International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:12, No:6, 2018

Adaption Model for Building Agile Pronunciation Dictionaries Using Phonemic Distance Measurements

Akella Amarendra Babu, Rama Devi Yellasiri, Natukula Sainath

***Abstract*—**Where human beings can easily learn and adopt pronunciation variations, machines need training before put into use. Also humans keep minimum vocabulary and their pronunciation variations are stored in front-end of their memory for ready reference, while machines keep the entire pronunciation dictionary for ready reference. Supervised methods are used for preparation of pronunciation dictionaries which take large amounts of manual effort, cost, time and are not suitable for real time use. This paper presents an unsupervised adaptation model for building agile and dynamic pronunciation dictionaries online. These methods mimic human approach in learning the new pronunciations in real time. A new algorithm for measuring sound distances called Dynamic Phone Warping is presented and tested. Performance of the system is measured using an adaptation model and the precision metrics is found to be better than 86 percent.

***Keywords*—**Pronunciation variations, dynamic programming, machine learning, natural language processing.

I. Introduction

PATTERN recognition in human is inherent ability to process the sounds and convert them into words. The input stream of speech is segmented into small segments of speech frames and converted into phonemes. The sequence of phonemes thus obtained, are converted into words. Due to pronunciation variability, a particular word will have many sequences of phonemes. These sequences are called pronunciation variations or accents.

The pronunciation varies from person to person, and a person pronounces the same word in a different way under different conditions of emotion, and thus, it results different speech patterns and pronunciations. For example, the speaking style changes when asking for cup of tea in the board room and asking for the same at home. Thus, pronunciation variability depends on the speaker's speaking style, mood, emotions [1], [2] and speaking habits like disfluencies [3]. The length of the vocal cords in the humans differs from person to person. Therefore, the frequencies generated would differ resulting in different pronunciation. The reasons for pronunciation variations are summarized in Fig. 1.

The articulators in humans position themselves in different ways to produce a sound wave. The articulators move continuously to produce different combination of sounds. As the articulators move in anticipation of the next sound, co-articulation effect takes place causing pronunciation variability.

The native language will influence the pronunciation of a person. For example, there are many languages spoken in India. The native language of the language has influence on the pronunciation of English language. Therefore, the pronunciation of English language spoken by different Indians will be different due to influence of the native language.

Fig. 1 Causes for pronunciation variability

In real-time scenario, humans use only a limited numbers of words for communication and therefore, those words and their pronunciation variations are remembered by them. The number of words will vary from person to person and it will vary between 600 to 200 words [4]. The pronunciations will be around 2 to 3 on average per word, and therefore, humans remember around 1200 to 6000 pronunciation variations. It is reasonable to expect the same number of words and their pronunciation in the front-end memory of machines as well [5], [6]. It keeps the size of the pronunciation dictionaries lean and agile.

Related literature is discussed in the next section. The human articulatory system is explained in Section III. Theory related to measuring acoustic distance between two phonemes is explained in Section IV. Process to measure the distance between various phonemes of a language is detailed in this section. Dynamic Phone Warping (DPW) algorithm is explained in Section V. Experiments for measuring acoustic distance between two words or two pronunciations are detailed in this section. Section VI covers the adaptation model which covers the building up of the pronunciation dictionaries using an adaption model. The adaptation model is described in detail. A parameter called critical distance is covered. The results are analyzed in this section.

Akella Amarendra Babu is with St. Martin's Engineering College, Dhulapally, Secunderabad, Telangana State, India – 500100 (corresponding author, phone: +919849934000; e-mail: aababu.akella@ gmail.com).

Rama Devi Yellasiri is with CBIT, Gandipet, Hyderabad, Telangana State, India (e-mail: yrd@cbit.ac.in).

Natukula Sainath (Associate Professor) is with St. Martin's Engineering College, Dhulapally, Secunderabad, Telangana State, India – 500100 (e-mail: nsaainath@gmail.com).

II. Related Works

Pronunciation dictionaries are manually generated using linguistic knowledge. The original speech is recorded, and linguistic experts listen to the recordings and write the transcriptions. For example, The TIMIT speech corpus consists of speech data read by 630 different speakers. There are 438 males and 192 females, from eight different dialect regions. The training set has 420 speakers, and the test set has 210 speakers, all read 10 sentences each. The training and test datasets are mutually exclusive to ensure data isolation and cross validation.

Hahn et al. have used G2P methods for comparing large pronunciation dictionaries [7]. Algorithms are developed for grapheme to phoneme translation in [8]. It is used in applications used for searching the databases and speech synthesis. Adda-Decker and Lamel developed different algorithms for producing pronunciation variants depending on language and speaking style of the speakers [9]. Wester suggested pronunciation models which use both based on knowledge and data-driven [10]. Strik and Cucchiarini surveyed the literature covering various methods for modeling pronunciation variation [11].

Supervised methods are used for preparing pronunciation dictionaries [12]-[15]. These methods are manual, and therefore, they are slow, expensive, and manpower oriented. It is not possible to prepare a comprehensive pronunciation dictionary which covers all accents of all humans. The pronunciation dictionary should cover minimum vocabulary with a capability to enhance and update it as and when required automatically.

Unsupervised methods are used for achieving such a capability. One such method was suggested by Park and Glass [16]. Similar patterns are extracted from raw speech and grouped together and labeled. The new patterns are matched with existing patterns and grouped into the class based on minimum distance criterion. Other methods included vector quantization, deep neural networks, and clustering methods [17].

III. Human Speech Production System

The human organs which participate in the production of speech sounds are called articulators. These are lips, nose, mouth cavity, nasal cavity, velum, tongue, trachea, vocal cords, esophagus, and pharynx cavity. Each articulator has certain preset positions. For example, tongue has three positions – front, mid and back. Vocal cords vibrate while producing voiced sounds like /b/, /d/, /g/, whereas they do not vibrate while producing unvoiced sounds /p/, /t/, /k/.

The mechanism which produces sound is shown in Fig. 2. When humans breathe in, the lungs are filled with air. The air is released through the articulators to produce various sounds. The articulators change their positions temporarily and continuously and produce a sequence of sounds while the humans utter a word [18].

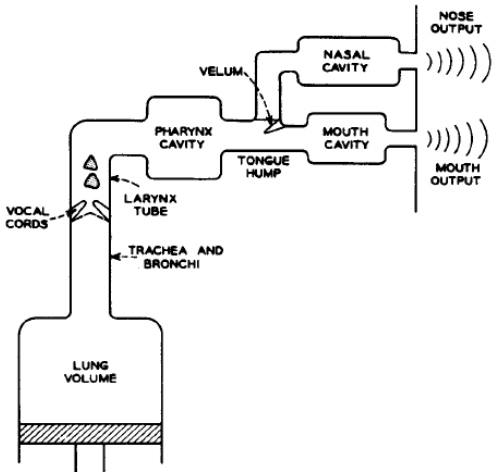
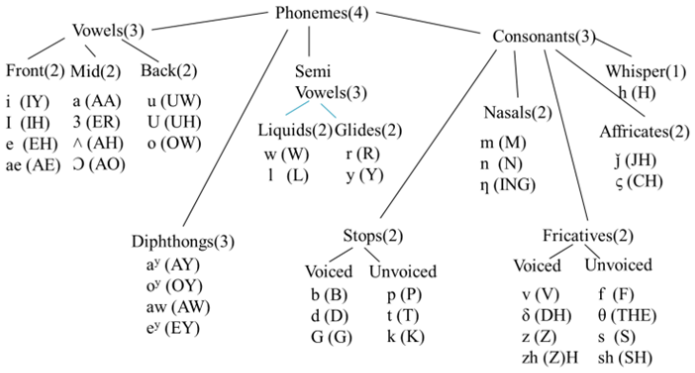


Fig. 2 Speech production mechanism

The sounds which are produced through speech production system are called phonemes. A set of basic sounds characterizes a language. The Standard English language is composed of 39 phonemes. These phonemes are shown in Fig. 3.



Note: The figures within the brackets indicate the weight assigned to the attached feature tag. Weight '1' is assigned to the tags where the figures are not indicated.

Fig. 3 Classification of phonemes with weightages

IV. Inter-Phoneme Distances

There are three steps in computing phoneme to phoneme distance.

a) Allot weights to various levels.
b) Prepare feature sets of the articulators for each phoneme.
c) Compute phonemic distance between various phoneme pairs.

A. Allot of Weightage for Each Level

Table I gives the weightages given to various levels.

TABLE I
Weightage Assigned to Features at Various Levels

Level No.	Features	Weight-age
1	Phoneme (Root level)	4
2	Vowel, diphthong, semi-vowel, consonant	3
3	Front, mid, back, liquids, glides, nasals, stops, fricatives, affricates	2
4	All other features	1

The weights are extrapolated over various phonemes as shown in Fig. 3.

B. Build Feature Sets for Each Phoneme

The second step is to prepare a set of features for producing various sounds. Fig. 4 gives the set of feature for vowel IY and nasal M.

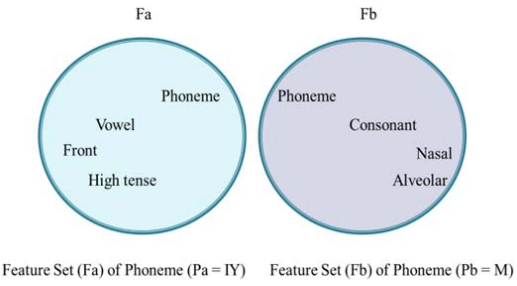


Fig. 4 Feature sets of phonemes IY and M

C. Compute Phonemic Distance

The third step is to find the phonemic distance JD between the phonemes Pa and Pb using the following Jaccard similarity method.

$$JD(Pa, Pb) = 1 - (Fa \cap Fb) / (Fa \cup Fb) \quad (1)$$

The factors in the above equation are computed as shown in Fig. 5.

An example of the above computations is given below.

Phonetic distance between a front vowel IY and a nasal M is computed as follows.

- Feature set Fa for the front vowel (Pa = IY) = {phoneme, vowel, front, high tense}.
- Feature set Fb for the nasal (Pb = M) = {phoneme, consonant, nasal, alveolar}.
- Features common to the feature sets Fa and Fb= (Fa ∩ Fb) = {Phoneme}
- Weightage of the features common to both the feature sets W (Fa ∩ Fb) = 4.
- Total features in both feature sets Fa and Fb = (Fa ∪ Fb) = {phoneme, vowel, front, high tense, consonant, nasal, alveolar).
- Weightage of total features in both the feature sets W ((Fa) ∪ (Fb)) = {4 + 3 + 2 + 1 + 3 + 2 + 1) = 16.
- Jaccard Similarity Coefficient JC (Pa, Pb) = W (Fa ∩ Fb) / W (Fa ∪ Fb) = 4 / 16 = 0.25.
- Jaccard Distance JD (Pa, Pb) = 1 – JC = 0.75.

In the same way, the phonemic distances between all 1521 pairs are computed and shown in Fig. 6.

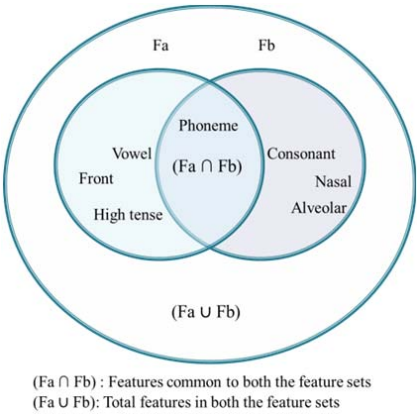


Fig. 5 Common and total feature sets of phonemes IY & M

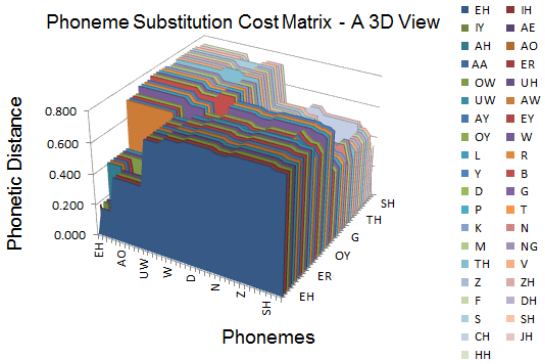


Fig. 6 Phonetic substitution cost matrix – A 3D view

Average of all 1521 phonemic distances is computed and is considered as substitution cost to replace one phoneme by the other in the edit operations. Half of the substitution cost is considered as the cost of deletion or insertion operation and is called an indel.

V. Dynamic Phone Warping (DPW)

DPW algorithm is designed to compute the acoustic distance between a pair of given words. The given words are converted into sequence of phonemes. Dynamic programming technique is used for temporal global alignment. Edit operations are used to compute the phonetic distance. The similarity index between two sequences of phonemes is normalized by the length of the phoneme sequence. The inter-

phonemic distances computed in previous section are used as the cost of edit operations.

The above process of computing the phonetic distance is termed as Dynamic Phone Warping (DPW). Flow chart for DPW algorithm is given in Fig. 7.

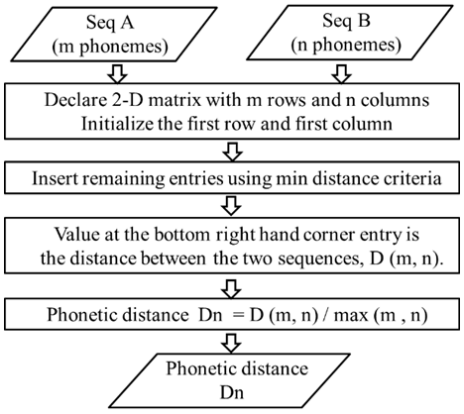


Fig. 7 Flow chart of DPW algorithm

Test setup is given in Fig. 8. Distance between a pair of words is computed using this setup.

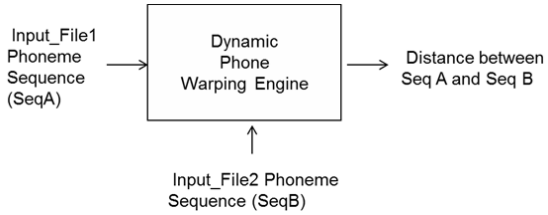


Fig. 8 Test setup to compute phonetic distance using DPW algorithm

The process of distance computations is illustrated with the help of two examples. Results of experiment 1 are given in Fig. 9 and results experiment 2 is given in Fig. 10.

		T	O	M	A	T	O
			T	AH	M	**EY**	T	OW
T O M A T O (1)		**0**	0.31	0.62	0.93	1.24	1.55	1.86
	T	0.31	**0**	0.31	0.62	0.93	1.24	1.55
	AH	0.62	0.31	**0**	0.31	0.62	0.93	1.24
	M	0.93	0.62	0.31	**0**	0.31	0.62	0.93
	AA	1.24	0.93	0.62	0.31	**0.37**	0.68	0.99
	T	1.55	1.24	0.93	0.62	0.68	**0.37**	0.68
	OW	1.86	1.55	1.24	0.93	0.99	0.68	**0.37**

Fig. 9 Phonetic distance between a pair of pronunciations

		T	O	M	A	T	O
			T	AH	M	EY	T	OW
BECAUSE		**0**	**0.31**	0.62	0.93	1.24	1.55	1.86
	B	0.31	0.62	**0.93**	1.24	1.55	1.86	2.17
	IH	0.62	0.68	0.99	**1.3**	1.61	1.92	2.23
	K	0.93	0.99	1.3	1.61	**1.92**	2.23	2.54
	AH	1.24	1.3	1.36	1.67	1.98	**2.29**	2.6
	Z	1.55	1.61	1.67	1.98	2.29	2.6	**2.91**

Fig. 10 Phonetic distance between a pair of words

Two pronunciation variants of the word "TOMATO" are compared in experiment 1. The absolute distance is 0.37. The length of both the sequences is equal to 6. The absolute distance is normalized by dividing the same with the length of the longest sequence. Therefore, the normalized phonetic distance is equal to 0.062.

Experiment 2 gives the results of distance measurements between a pair of phoneme sequences of two words. The absolute distance is equal to 2.91. The length of the longest sequence is equal to 6. Therefore, the normalized phonetic distance is equal to 0.485.

A. Analysis

Results from above two experiments reveal that the inter-pronunciation phonetic distance is less than inter-word phonetic distance.

Exhaustive experiments are conducted and precision of the classification into pronunciations and distinctive words is found to be 86.07%. The results are given in Fig. 11.

Total Number of input word pairs Analyzed	=	201
Total Number of errors	=	28
Classification Error Rate	=	13.93%

Fig. 11 Results of comparison of 201 pairs of words

Statistical hypothesis tests are conducted using z statistic, and the results confirm that the phonemic distance can be used to classify the pronunciations from distinctive words.

VI. Adaptation Model

Architecture of adaptation for preparation online pronunciation dictionaries from input speech is given in Fig. 12.

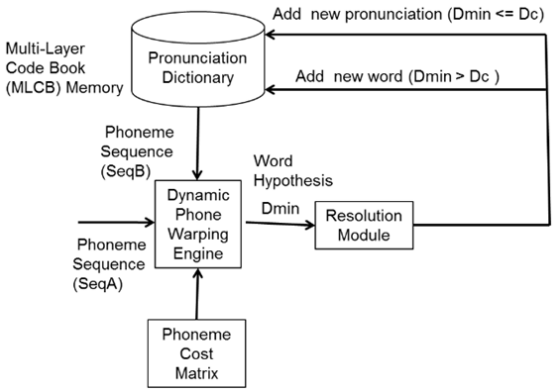


Fig. 12 Architecture of the adaptation model

The input speech is converted into sequence of phonemes and given to adaptation model as sequence SeqA. It is compared with all the phoneme sequences existing in the pronunciation dictionary. The pronunciation dictionary is organised as a multi-layer code book and the pronunciations are grouped into various word classes. A resolution module is designed to classify the input phoneme sequence into a

pronunciation variant of an existing word in the dictionary or a new word based on a parameter called Dc. The parameter is estimated empirically from the data available in the pronunciation dictionary.

The new pronunciations are added as and when a new speaker's speech is input into the system. In case where the pronunciation already exists, the computed phonemic distance Dmin is zero and no action is taken. The input sequence is ignored. Precision measurements show that the performance of the classifier is better than 86%.

VII. Conclusion

Phonemic distance measurements are used in this paper to build online dynamic pronunciation libraries. Firstly, the distances between basic phonemes are computed. The inter-phonemic distances are used to compute the distance between the words. The pronunciations are classified using a new algorithm called Dynamic Phone Warping algorithm. Statistical testing showed that the phonemic distance computation can be used for building dynamic pronunciation dictionaries with 86% accuracy.

Acknowledgment

We thank the staff of speech laboratories at St. Martin's Engineering College for cooperating in conducting the experiments.

References

[1] S. H. Dumpala, K. V. Sridaran, S. V. Gangashetty, B. Yegnanarayana, "Analysis of laughter and speech-laugh signals using excitation source information", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, (2014) 975 – 979.

[2] B. Y. Thati, B. Yegnanarayana, Analysis of breathy voice based on excitation characteristics of speech production, International Conference on Signal Processing and Communications (SPCOM), Bangalore, (2012) 1 – 5.

[3] Jennifer E. Arnold, Michael K. Tanenhaus, "Disfluency effects in comprehension: how new information can become accessible", In Gibson, E., and Perlmutter, N. (Eds.), The processing and acquisition of reference, MIT Press, January 2011, pp. 1-30.

[4] Akella Amarendra Babu, Y. Ramadevi, A. Ananda Rao, "Dynamic pronunciation modeling for unsupervised learning of ASR systems", IETE Journal of Research, vol. 62, no 5, pp. 546-556, 2016.

[5] Janet M. Baker, Li Deng, James Glass, Sanjeev Khudanpur, Chin-Hui Lee, Nelson Morgan, and Douglas O'Shaughnessy, "Research developments and directions in speech recognition and understanding, part 1," IEEE Signal Processing Magazine, vol. 75, May 2009.

[6] Hesham Tolba, Douglas O'Shaughnessy, "Speech recognition by intelligent machines," IEEE Canadian Review – Summer, 2001.

[7] Stefan Hahn, Paul Vozila, Maximilian Bisani, Comparison of Grapheme-to-Phoneme Methods on Large Pronunciation Dictionaries and LVCSR Tasks, IEEE proceedings of INTERSPEECH 2012.

[8] M. Divay and A.-J. Vitale. Algorithms for grapheme-phoneme translation for English and French: Applications for database searches and speech synthesis. Computational linguistics, 23(4):495–523, 1997.

[9] M. Adda-Decker and L Lamel. Pronunciation variants across system configuration, language and speaking style. Speech Communication, 29:83–98, 1999.

[10] M. Wester. Pronunciation modeling for ASR- knowledge-based and data-driven methods. Computer Speech and Language, pages 69–85, 2003.

[11] H. Strik and C. Cucchiarini. Modeling pronunciation variation for ASR: A survey of the literature, Speech Communication, vol. 29 no. 4, pp. 225–246, 1999.

[12] Martijn Wieling, Eliza Margaretha, John Nerbonne, "Inducing phonetic distances from dialect variation," Computational Linguistics in the Netherlands Journal 1, 2011, pp. 109-118.

[13] Ben Hixon, Eric Schneider, Susan L. Epstein, "phonemic similarity metrics to compare pronunciation methods", INTERSPEECH 2011.

[14] Michael Pucher, Andreas Türk1, Jitendra Ajmera, Natalie Fecher, "Phonetic distance measures for speech recognition vocabulary," 3rd Congress of the Alps Adria Acoustics Association 27–28 September 2007, Graz – Austria.

[15] Maider Lehr et al., "Discriminative pronunciation modeling for dialectal speech recognition," INTERSPEECH 2014, Singapore.

[16] Alex, S. Park. and James R. Glass, "Unsupervised Pattern Discovery in Speech," IEEE Transactions On Audio, Speech, And Language Processing, Vol. 16, No. 1, January 2008.

[17] Li Deng, Xiao Li, "Machine learning paradigms for speech recognition: An Overview", IEEE Transactions on audio, speech, and language processing, vol. 21, no. 5, May 2013, pp. 1–30.

[18] L. Rabiner, B. Juang and B. Yegnanarayana, Fundamentals of speech recognition, second ed., Prentice Hall, Englewood Cliffs, N.J., 2010.