Relationship between spoken Indian languages by clustering of long distance bigram features of speech

K V Vijay Girish
Department of Electrical Engineering,
Indian Institute of Science,
Bangalore, India
Email: vijay.girish@gmail.com

Veena Vijai
Electrical and Electronics Engineering,
BITS Pilani, K. K. Birla Goa Campus,
Goa, India
Email: veena.vijai42@gmail.com

A G Ramakrishnan
Department of Electrical Engineering,
Indian Institute of Science,
Bangalore, India
Email: ramkiag@ee.iisc.ernet.in

***Abstract*—In this paper, a novel method of identifying relationships between languages has been proposed. Our analysis deals with four major Indian languages, as well as Sanskrit and English. We have made use of long distance bigram Mel Frequency Cepstrum Coefficient features and different linkage measures to test the similarities between the clusters formed. Phylogenetic trees have been constructed to provide a visual understanding of the same. The results obtained match with already existing knowledge about language families. For all types of linkage measures, the closest language to Hindi is Marathi and for Tamil, it is Telugu. Since K-medoids give expected language relationships, they are used to learn dictionaries in order to see if they are useful in language identification as well. We have reported the results of one-vs-one classification and found that accuracy improves in the case of English when the weights recovered are multiplied with joint probability of the cluster associated with that medoid.**

***Index Terms*—k-medoid, bigram, distance, phylogenetic, language.**

I. Introduction

A. Motivation

Language is the medium through which we articulate our thoughts. It is a defining element of culture, as languages have evolved to their present form by expanding their vocabularies to remain culturally relevant. Since languages are such a vital part of culture, anthropological history is closely tied to the evolution of language. Presently, there are somewhere between 6000 and 7000 languages spoken around the world. The field of comparative linguistics has two major aims:- (1) to study the similarities and differences between languages, and (2) to use the findings to identify a common parent language. Relationships between languages are an important area of analysis as they shed more light on the spread of culture, especially in cases where two languages are similar but the areas they are being spoken are not in geographic proximity. They can be used to trace civilizations and better understand how languages have evolved to their present form.

India is known to be a linguistically diverse country. Greenberg's diversity index gives India a value of 0.914, i.e. two people selected at random from the country will have different native languages in 91.4% of cases [1]. According to the results of the latest Census in 2001, India has 234 identifiable mother tongues, which mostly belong to two distinct language families, namely Indo-Aryan and Dravidian. Northeast India constitutes a single linguistic region with about 220 languages in multiple language families (Indo-European, Sino-Tibetan, TaiKadai, Austroasiatic) [2] which share common structural features. The Indian Constitution lists 22 official languages.

We have restricted our experiments to six languages in total. We have selected five Indian languages based on the proportion of their native speakers in India and availability of their speech corpora. Hindi, Telugu, Marathi, and Tamil are four of the five most spoken languages in India, in that order. We pick Sanskrit as our fifth language, the reason being that Sanskrit has been accorded the status of a classical language by the Indian Government since it is thousands of years old with a vast and original literary tradition to its name. Moreover, Sanskrit is known to be the etymological source of many other Indian languages, and it is an interesting problem to analyze their relationship with Sanskrit. Lastly, we pick English as our sixth language. English is becoming more ubiquitous by the day, and we wanted our experiments to include a language which has its origins outside the Indian subcontinent to see its correlation with Indian languages and whether it is more easily identifiable. Furthermore, English has been the official state language of Nagaland since 1967 [3]; India is among the top 10 countries in the world in publishing the maximum number of English books every year [4].

Since speech is the primary form of communication, the application of machine learning to speech signals may help to discover or verify similarities between languages. Several organizations are attempting to analyze the evolution of human languages by tracing their historical relationships. The use of phylogenetic trees may prove to be very useful in doing so, since languages closer to one another are more likely to have had a common source. Analyzing the relationships between languages may prove useful in the problem of language identification (LID) as well. The quantification of similarity between languages has a direct link to the level of difficulty

that is faced while performing LID. In addition, phylogenetic trees provide a representation which can be used to derive a hierarchical classifier in order to identify languages at multiple levels. LID has several applications - in information retrieval systems, where queries may be posed in multiple languages, as a preprocessing step in ASR systems, and in matters of national security.

B. Literature Review

The problem of classifying languages using text data has been approached by many researchers but that of determining the relationship between languages by using features extracted from speech signals has not been addressed explicitly. Swadesh [5] approached language comparison on handcrafted word lists whereas Dyen et. al. [6] compared languages based on cognates. Ellison et. al. [7] compared a matrix of inter-language similarities to find the distance between languages. Singh et. al. [8] studied languages using corpus based measures. Rama et. al. [9] constructed matrices of distances between Indian languages using four different distance measures and phylogenetic trees which were given by Saitou [10]. Ghosh et. al. [11] explored the relationship between twelve Indian languages using bigram probabilistic models of a common set of graphemes from each language. All of the above work is focused on text-based language processing.

Dictionary learning (DL) methods used in the past are random selection of features [12] and K-means clustering [13]. The relation between vector quantization and DL was shown by Delgado et. al. [14]. Recursive least squares dictionary learning [15] and K-SVD [16] are other DL algorithms.

We have used orthogonal matching pursuit (OMP) [17] to recover the weights from the concatenated dictionary. Other source recovery algorithms are matching pursuit [18], basis pursuit [19], and active-set Newton algorithm [12]. Girish et. al. [20] approached speaker classification using sparse representation of spectral features as a linear combination of atoms from concatenation of speaker dictionaries.

Language identification from speech has been attempted previously using several methods. Kondrak et. al. [21] have investigated phonetic similarity in speech for different languages. Koolagudi et. al. [22] used Mel frequency cepstral coefficients (MFCCs) to identify fifteen Indian languages. Carrasquillo et. al. [23] used shifted delta cepstral features and Gaussian mixture models (GMMs) for language identification of speech. Zissman et. al. [24] described a set of available cues. Verma et. al. [25] used K-means clustering and support vector machine, while Singh et. al. [26] applied sparse representation to GMM mean shifted supervectors.

In this work, we have clustered distance bigram features, which are extracted from the speech databases of six languages. Similarities between the languages chosen and their relationships are explored using various linkage measures applied on the cluster centroids. Additionally, the centroids are used as dictionary atoms, and the dictionaries used for testing are the concatenated dictionaries of two languages taken at a time. The test bigram features have been expressed as a sparse linear combination of atoms from the concatenated dictionary in order to study the distribution of weights for one-vs-one language classification.

C. Contributions

The main contributions of this paper are: (1) Using multiple distance bigram features, (2) Forming clusters which characterize the sound sequences in a particular language, (3) Using different linkage measures to see the relationship between major Indian languages and creating phylogenetic trees [27] to obtain language groups, (4) Testing whether the features used to obtain the language relationships are also useful in classification, (5) Using a dictionary learning and sparse representation based approach to classify Indian languages.

II. Problem Formulation

A. Feature Extraction

Frames of 40 ms duration are extracted with a shift of 15 ms from the speech utterances taken from the training set of speaker sources. It is assumed that the speech signal remains stationary for 40 ms duration. A uniform sampling frequency of 16 kHz was used for every file. MFCCs are extracted for each frame.

MFCCs were proposed by Davis et. al. [28] in the 1980s. The technique finds the energy associated with frequency bands, which are created according to the mel scale. The logarithmic nature of the mel scale is known to better model the perception of the human ear. MFCCs also contain information about the shape of the vocal tract, which is very useful to analyze speech. From each frame, 42 features are extracted. 14 are MFCCs which include the log energy. The other 28 are the delta and delta-delta coefficients, which model the frame to frame changes in MFCCs, i.e are related to the dynamics of speech.

After the MFCCs for each frame are extracted, two feature vectors are concatenated to create bigram features. Bigram models are a type of probabilistic language model. Each bigram feature represents a particular sound sequence. One bigram feature may contain two different phonemes, the same phoneme, or a phoneme and a silence frame. However, if only consecutive frames are concatenated to form a feature, not all information in the speech utterance is used. To elaborate, a significant drawback of using fixed distance bigram features is that the phoneme duration varies not only with the type of phoneme and variation between languages, but also with context, speaking rate, etc. Therefore, we use the concept of long distance bigram features. For example, a 3-distance bigram feature would be a concatenation of frame 1 and frame 4. For training, we create bigram matrices from 1 up to 8 distance. Each bigram feature, regardless of distance, has 84 dimensions. Figure 1 illustrates the long distance bigram frames at a distance of 3 and 8. It is seen in the figure that frames k and k+8 (8-distance) contain two different phonemes and form a vowel-fricative sequence.

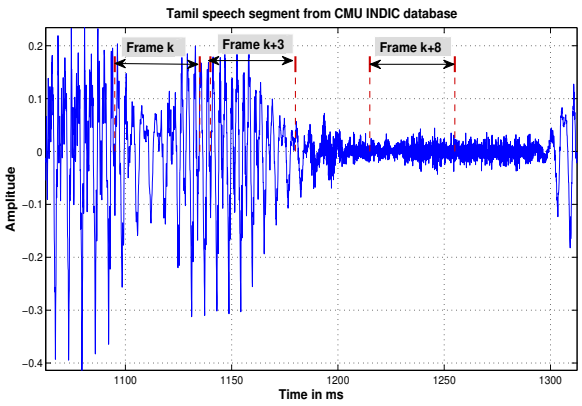


Fig. 1: Illustration of long distance bigram frames at distance of 3 and 8 taken from a Tamil speech segment.

Features having very low energy relative to the average energy of the features are removed - we consider such features to be silence bigrams, which don't provide any information about the language. Features for each language are extracted separately. Dictionaries for each language are learnt separately by the method given in Section II.B.

B. Dictionary Learning

A dictionary is a set of features used to represent large training data. Each column vector of the dictionary is referred to as an atom. Linear combinations of dictionary atoms are used to represent training and test features - such a representation may be exact or approximate within some error margin.

A dictionary is defined as a matrix $\mathbf{D} \in \mathbb{R}^{\mathrm{p}\times\mathrm{K}}$, where p is the dimension of each feature vector and K is the number of atoms. Each atom of the dictionary is normalized to unit L_2 norm. Any real valued feature vector can be represented as $\mathbf{y} \approx \mathbf{D}\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^{\mathrm{K}}$ contains the weights for each atom in vector form. The vector $\mathbf{x}$ is estimated by minimizing the distance $dist(\mathbf{y}, \mathbf{D}\mathbf{x})$, where $dist()$ is a distance metric between $\mathbf{y}$ and $\mathbf{D}\mathbf{x}$ such as L_2 norm. If the dictionary $\mathbf{D}$ is overcomplete, the weight vector $\mathbf{x}$ tends to be sparse. Dictionary learning is the method of constructing the dictionary $\mathbf{D}$, given the training features for each source.

1) K-Medoid Clustering: Clustering is the process of grouping a set of objects into clusters so that objects within a cluster are similar to each other but are dissimilar to objects in other clusters. We have chosen to use K-medoids clustering [29] instead of the well-known K-means clustering for two reasons:- (1) the medoids themselves are representative features from the training and (2) K-medoids is less sensitive to outliers.

Clustering is meant to identify particular sound sequences, i.e. bigram features, which occur often in a language. A value of 200 is chosen for K, with the assumption that it would represent most or all of the sound sequences in a language. Thus, 200 medoids are chosen as atoms to build the dictionary of each language. Therefore, 1200 medoids in total were obtained for six languages.

C. Sparse Representation of Test Features

Given a fixed dictionary and a test feature vector, the process of representing that test vector by estimating the weights corresponding to the dictionary atoms is known as source recovery. Concatenated dictionaries are formed by concatenating all the 200 medoids from two languages at a time. The reasoning is that the test language will be better mapped to atoms of the same language. In this way, all possible pairs of dictionaries were created, and the 15 dictionaries so formed are used for testing.

We have chosen to use OMP [18], since the MFCC feature vectors have both positive and negative values. Sparsity is fixed at 10, so that each feature vector is associated with ten medoids. Therefore, each feature, $\mathbf{y}$ is mapped to medoids of both languages (say, $L1$ and $L2$) by weights recovered using a dictionary obtained by the concatenation of the dictionaries of the two languages, namely $\mathbf{D}_{L1}$ and $\mathbf{D}_{L2}$:

$$\mathbf{y} \approx [\mathbf{D}_{L1}\mathbf{D}_{L2}][\mathbf{x}_{L1}\mathbf{x}_{L2}]^T \tag{1}$$

For each test feature, the sums of the absolute values of the weights associated with the medoids of $L1$ and $L2$ are computed separately as $\sum|\mathbf{x}_{L1}|$ and $\sum|\mathbf{x}_{L2}|$. The feature is classified as belonging to that language with the higher sum. This procedure is repeated for all the features in the files used for testing, and feature-wise accuracies for one-vs-one language classification are computed. Since two frames at a specific distance form a feature vector for classification, the accuracy reported refers to the number of pairs of frames correctly classified, which we call as feature-wise accuracy.

Another approach is to multiply each OMP-recovered weight by a particular factor which can better help determine the relevance of that medoid. The factor in question is calculated using the following algorithm:

1) For one feature, consider the first medoid $M1$ with non-zero weight.
2) Find the number of features in the cluster associated with $M1$.
3) Divide this number by the total number of features in all the clusters of that language.

The result is the joint probability of elements in that cluster i denoted by w_i, which can be used as an additional weight to multiply the OMP-recovered weights. Therefore, languages having higher occurrences of a phoneme sequence will have higher weights if the test feature is similar to that phoneme sequence, and hence there is a higher chance of the test feature being classified correctly.

III. Experimental Setup

A. Databases Used

The speech databases compiled by CMU INDIC, TIMIT, IIT Madras TTS, IIIT Blizzard, Speech Ocean, and MILE Lab (IISc) are used for training, and 10 minutes of data is used to train each language. Although the number of speakers and utterances from each speaker vary between each language, the same number of utterances were taken from each speaker

for a fixed language. For Indian languages, we have 5 to 7 speakers, whereas we have used 20 speakers from the TIMIT database for English. The purpose of using multiple speakers for training is to capture underlying information about the language rather than the speaker.

B. Testing Setup

In the testing phase, one speaker from each language is chosen. It is ensured that the speakers used for testing are completely different from those used for training. The aim is to check whether the model built is speaker-independent. The databases used for testing are TIMIT for English, MILE Lab for Sanskrit, and clips downloaded from All India Radio (AIR) for the other four languages. For the AIR recordings, it is ensured that the clips chosen are full sentences with no music or words from English in between. Ten clips are used for testing each language. The total time of test speech for each language varies from 30 to around 80 seconds.

IV. Experiments, Results and Discussion

We use linear discriminant analysis to find the significant directions that maximize the separation between the language classes. The features corresponding to all the languages projected onto two significant eigenvectors corresponding to the two highest eigenvalues are shown in Figure 2. It is seen that English and Sanskrit have less overlap with other languages than all the other languages.

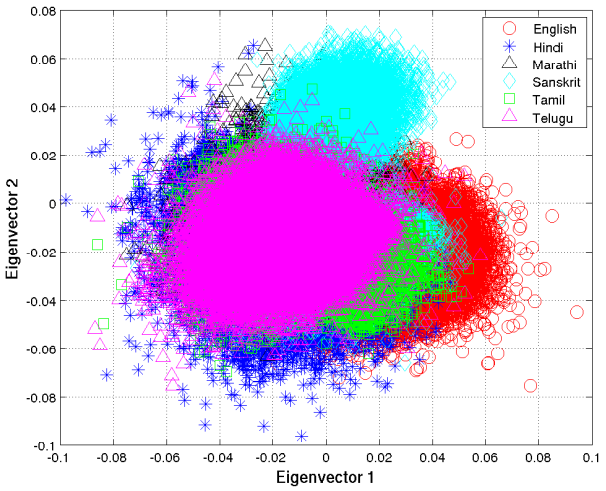


Fig. 2: Illustration of features projected onto two significant eigenvectors using linear discriminant analysis. The top two eigenvalues account for 80%, while the top three account for 91.6% of the discrimination information.

A. Cluster Relationships Using Linkage

Given the dictionary of medoids and the number of cluster elements learnt from different languages, the relationship between the six languages has been shown using various linkages (similar to [30]) as

- *Single linkage:* the smallest distance between medoids $\mathbf{m}_i$, $\mathbf{m}_j$ in the two languages:

$$d_s = min(dist(\mathbf{m}_i, \mathbf{m}_j));\ \mathbf{m}_i \in S^p,\ \mathbf{m}_j \in S^q \quad (2)$$

  S^p, S^q are the set of medoids belonging to two different languages p, q and $dist()$ is L_1 or L_2 norm
- *Complete linkage:* the largest distance between medoids in the two languages: $d_c = max(dist(\mathbf{m}_i, \mathbf{m}_j))$; $\mathbf{m}_i \in S^p,\ \mathbf{m}_j \in S^q$
- *Average linkage:*

$$d_{av} = \frac{1}{n_{S^p} n_{S^q}} \sum_i \sum_j dist(\mathbf{m}_i, \mathbf{m}_j) \quad (3)$$

  n_{S^p} and n_{S^q} are the number of elements in S_p, S_q
- *Centroid linkage*: the distance between the means of the medoids of the two languages:

$$d_{ce} = dist\left(\frac{\sum_i \mathbf{m}_i}{n_{S^p}}, \frac{\sum_j \mathbf{m}_j}{n_{S^q}}\right) \quad (4)$$

- *Weighted Centroid linkage*: the distance between the weighted means of the medoids of the two languages:

$$d_{wc} = dist\left(\frac{\sum_i w_i \mathbf{m}_i}{n_{S^p}}, \frac{\sum_j w_j \mathbf{m}_j}{n_{S^q}}\right) \quad (5)$$

  where w_i, w_j are the joint probabilities of elements in the clusters corresponding to $\mathbf{m}_i, \mathbf{m}_j$ respectively. This measure gives higher weightage to medoids corresponding to clusters containing higher number of elements.

It is observed that L_1 and L_2 norms give similar results. Table I lists the distances between language pairs using various linkages at a bigram distance of 3 (all other bigram distances give similar relationships).

B. Analysis of Cluster Relationships

The medoids found after clustering show expected relationships between languages. We expect Tamil and Telugu to be closer since they both belong to the Dravidian family of languages. Further, speakers of Tamil and Telugu are in closer geographical proximity. Similarly, Hindi and Marathi belong to the Indo-Aryan family of languages and have several common root words. Average, centroid, and weighted linkages capture the best relationships.

For the same language pair, the distance between the medoids is higher when complete and average linkages are used. In the case of Hindi-Hindi, complete linkage has a value of 1.933 and average linkage is 0.757; thus, the distance between the medoids of the same language is large, indicating that there is high variability within every language. The inference that can be drawn is that each medoid represents a different sound sequence, and consequently, the phonetic sequence variability is high. However, the value of single, centroid, and weighted linkage for Hindi-Hindi or any other same language pair is always zero.

Eight phylogenetic trees were constructed from each n-distance bigram feature matrix of medoids where $1 \leq n \leq 8$. All the trees gave very similar results, with only slight changes

TABLE I: Distance matrix of various Indian languages using five different linkages i.e. Single (Sin.), Complete (Comp.), Average (Avg.), Centroid (Cent.) and Weighted (Wei.) using L2 norm. The numbers highlighted in bold correspond to the closest among the five other languages.

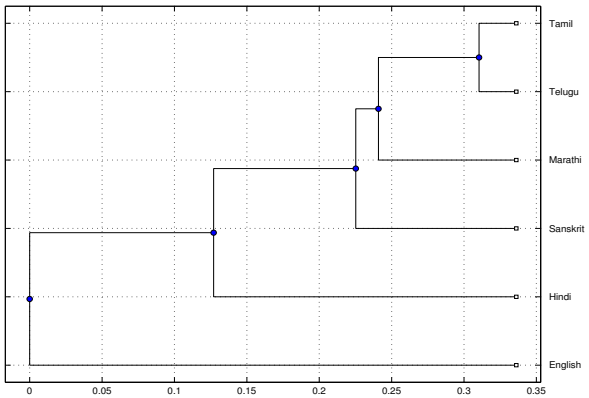
Language	Linkage	English	Hindi	Marathi	Sanskrit	Tamil	Telugu
English	Sing.	0.000	0.409	0.425	**0.252**	0.289	0.305
	Comp.	1.604	1.969	1.918	**1.890**	1.937	1.911
	Avg.	0.619	1.210	1.134	1.377	1.085	**1.010**
	Cent.	0.000	0.977	0.891	1.214	0.783	**0.729**
	Wei.	0.000	1.089	1.035	1.422	0.998	**0.850**
Hindi	Sing.	0.409	0.000	**0.136**	0.222	0.233	0.245
	Comp.	1.969	1.933	**1.936**	1.955	1.938	1.943
	Avg.	1.210	0.757	**0.778**	1.095	0.893	0.866
	Cent.	0.977	0.000	**0.176**	0.808	0.366	0.433
	Wei.	1.236	0.000	**0.440**	0.944	0.639	0.708
Marathi	Sing.	0.425	0.136	0.000	0.124	0.094	**0.096**
	Comp.	**1.918**	1.936	1.936	1.932	1.944	1.941
	Avg.	1.134	**0.778**	0.738	1.086	0.841	0.805
	Cent.	0.891	**0.176**	0.000	0.800	0.261	0.325
	Wei.	1.025	**0.321**	0.000	0.899	0.364	0.421
Sanskrit	Sing.	0.252	0.222	0.124	0.000	0.105	**0.104**
	Comp.	**1.890**	1.955	1.932	1.924	1.920	1.921
	Avg.	1.377	1.095	**1.086**	0.596	1.144	1.183
	Cent.	1.214	0.808	**0.800**	0.000	0.824	0.927
	Wei.	1.273	0.909	**0.860**	0.000	0.901	0.958
Tamil	Sing.	0.289	0.233	0.094	0.105	0.000	**0.026**
	Comp.	1.937	1.938	1.944	1.920	1.923	**1.922**
	Avg.	1.085	0.893	0.841	1.144	0.821	**0.773**
	Cent.	0.783	0.366	0.261	0.824	0.000	**0.138**
	Wei.	1.021	0.558	0.525	0.922	0.000	**0.450**
Telugu	Sing.	0.305	0.245	0.096	0.104	**0.026**	0.000
	Comp.	**1.911**	1.943	1.941	1.921	1.922	1.927
	Avg.	1.010	0.866	0.805	1.183	**0.773**	0.700
	Cent.	0.729	0.433	0.325	0.927	**0.138**	0.000
	Wei.	0.902	0.590	0.474	0.977	**0.291**	0.000

in the value of the distance computed through the linkage function. This proves that even pairs of frames, which are not consecutive, provide information about the language. Figure 3 shows the phylogenetic trees for various linkages except for complete linkage using n-distance bigram features with $n = 3$, since we get similar results for varying n. Since the relationship between languages is unaffected by the distance n, it may be said that this relationship remains intact even with variation in the speaking rate. It is observed that the average and centroid linkages best characterize the relationship between languages, i.e., Telugu-Tamil and Hindi-Marathi are under the same sub-group.

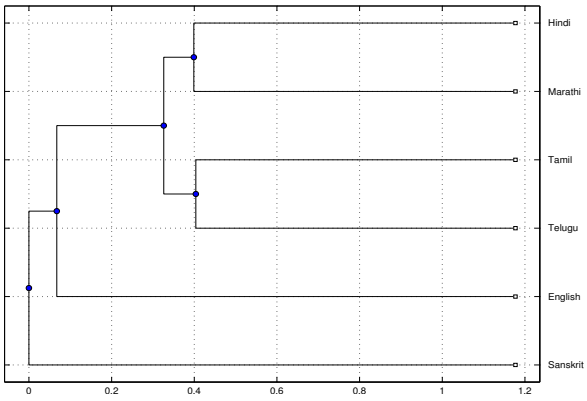
C. K-medoid dictionaries for Classification

The same medoids used to analyze language relationships are used for language classification. Table II shows the feature-wise, one vs one classification accuracies using sparse representation of bigram features. The feature-level accuracy is seen to be poor in some cases. This may be explained by observing that certain phoneme sequences are likely to occur almost equally in both the languages being compared; so the weights may be mapped to the wrong language, resulting in wrong classification of that feature.

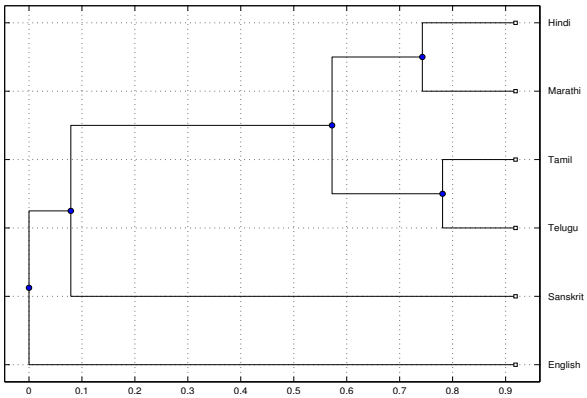
Table III shows the improvement in feature-wise accuracy of various languages against English using joint probability weights over using no weights. It can be inferred that some features from the five Indian languages which are originally misclassified, are properly classified after weighting as the occurrence of those features is higher in the correct language.



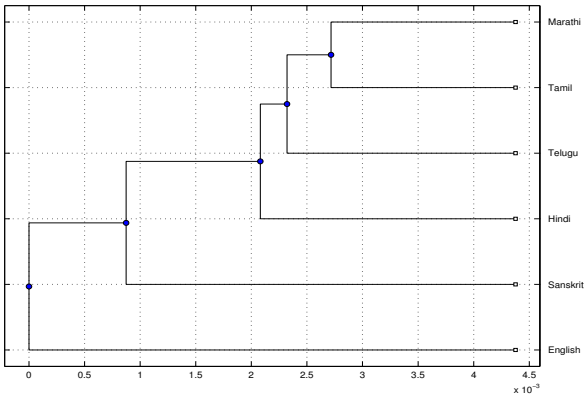
(a) Single linkage



(b) Average linkage



(c) Centroid linkage



(d) Weighted centroid linkage

Fig. 3: Illustration of phylogenetic trees using hierarchical clustering of 6 languages with L_2 norm

TABLE II: Two class feature-wise language classification accuracies obtained using 1 to 8 distance bigram features. The best accuracy is shown at the corresponding distance

Dict. comb.	Test	Accuracy	Distance	Test	Accuracy	Distance
En - Hi	English	48.52	1	Hindi	90.99	4
En - Ma	English	58.47	5	Marathi	76.54	1
En - Sa	English	83.86	5	Sanskrit	78.32	2
En - Ta	English	66.38	1	Tamil	80.37	4
En - Te	English	56.16	6	Telugu	91.06	2
Hi - Ma	Hindi	63.17	4	Marathi	82.19	1
Hi - Sa	Hindi	90.83	6	Sanskrit	66.76	8
Hi - Ta	Hindi	64.17	1	Tamil	74.66	8
Hi - Te	Hindi	71.16	4	Telugu	24.05	1
Ma - Sa	Marathi	59.11	5	Sanskrit	59.62	4
Ma - Ta	Marathi	74.39	1	Tamil	60.15	8
Ma - Te	Marathi	80.84	2	Telugu	67.91	7
Sa - Ta	Sanskrit	65.66	1	Tamil	89.55	5
Sa - Te	Sanskrit	60.25	2	Telugu	83.95	7
Ta - Te	Tamil	79.62	4	Telugu	66.37	2

∗ En: English, Ma: Marathi, Sa: Sanskrit, Hi: Hindi, Ta: Tamil, Te: Telugu.

TABLE III: Improvement of feature-wise accuracy of various languages against English using joint probability weights over using no weights

Test language	Hindi	Marathi	Sanskrit	Tamil	Telugu
Without weights	90.99	76.54	78.32	80.37	91.06
With weights	95.23	84.29	90.28	88.75	90.83

V. Conclusion and Future Work

A novel approach to determining relationships between languages using distance bigram MFCC features is explored. It is seen that centroid and average linkages give the best results in terms of relationships. One-vs-one classification shows good results for some language pairs. However, in other cases, the feature-level accuracy is poor. The results indicate that the features and the methodology used have not been fully effective in discriminating between the languages and that this problem requires more exploration. As future work, we plan to (1) identify only discriminative features from both testing and training features, (2) obtain adaptive distance bigram features to better capture the phonetic variability irrespective of phoneme duration and use them for clustering or dictionary learning, (3) cluster only those bigram features which are composed of two different phonemes, (4) cluster bigram or any n-gram vectors created from feature extraction methods other than MFCC to see if similar or better relationships are obtained, and (5) test other approaches after source recovery is performed for a test segment on the dictionary of medoids, by creating an accumulated classification scheme.

References

[1] Greenberg, [Online] *https://www.ethnologue.com/statistics/country*

[2] [Online] *http://www.iitg.ernet.in/rcilts/phaseI/n_e.html*

[3] [Online] *http://mdoner.gov.in/content/nagaland-2*

[4] [Online] *http://www.buchmesse.de/images/fbm/dokumente-ua-pdfs/2013/book_market_india_2013.pdf_40366.pdf*

[5] M. Swadesh, "Towards greater accuracy in lexicostatistic dating," *International Journal of American Linguistics*, vol-21, pp. 121-137.

[6] I. Dyen, J. Kruskal and P. Black, "An Indoeuropean classification: a lexicostatistical experiment," *American Philosophical Society*, 1992.

[7] T. Ellison and S. Kirby, "Measuring language divergence by intra-lexical comparison," *In Proceedings of the 44th annual meeting of the ACL*, pp. 273-280, NJ, USA, 2006.

[8] A.K. Singh and H. Surana, "Can corpus based measures be used for comparative study of languages?" *In Proceedings of the ACL Workshop Computing and Historical Phonology*, Prague, Czech Republic, 2007.

[9] T. Rama and A.K. Singh, "From bag of languages to family trees from noisy corpus," *Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 2009.

[10] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular biology and evolution*, pp. 406-425, 1987.

[11] S. Ghosh, K. V. V Girish, and T. V. Sreenivas, "Relationship between Indian languages using long distance bigram language models," *Proceedings of ICON-2011: 9th International Conference on Natural Language Processing*, 2011.

[12] T. Virtanen, J. F. Gemmeke, B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, pp. 2277 - 2289, 2013.

[13] A. Coates, and Andrew Y. Ng, "Learning feature representations with K-means," *Lecture Notes in Computer Science, Neural Networks: Tricks of the Trade*, vol. 7700, pp. 561-580, 2012.

[14] K. Kreutz-Delgado, J. Murray, D. Rao, K. Engan, T. Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representations," *Neural Computation*, vol. 15, pp. 349-396, 2003.

[15] K. Skretting, and K. Engan, "Recursive least squares dictionary learning algorithm," *IEEE Trans. Sig. Process.*, vol. 58, pp. 2121-2130, 2010.

[16] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representations," *IEEE Trans. Sig. Process.*, vol. 54, pp. 4311-4322, 2006.

[17] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *Proceedings of Asilomar Conference on Signals, Systems and Computers*, 1993.

[18] S. G. Mallat, and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Sig. Process.*, vol. 41, pp. 3397-3415, 1993.

[19] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43 (1), pp. 129-159, 2001.

[20] K. V. V. Girish, A. G. Ramakrishnan and T. V. Ananthapadmanabha, "Hierarchical classification of speaker and background noise and estimation of SNR using sparse representation," *accepted INTERSPEECH, 2016*

[21] G. Kondrak and T. Sherif, "Evaluation of several phonetic similarity algorithms on the task of cognate identification," *Proceedings of the Workshop on Linguistic Distances*, pp. 43–50, 2006.

[22] S. G. Koolagudi, D. Rastogi, and K. S. Rao, "Identification of language using mel-frequency cepstral coefficients (MFCC)," Procedia Engineering, vol. 38, pp. 3391-3398, 2012.

[23] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds and J. R. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *INTERSPEECH*, 2002.

[24] M. A. Zissman and K. M. Berkling. "Automatic language identification," *Speech Communication*, vol-25, pp. 115-124, 2001.

[25] V. K. Verma and N. Khanna, "Indian language identification using k-means clustering and support vector machine (SVM)," *Engineering and Systems (SCES), 2013 Students Conference on IEEE*, 2013.

[26] O. P. Singh, B. C. Haris and R. Sinha, "Language identication using sparse representation: A comparison between GMM supervector and i-vector based approaches," *IEEE India Conference (INDICON)*, 2013

[27] J. Huelsenbeck, F. Ronquist, R. Nielsen, and J. Bollback, "Bayesian inference of phylogeny and its impact on evolutionary biology," *Science*, vol. 294 (5550), pp. 2310-2314, 2001.

[28] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

[29] H.S. Park and C.H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications*, vol. 36, pp. 3336-3341, 2009.

[30] R Turner, [Online] *http://in.mathworks.com/help/stats/linkage.html*.