	0
Neutral	2	0	4	86	8
Sad	3	2	4	10	81

Figure 20. Recognition performance for different combinations of weighting factors associated with the ER systems based on spectral and overall prosodic features.

the emotions using discrete hidden Markov model (Nwe *et al* 2003). In addition to pitch-related information, log energy, formants and MFCC are used for classifying the emotions (Kwon *et al* 2003). They have achieved 96.3% success in classifying the stressed and neutral speech, and 70% for classifying the four emotions (Kwon *et al* 2003). Fifty-five acoustic features comprising 25-prosodic, 24-MFCCs and 6-formant frequencies are used for discriminating six emotions using Fisher's linear discriminate analysis (Wang & Guan 2004). Features extracted from glottal air flow signal are used in Iliev *et al* (2010) for classifying the emotions using optimal path classifier. Modulation spectral features are proposed in Wua *et al* (2010) for discriminating the emotions along with prosodic features. Articulatory features are explored in Lee & Narayanan (2005a) for recognizing emotions from speech.

Table 31. Performance of the combined emotion recognition system by combining the evidences of individual ER systems developed using spectral and prosodic features. The weighting factors corresponding to the following performance are 0.4 and 0.6 for the ER systems based on spectral and prosodic features, respectively.

	Emotion recognition performance (%)				
	Anger	Fear	Happy	Neutral	Sad
Anger	87	2	3	3	5
Fear	4	82	5	6	3
Happy	1	6	81	7	5
Neutral	0	4	3	90	3
Sad	1	1	5	7	86

7. Dialect identification from speech

We have explored speech features to identify Hindi dialects. A dialect is any distinguishable variety of a language spoken by a group of people. In this study, five prominent dialects of Hindi are considered for the identification task. They are Chattisgarhi (spoken in central India), Bengali (Bengali-accented Hindi spoken in the eastern region), Marathi (Marathi-accented Hindi spoken in the western region), General (Hindi spoken in the northern region) and Telugu (Telugu-accented Hindi spoken in the southern region). Speech database considered for this study consists of spontaneous speech spoken by male and female speakers. Prosodic and spectral features extracted from speech are used for discriminating the dialects. Spectral features are represented by MFCC and prosodic features are represented by durations of syllables, pitch and energy contours. AANN models are used to capture the dialect-specific information from the distributions of the feature vectors.

Dialects of a given language are the differences in speaking styles of a particular language, because of geographical and ethnic differences of the speakers. A number of studies have shown that the acoustic space spanned by phonemes for native speakers will shift when speakers are non-native. Other factors such as voice onset time, voiced stop release time, durations of the sound units and pitch contours are also important in identifying the dialect (Arslan & Hansen 1996, 1997). Recent studies have considered the features extracted from spectral trajectories for a dialect classification (Angkitrakul & Hansen 2002, 2003).

Automatic dialect classification has several applications. For increasing the performance of the speech systems (such as speech recognition and speaker recognition), dialect identification at the front-end will narrow down the search space and improve the performance further. For the natural human machine interface, dialect identification system will help the machine in understanding the speech spoken by the human and to synthesize the speech in the appropriate dialect of the person communicating with the machine (Blackburn *et al* 1993; Arslan & Hansen 1997).

We have explored segmental and suprasegmental features for the identification of dialects of Hindi language. Usually, segmental features are extracted by analysing the speech segments of duration 20–30 ms. Mostly, these features are extracted from the frequency spectrum of the speech segment, hence these features are known as spectral features. Suprasegmental features also known as prosodic features extracted from the speech segments of duration greater than 100 ms. Subsegmental features are extracted from the speech segments of duration less than 3 ms.

Automatic dialect identification studies were carried out for the languages of western and eastern countries such as USA and Japan (Arslan & Hansen 1997; Itahashi & Tanaka 1993). In literature, most of the dialect identification systems are developed using Gaussian mixture models and support vector machines (Torres-Carrasquillo *et al* 2004, 2008; Zissman *et al* 1996). GMM techniques yield an average of 30% equal error rate for the dialects in the Miami corpus and about 13% equal error rate for the dialects in the CallFriend corpus (Torres-Carrasquillo *et al* 2004). No systematic study has been carried out on the dialects of any Indian language using the features derived from speech. Hence, we are exploring the features derived from speech for identifying the dialects of Hindi (Rao *et al* 2010). At present, hundreds of dialects of Hindi are being used in different geographical regions of India. We have considered five prominent dialects of Hindi spoken in central, eastern, western, northern and southern regions of India. We named these five Hindi dialects using their local language names. Chattisgarhi is the local language in the central part of India. Hence, the dialect of Hindi spoken in that region is named as Chattisgarhi. Similarly, the dialects of Hindi spoken in eastern, western and southern regions of India are named as Bengali, Marathi and Telugu. The local language of northern

region of India is mainly Hindi, hence the northern dialect is named as General. AANN are explored for capturing the dialect specific information from the proposed spectral and prosodic features.

7.1 Database

Speech data are collected from five different geographical regions (central, eastern, western, northern and southern) of India, representing five dialects of Hindi. For each dialect, speech data are collected from five male and five female speakers. Speech data are collected from a speaker, by posing arbitrary questions on topics such as the speaker's childhood, history of the hometown, details of the career, habits and so on. From each speaker, 5–10 min of speech are collected from the spontaneous responses to the above questions. Altogether, for each dialect, the duration of total speech materials is about 1–1.5 h. Instead of reading some study material or uttering small fixed text sentences, responses to the general questions usually contain the natural accent of the language. Hence, we have used the spontaneous responses to the questions as the speech material for the identification of dialects.

7.2 Development of dialect identification systems

For each dialect, four AANN models are developed using spectral, duration, pitch and energy features. MFCCs are used for representing the spectral features (Benesty *et al* 2008; Rabiner & Juang 1993). MFCCs are extracted from a speech frame of 20 ms, with a frame shift of 10 ms. We have used a 13-dimensional MFCC feature vector to represent the speech frame. The derived MFCC feature vectors of a particular dialect are given as input and output of the AANN model. The reason for feeding the feature vectors as input and output is to capture the distribution of the feature vectors. The number of epochs needed for training depends on the behaviour of the training error. It is found that 100 epochs are adequate for the AANN models used in this work. For developing the AANN models using prosodic parameters, the sizes of all feature vectors should be the same. For deriving the prosodic feature vectors, speech data are segmented into phrases at pauses longer than 250 ms. The average phrase duration is observed to be about 2.5 s, maximum and minimum phrase durations are observed to be 4.2 and 0.9 s, respectively. Maximum number of syllables in a phrase is found to be 23. Therefore, the size of the duration vector is fixed to 23 dimensions indicating 23 duration values. If the number of syllables in a phrase is less than 23, then the tail portion of the duration vector is appended with zeros to maintain the size of the duration vector to be 23. Syllable durations are determined automatically, using vowel onset points (Prasanna & Zachariah 2002; Prasanna 2004). The sequence of fundamental frequency values constitutes the pitch contour. Pitch contours are extracted from speech using the autocorrelation of the Hilbert envelope of the linear prediction residual signal (Prasanna & Yegnanarayana 2004a). The energy contour of a speech signal is derived from the sequence of frame energies. Frame energies are computed by summing the squared sample amplitudes. We have chosen the frame size of 20 ms and a frame shift of 10 ms. The size of pitch and energy contours of the phrases are proportional to the length of the utterance. To obtain a fixed dimensional vector, we have used resampling. The dimension of pitch and energy contours is chosen to be 23. Here, the dimension of 23 for pitch and energy contours is not crucial. The reduced size of pitch and energy contours has to be chosen such that the dynamics of the original contours have to be retained in the resampled versions. The structures of the AANN models used in this study are 13L 26N 6N 26N 13L and 23L 40N 10N 40N 23L for capturing the distributions of

Figure 21. Dialect identification system (DIS) using AANN models.

spectral and prosodic feature vectors, respectively. For training the AANN models, speech data of 7 (3 female and 4 male) speakers is used. Speech data from the other 3 (2 female and 1 male) speakers is used for evaluating the models.

7.3 Evaluation of dialect identification systems

We have developed four dialect identification (DI) systems using (i) spectral features, (ii) durations of syllables in the utterance, (iii) pitch contour (sequence of pitch (F_0) values) and (iv) energy contour. Each DI system consists of five AANN models representing the five dialects: Chattisgarhi (C), Bengali (B), Marathi (M), General (G) and Telugu (T). The block diagram of the basic DI system using AANN models is shown in figure 21. We first analysed the performance of the four DI systems separately, and then they are combined using score-level fusion.

Table 32. Performance of the dialect identification system developed using spectral features. The entries in the table indicate the percentage of recognition. (C: Chattisgarhi, B: Bengali, M: Marathi, G: General and T: Telugu).

	Identification performance (%)				
	C	B	M	G	T
C	58	8	4	19	11
B	6	68	4	13	9
M	12	6	52	17	13
G	6	10	6	70	8
T	8	6	10	15	61

Table 33. Performance of the dialect identification systems developed using duration, pitch and energy features. The entries in the table indicate the percentage of recognition. (C: Chattisgarhi, B: Bengali, M: Marathi, G: General and T: Telugu).

	Duration (D)					Pitch (P)					Energy (E)					Combined (D+P+E)				
	C	B	M	G	T	C	B	M	G	T	C	B	M	G	T	C	B	M	G	T
C	52	10	18	12	8	63	7	9	11	10	43	11	9	17	20	69	10	5	7	9
B	9	62	7	10	12	10	59	8	12	11	12	50	9	15	14	6	69	6	9	10
M	10	8	54	12	16	10	8	62	7	13	8	17	42	10	23	6	10	66	10	8
G	4	10	8	60	18	6	8	6	73	7	9	11	10	54	16	7	4	6	79	4
T	8	12	10	17	53	8	15	9	19	49	14	16	8	12	50	7	10	8	12	63

Performance of the DI system using spectral features is given in table 32. The average identification performance is observed to be 62%. The diagonal entries of the table indicates the correct classification, and the rest indicates the misclassification. Performance of the DI system using prosodic features is given in table 33. Columns 2–6, 7–11 and 12–16 show the performance of the DI systems developed by duration, pitch and energy features, respectively. The average performance is observed to be 55%, 61% and 48% for the DI systems developed using duration, pitch and energy features, respectively. Columns 17–21 show the performance of the DI system by combining the confidence scores of the individual prosodic systems. The average performance of the combined prosodic system is observed to be much better (69%) compared to the individual prosodic systems. Finally, the evidence of the spectral-based DI system is combined with the evidence of the combined prosodic system. The block diagram of the combined DI system (spectral+prosodic) is shown in figure 22. The results of this combination show drastic improvement with respect to the performances of individual systems. The average performance of the combined system (i.e., spectral + prosodic) is found to be 78% (see table 34). The reason for the improved performance of the combined system may be due to the complementary nature of the features.

Figure 22. Combined dialect identification system using the evidences of spectral and prosodic DI systems.

Table 34. Performance of the dialect identification system by combining the evidences from the DI systems developed using spectral and prosodic features. The entries in the table indicate the percentage of recognition. (C: Chattisgharhi, B: Bengali, M: Marathi, G: General and T: Telugu).

	Identification performance (%)				
	C	B	M	G	T
C	80	6	4	7	3
B	5	77	8	4	6
M	6	9	74	5	6
G	2	2	4	86	6
T	4	9	6	8	73

8. Summary and conclusions

We have discussed the applications of neural networks for developing various speech systems. Among several neural network models, application of different configurations of feedforward neural network models were explored. Four-layer FFNN models were used for predicting durations and F_0 values of the syllables using the linguistic and production constraints. The neural network-based prosody models were further demonstrated to capture the prosodic information specific to speaker, language and sound unit categories. Therefore, the developed prosodic models can also be explored along with the conventional models for enhancing the performance of speech, speaker and language recognition systems. The developed prosodic models were also demonstrated to be useful in text to speech synthesis systems, by improving the quality of the synthesized speech. Five-layer FFNN models with symmetric structure were explored for capturing the mapping functions between source and target speakers in the context of a voice conversion application. Separate mappings functions were developed for mapping vocal tract, excitation source and prosodic features. The developed mapping functions were evaluated using both subjective and objective measures. Autoassociative neural network models were explored for capturing the emotion specific information from speech using spectral and prosodic features. Finally, AANN models were also used for discriminating five Hindi dialects using spectral and prosodic features.

References

- Abe M, Nakanura S, Shikano K, Kuwabara H 1998 Voice conversion through vector quantization. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 1: 655–658
- Angkititrakul P, Hansen J L 2002 Stochastic trajectory model analysis for accent classification. *ICSLP*, Denver, CO, USA, pp. 493–496
- Angkititrakul P, Hansen J L 2003 Use of trajectory models for automatic accent classification. *Proc. Eurospeech*, Geneva, Switzerland, pp. 1353–1356
- Anjani A V N S 2000 *Autoassociate neural network models for processing degraded speech*. Master's thesis, MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India
- Arslan L, Hansen J 1996 Language accent classification in American English. *Speech Commun.* 18(4): 353–367

- Arslan L, Hansen J 1997 A study of temporal features and frequency characteristics in American English foreign accent. *J. Acoust. Soc. Am.* 102: 28–40
- Barbosa P A, Bailly G 1992 Generating segmental duration by p-centers. *Proc. of the Fourth Workshop on Rhythm Perception and Production*, Bourges, France, pp. 163–168
- Barbosa P A, Bailly G 1994 Characterization of rhythmic patterns for text-to-speech synthesis. *Speech Commun.* 15: 127–137
- Batliner A, Mobius B, Mohler G, Schweitzer A, Noth E 2001 Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground. *Proc. Eurospeech*, Scandinavia
- Benesty J, Sondhi M M, Huang Y (eds) 2008 *Springer handbook on speech processing*, (Berlin, Germany: Springer Publishers)
- Black A W, Taylor P, Caley R 2000 The festival speech synthesis system: System documentation. The Centre for Speech Technology Research (CSTR), University of Edinburgh, 1.4.0 edition. http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html
- Blackburn C S, Vonwiller J P, King R W 1993 Automatic accent classification using artificial neural networks. In *Proc. Eurospeech*, vol. 2, pp. 1241–1244
- Blin L, Boeffard O, Barraud V 2008 Web-based listening test system for speech synthesis and speech conversion evaluation. In *Proc. LREC*, Marrakech (Morocco)
- Breiman L, Friedman N, Olshen R 1984 *Classification and regression trees* (Pacific Grove, CA, USA: Wadsworth and Brooks)
- Buhmann J, Vereecken H, Fackrell J, Martens J P, Coile B V 2000 Data driven intonation modeling of 6 languages. *Proc. Int. Conf. Spoken Language Processing*, vol. 3, Beijing, China, pp. 179–183
- Busso C, Deng Z, Yildirim S, Bulut M, Lee C M, Kazemzadeh A, Lee S, Neumann U, Narayanan S 2004 Analysis of emotion recognition using facial expressions, speech and multimodal information. In *ACM 6th Int. Conf. on Multimodal Interfaces (ICMI 2004)*, ACM, State College, PA
- Busso C, Lee S, Narayanan S 2009 Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Trans. Speech Audio Process.* 17: 582–596
- Campbell W N 1990 Analog i/o nets for syllable timing. *Speech Commun.* 9(1): 57–61
- Campbell W N 1992 Syllable based segment duration. In (eds) G Bailly, C Benoit, T R Sawallis, *Talking machines: Theories, models and designs*, Elsevier, Amsterdam, pp. 211–224
- Campbell W N 1993 Predicting segmental durations for accommodation within a syllable-level timing framework. In *Proc. European Conf. Speech Communication and Technology*, vol. 2, Berlin, Germany, pp. 1081–1084
- Campbell W N, Isard S D 1991 Segment durations in a syllable frame. *J. Phonetics: Special issue on speech synthesis* 19: 37–47
- Chopde A 2001 Itrans Indian language transliteration package version 5.2 source. <http://www.aczone.com/itrans/>
- Chung H 2002a Duration models and the perceptual evaluation of spoken Korean. In *Proc. Speech Prosody*, Aix-en-Provence, France, pp. 219–222
- Chung H 2002b Perceptual evaluation of duration models in spoken Korean. *Korean J. Speech Sci.* 9: 207–215
- Cordoba R, Vallejo J A, Montero J M, Gutierrezarriola J, Lopez M A, Pardo J M 1999 Automatic modeling of duration in a Spanish text-to-speech system using neural networks. In *Proc. European Conf. Speech Communication and Technology*, Budapest, Hungary
- Cosi P, Tesser F, Gretter R 2001 Festival speaks Italian. In *Proc. Eurospeech 2001*, Aalborg, Denmark, pp. 509–512
- Davis S, Mermelstein P 1980 Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Speech Audio Process.* 28(4): 357–366
- Dellaert F, Polzin T, Waibel A 1996 Recognising emotions in speech. In *International Conference on Spoken Language Processing (ICSLP) 96*, vol. 3, Philadelphia, PA, USA, pp. 1816–1819
- Deller J R, Proakis J G, Hansen J H L 1993 *Discrete-time processing of speech signals* (New York, USA: Macmilan Publishing Company)
- Desai S, Black A W, Yegnanarayana B, Prahlad K 2010 Spectral mapping using artificial neural networks for voice conversion. *IEEE Trans. Audio Speech Lang. Process.* 18(5): 954–964

- Diamantaras K I, Kung S Y 1996 *Principal component neural networks: Theory and applications* (New York: John Wiley and Sons)
- Douglas-Cowie E, Campbell N, Cowie R, Roach P 2003 Emotional speech: Towards a new generation of databases. *Speech Commun.* 40: 3360
- Douglas-Cowie R, Tsapatsoulis E, Votsis N, Kollias G, Fellenz S, Felling W, Taylor J 2001 Emotion recognition in human computer interaction. *IEEE Signal Process. Mag.*, Stockholm, Sweden, 23–25 April 2001
- Dusterhoff K E, Black A W, Taylor P A 1999 Using decision trees within the Tilt intonation model to predict F0 contour. In *Proc. Eurospeech*, Budapest, Hungary
- Fujisaki H 1983 Dynamic characteristics of voice fundamental frequency in speech and singing. In (ed) P F MacNeilage, *The production of speech*, New York, USA: Springer-Verlag, pp. 39–55
- Fujisaki H 1988 A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In (ed) O Fujimura, *Vocal physiology: Voice production, mechanisms and functions*, New York, USA: Raven Press, pp. 347–355
- Gangashetty S V, Sekhar C C, Yegnanarayana B 2004 Extraction of fixed dimension patterns from varying duration segments of consonant-vowel utterances. In *Proc. IEEE Int. Conf. Intelligent Sensing and Information Processing*, Chennai, India, pp. 159–164
- Giannakopoulos T, Pikrakis A, Theodoridis S 2009 A dimensional approach to emotion recognition of speech from movies. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 1: 65–68
- Goubanova O, King S 2008 Bayesian networks for phone duration prediction. *Speech Commun.* 50: 301–311
- Goubanova O, Taylor P 2000 Using bayesian belief networks for modeling duration in text-to-speech systems. In *Proc. Int. Conf. Spoken Language Processing*, vol. 2, Beijing, China, pp. 427–431
- Grimm M, Kroschel K, Narayanan S 2007 Support vector regression for automatic recognition of spontaneous emotions in speech. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 4: 1085–1088
- Gunes H, Pantic M 2010 Automatic, dimensional and continuous emotion recognition. *Int. J. Synthetic Emotions* 1(1): 68–99
- Gupta C S 2003 *Significance of source features for speaker recognition*. Master's thesis, MS thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India
- Haykin S 1999 *Neural networks: A comprehensive foundation* (New Delhi: Pearson Education Aisa, Inc.)
- Hifny Y, Rashwan M 2002 Duration modeling of Arabic text-to-speech synthesis. In *Proc. Int. Conf. Spoken Language Processing*, Denver, Colorado, USA, pp. 1773–1776
- Hogg R V, Ledolter J 1987 *Engineering statistics* (New York: Macmillan Publishing Company)
- Hwang S H, Chen S H 1994 Neural-network-based F0 text-to-speech synthesizer for Mandarin. *IEE Proc. Image Signal Process.* 141(6): 384–390
- Hwang S-H, Chen S-H 1995 A prosodic model for mandarin speech and its application to pitch level generation for text-to-speech. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 1: 616–619
- Ikbāl M S, Misra H, Yegnanarayana B 1999 Analysis of autoassociative mapping neural networks. In *Int. Joint Conf. Neural Networks*, USA, pp. 854–858
- Iliev A I, Scordilis M S, Papa J P, Falcão A X 2010 Spoken emotion recognition through optimum-path forest classification using glottal features. *Comput. Speech. Lang.* 24(3): 445–460
- Inanoglu Z 2003 *Transforming pitch in a voice conversion framework*. Master's thesis, St. Edmunds College, University of Cambridge
- Itahashi S, Tanaka K 1993 A method of classification among Japanese dialects. *Proc. Eurospeech* 1: 639–642
- Itakura F 1975 Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Speech Audio Process.* 23(1): 67–72
- Kain A 2001 *High resolution voice transformation*. PhD thesis, OGI School of Science and Engineering, Oregon Health and Science University, USA
- Kain A, Macon M W 2001 Design and evaluation of a voice conversion algorithm based on spectral envelop mapping and residual prediction. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, Salt Lake City, UT, USA, pp. 813–816

- Khan A N, Gangashetty S V, Yegnanarayana B 2003 Syllabic properties of three Indian languages: Implications for speech recognition and language identification. In *Int. Conf. Natural Language Processing*, Mysore, India, pp. 125–134
- Kishore S P, Kumar R, Sangal R 2002 A data-driven synthesis approach for Indian languages using syllable as basic unit. In *Int. Conf. Natural Language Processing*, Mumbai, India, pp. 311–316
- Kishore S P, Yegnanarayana B 2001 Online text-independent speaker verification system using autoassociative neural network models. In *Int. Joint Conf. Neural Networks*, Washington, DC, USA
- Krishna N S, Murthy H A 2004 Duration modeling of Indian languages Hindi and Telugu. In *5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, pp. 197–202
- Krishna N S, Murthy H A, Gonsalves T A 2002 Text-to-speech (tts) in Indian languages. In *Int. Conf. Natural Language Processing*, Mumbai, pp. 317–326, 18–21 December 2002
- Krishna N, Tulukdar P, Bali K, Ramakrishnan A 2004 Duration modeling for Hindi text-to-speech synthesis system. In *Proc. Int. Conf. Spoken Language Processing*, Denver, USA
- Kwon O, Chan K, Hao J, Lee T 2003 Emotion recognition by speech signals. In *Eurospeech*, Geneva, pp. 125–128
- Lee C M, Narayanan S 2005a Toward detecting emotions in spoken dialogs. *IEEEAUP* 13(2): 293–303
- Lee C M, Narayanan S S 2005b Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process.* 13(2): 293–303
- Lee K 2007 Statistical approach for voice personality transformation. *IEEE Trans. Audio Speech Lang. Process.* 15: 641–651
- Lee S, Yildirim S, Kazemzadeh A, Narayanan S 2007 An articulatory study of emotional speech production. *Proc. Interspeech* 4: 497–500
- Leung C-C, Ferras M, Barras C, Gauvain J-L 2008 Comparing prosodic models for speaker recognition. In *Interspeech*, vol. 1, Brisbane, Australia, pp. 1945–1948, September 2008
- Lugger M, Yang B 2008 Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 4: 4945–4948
- Mary L, Yegnanarayana B 2008 Extraction and representation of prosodic features for language and speaker recognition. *Speech Commun.* 50(10): 782–796
- McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S 2000 Approaching automatic recognition of emotion from voice: A rough benchmark. In *ISCA Workshop on Speech and Emotion*, Belfast
- Mei X, Sun S 2000 An efficient method to compute lsfs from lpc coefficients. In *ICSP-2000*, pp. 655–658
- Mixdorff H, Jokisch O 2001 Building an integrated prosodic model of German. In *Proc. European Conf. Speech Communication and Technology*, vol. 2, Aalborg, Denmark, pp. 947–950
- Murthy P S, Yegnanarayana B 1999 Robustness of group-delay-based method for extraction of significant excitation from speech signals. *IEEE Trans. Speech Audio Process.* 7: 609–619
- Narendranadh M, Murthy H A, Rajendran S, Yegnanarayana B 1995 Transformation of formants for voice conversion using artificial neural networks. *Speech Commun.* 16(2): 206–216
- Nicholson J, Takahashi K, Nakatsu R 1999 Emotion recognition in speech using neural networks. In *6th International Conference on Neural Information Processing, ICONIP-99*, pp. 495–501
- Nwe T L, Foo S W, Silva L C D 2003 Speech emotion recognition using hidden Markov models. *Speech Commun.* 41(4): 603–623
- Oppenheim A V, Schaffer R W, Buck J R 1999 *Discrete-time signal processing* (NJ: Prentice-Hall)
- Ostendorf M, Shafran I, Bates R 2003 Prosody models for conversational speech recognition. In *Symposium on Prosody and Speech*, Tokyo, Japan
- Pantic M, Bartlett M 2007 Machine analysis of facial expressions. In (eds) K Delac, M Grgic, *Face recognition* (Vienna: I-Tech Education) pp. 377–416
- Pantic M, Rothkrantz L J M 2003 Toward an affect-sensitive multimodal human–computer interaction. *Proc. IEEE* 91: 1370–1390
- Percybrooks W S, Moore-II E 2008 Voice conversion with linear prediction residual estimation. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 4: 4673–4676

- Pierrehumbert J B 1980 *The phonology and phonetics of English intonation*. PhD thesis, MIT, MA, USA
- Prasanna S R M 2004 *Event-based analysis of speech*. PhD thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India
- Prasanna S R M, Yegnanarayana B 2004a Extraction of pitch in adverse conditions. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, Montreal, Canada
- Prasanna S R M, Yegnanarayana B 2004b Extraction of pitch in adverse conditions. In *IEEE Int. Conf. Acoust. Speech Audio Process.*, vol. 1, Montreal, Canada
- Prasanna S R M, Zachariah J M 2002 Detection of vowel onset point in speech. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 3, Orlando, Florida, USA
- Rabiner L R, Juang B H 1993 *Fundamentals of speech recognition* (Englewood Cliffs, New Jersey: Prentice-Hall)
- Rao K S 2011 Application of prosody models for developing speech systems in Indian languages. *Int. J. Speech Technol.* 14: 19–33
- Rao K S 2009 Voice conversion by mapping the speaker-specific features using pitch synchronous approach. *Comput. Speech Lang.* 23(2): 240–256
- Rao K S, Laskar R H, Koolagudi S G 2007 Voice transformation by mapping the features at syllable level. In *2nd International Conference on Pattern Recognition and Machine Intelligence (Premi-2007)*, Kolkata, India
- Rao K S, Nandy S, Koolagudi S G 2010 Identification of Hindi dialects using speech. In *WMSCI-2010*, Orlando, Florida, USA
- Rao K S, Reddy R, Maity S, Koolagudi S G 2010 Characterization of emotions using dynamics of prosodic features. In *Speech Prosody*, Chicago, USA
- Rao K S, Yegnanarayana B 2003 Prosodic manipulation using instants of significant excitation. In *Proc. IEEE Int. Conf. Multimedia and Expo*, Baltimore, Maryland, USA, pp. 389–392
- Rao K S, Yegnanarayana B 2006a Prosody modification using instants of significant excitation. *IEEE Trans. Audio Speech Lang. Process.* 14(3): 972–980
- Rao K S, Yegnanarayana B 2006b Voice conversion by prosody and vocal tract modification. In *9th Int. Conf. Information Technology*, Bhubaneswar, Orissa, India
- Rao K S, Yegnanarayana B 2007 Modeling durations of syllables using neural networks. *Comput. Speech Lang.* 21: 282–295
- Rao K S, Yegnanarayana B 2009 Intonation modeling for Indian languages. *Comput. Speech Lang.* 23: 240–256
- Riley M 1992 Tree-based modeling of segmental durations. In *Talking machines: Theories, models and designs*, Elsevier Publishers, Amsterdam, pp. 265–273
- Santen J P H V 1994 Assignment of segment duration in text-to-speech synthesis. *Comput. Speech Lang.* 8: 95–128
- Schröder M 2001 Emotional speech synthesis: A review. In *Eurospeech*, Aalborg, Denmark
- Schröder M 1996 *Can emotions be synthesized without controlling voice quality?* Phonus 4, Research report of the Institute of Phonetics, University of Saarland
- Schröder M, Cowie R 2006 Issues in emotion-oriented computing towards a shared understanding. *Workshop on Emotion and Computing*, Bremen, Germany
- Schröder M, Cowie R, Douglas-Cowie E, Westerdijk M, Gielen S 2001 Acoustic correlates of emotion dimensions in view of speech synthesis. In *7th European Conference on Speech Communication and Technology*. Eurospeech 2001 Scandinavia, 2nd Interspeech Event, Aalborg, Denmark
- Scordilis M S, Gowdy J N 1989 Neural network based generation of fundamental frequency contours. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, Glasgow, Scotland, pp. 219–222
- Shriberg E, Stolcke A 2001 Prosody modeling for automatic speech understanding: An overview of recent research at SRI, In *Prosody in Speech Recognition and Understanding*, ISCA Tutorial and Research Workshop (ITRW), Molly Pitcher Inn, Red Bank, NJ, USA
- Shriberg E, Stolcke A 2004 *Mathematical foundations of speech and language processing* (Philadelphia, PA, USA: Springer)

- Smits R, Yegnanarayana B 1995 Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. Speech Audio Process.* 3(5): 325–333
- Sonntag G P, Portele T, Heuft B 1997 Prosody generation with a neural network: Weighing the importance of input parameters. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Munich, Germany, pp. 931–934
- Sontag E D 1992 Feedback stabilization using two hidden layer nets. *IEEE Trans. Neural Networks* 3: 981–990
- Srikanth S, Kumar S R R, Sundar R, Yegnanarayana B 1989 *A text-to-speech conversion system for Indian languages based on waveform concatenation model*. Technical report no.11, Project VOIS, Department of Computer Science and Engineering, Indian Institute of Technology Madras
- Stylianou Y, Cappe O, Moulines E 1995 Statistical methods for voice quality transformation. In *Eurospeech*, Madrid, Spain, pp. 447–450
- Stylianou Y, Cappe Y, Moulines E 1998 Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech Audio Process.* 6: 131–142
- Sun R, Moore E, Torres J F 2009 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 4: 4509–4512
- Sundermann D 2005 Voice conversion: State-of-the-art and future work. In *Proc. DAGA: 31st German Annual Conf. on Acoustics*, Munich, Germany
- Sundermann D, Bonafonte A, Ney H 2005a A study on residual prediction techniques for voice conversion. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 1: 13–16
- Sundermann D, Bonafonte A, Duxans H, Hoege H 2005b Tc-star: Evaluation plan for voice conversion technology. In *Proc. DAGA: 31st German Annual Conf. on Acoustics*, Munich, Germany
- Taylor P A 2000 Analysis and synthesis of intonation using the Tilt model. *J. Acoust. Soc. Am.* 107(3): 1697–1714
- Teixeira J P, Freitas D 2003 Segmental durations predicted with a neural network. In *Proc. European Conf. Speech Communication and Technology*, Geneva, Switzerland, pp. 169–172
- Tesser F, Cosi P, Drioli C, Tisato G 2004 Prosodic data driven modeling of a narrative style in Festival TTS. In *5th ESCA Speech Synthesis Workshop*, Pittsburgh, USA, pp. 185–190
- t'Hart J, Collier R, Cohen A 1990 *A perceptual study of intonation* (Cambridge: Cambridge University Press)
- Toda T, Saruwatari H, Shikano K 2001 Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 2: 841–844
- Torres-Carrasquillo P A, Gleason T P, Reynolds D A 2004 Dialect identification using gaussian mixture models. In *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Toledo, Spain, pp. 297–300
- Torres-Carrasquillo P A, Sturim D, Reynolds D A, McCree A 2008 Eigen channel compensation and discriminatively trained gaussian mixture models for dialect and accent recognition. In *Interspeech*, Brisbane, Australia, 22–26 September 2008
- Toth A R, Black A W 2008 Incorporating durational modification in voice transformation. In *Interspeech*, Brisbane, Australia, vol. 2, pp. 1088–1091
- Turk O 2007 *Cross-lingual voice conversion*. Ph.D. thesis, Institute for Graduate Studies in Science and Engineering, Bogazii University, Berlin, Germany
- Turk O, Arslan L M 2005 Donor selection for voice conversion. In *Proc. EUSIPCO*, Antalya, Turkey
- Turk O, Arslan L M 2006 Robust processing techniques for voice conversion. *Comput. Speech Lang.* 20: 441–467
- Turk O, Schroder M 2010 Evaluation of expressive speech synthesis with voice conversion and copy re-synthesis techniques. *IEEE Trans. Speech Audio Process.* 18(5): 965–973
- Vainio M 2001 *Artificial neural network based prosody models for Finnish text-to-speech synthesis*. PhD thesis, Dept. of Phonetics, University of Helsinki, Finland
- Vainio M, Altosaar T 1998 Modeling the microprosody of pitch and loudness for speech synthesis with neural networks. In *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia

- Valbret H, Moulines E, Tubach J P 1992 Voice transformation using PSOLA techniques. *Speech Commun.* 11: 175–187
- Vegnaduzzo M 2003 *Modeling intonation for the Italian Festival TTS using linear regression*. Master's thesis, Dept. of Linguistics, University of Edinburgh
- Ververidis D, Kotropoulos C 2006a Emotional speech recognition: Resources, features, and methods. *Speech Commun.* 48: 1162–1181
- Ververidis D, Kotropoulos C 2006b A state of the art review on emotional speech databases. In *Eleventh Australasian International Conference on Speech Science and Technology*, Auckland, New Zealand
- Wang Y, Guan L 2004 An investigation of speech-based human emotion recognition. In *IEEE 6th Workshop on Multimedia Signal Processing*, pp. 15–18
- Watanabe T, Murakami T, Namba M, Hoya T, Ishida Y 2002 Transformation of spectral envelope for voice conversion based on radial basis function networks. In *Proc. Int. Conf. Spoken Language Processing*, Denver, CO, USA, pp. 285–288
- Weber F, Manganaro L, Peskin B, Shriberg E 2002 Using prosodic and lexical information for speaker identification. *Proc. ICASSP*, vol. 1, Orlando, pp. 141–144
- Wua S, Falk b T H, Chan W-Y 2010 Automatic speech emotion recognition using modulation spectral features. *Speech Commun.* doi: 10.1016/j.specom.2010.08.013
- Yoon W, Kyu-SikPark 2007 *Modelling decisions for artificial intelligency* (Berlin, Germany: Springer) pp. 455–462
- Ye H, Young S 2004 High quality voice morphing. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 1: 9–12
- Yegnanarayana B 1999 *Artificial neural networks* (New Delhi: Prentice-Hall)
- Yegnanarayana B, Kishore S P 2002 AANN an alternative to GMM for pattern recognition. *Neural Networks* 15: 459–469
- Yegnanarayana B, Reddy K S, Kishore S P 2001a Source and system features for speaker recognition using aann models. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Salt Lake City, UT
- Yegnanarayana B, Reddy K S, Kishore S P 2001b Source and system features for speaker recognition using AANN models. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, Salt Lake City, Utah, USA, pp. 409–412
- Yegnanarayana B, Veldhuis R N J 1998 Extraction of vocal-tract system characteristics from speech signals. *IEEE Trans. Speech Audio Process.* 6(4): 313–327
- Yin1 B, Ambikairajah E, Chen F 2006 Combining cepstral and prosodic features in language identification. In *18th Int. Conf. on Pattern Recognition (ICPR'06)*, Hong Kong, China, 20–24 August 2006
- Zissman M A, Gleason T P, Rekart D M, Losiewicz B L 1996 Automatic dialect identification of extemporaneous, conversational, latin american spanish speech. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, Atlanta, Georgia, USA