DYNAMIC PHONE WARPING – A METHOD TO MEASURE THE DISTANCE BETWEEN PRONUNCIATIONS

Akella Amarendra Babu[1], and Ramadevi Yellasiri[2]

[1]St. Martin's Engineering College, Dhulapally, Secunderabad, India
[2]CBIT, Hyderabad, Telangana, India

ABSTRACT

Human beings generate different speech waveforms while speaking the same word at different times. Also, different human beings have different accents and generate significantly varying speech waveforms for the same word. There is a need to measure the distances between various words which facilitate preparation of pronunciation dictionaries. A new algorithm called Dynamic Phone Warping (DPW) is presented in this paper. It uses dynamic programming technique for global alignment and shortest distance measurements. The DPW algorithm can be used to enhance the pronunciation dictionaries of the well-known languages like English or to build pronunciation dictionaries to the less known sparse languages. The precision measurement experiments show 88.9% accuracy.

KEYWORDS

Natural Language processing, word distance measurements, pronunciation dictionaries.

1. INTRODUCTION

Pronunciation dictionaries are not available for all languages and the accents of various regions. This paper aims to build online pronunciation dictionaries using sound distance measurements. Human beings hear a word; compare it with the words in the memory and select the word which highest similarity to the input word. The objective of this paper is to follow the technique adopted by the human beings and prepare the pronunciation dictionaries. The primary focus of this paper is to measure distances between and sounds and to use this data to measure the distances between the words.

The reasons for the pronunciation variability are as under:

1.1 Speaker's Accent: The accent of the speaker depends on his mother tongue [1, 2]. The difference is negligible in respect of the speakers of the same country. But the difference is glaring in respect of foreign speakers.

1.2 Speaker's Emotions: The pronunciation of the same word would be different when spoken with different emotions like joy, love, anger, sadness and shame [3, 4].

1.3 Speaking Style: The speaker style varies when speaking to various people. The same name is spoken with different pronunciation while addressing an office peon and while addressing your friend.

Dhinaharan Nagamalai et al. (Eds) : ACSIT, ICITE, SIPM - 2018
pp. 65–73, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80805

1.4 Speech Disfluencies: There will be lot of gaps and filler sounds while speaking. It interrupts the normal of the human beings. This phenomenon creates pronunciation variability [5].

The natural speech results in generating different formant frequencies for the same spoken phoneme due to above reasons. Therefore, the phoneme sequences generated for a word will vary and depend on the speaker's accent, mood and the context [6].

The next section reviews the literature related to this work. Section three covers the theoretical background to the proposed algorithm of Dynamic Phone Warping (DPW). Section four covers the measurement of distance between various phonemic sounds produced by human beings. DPW algorithm is described in section five. Experimental details and analysis of results are discussed in section six. Some of the applications which can be developed based on the DPW methods of phoneme distance measurements are discussed in section seven.

2. RELATED WORK

The methods proposed for preparation of pronunciation dictionaries are discussed in this section. Pronunciation dictionaries are manually generated using linguistic knowledge are covered knowledge based methods. They are Grapheme-to-Phoneme (G2P) and Phoneme-to-Phoneme (P2P) conversions.

Stefan Hahn, Paul Vozila and Maximilian Bisani 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. M. Adda-Decker and L Lamel developed different algorithms for producing pronunciation variants depending on language and speaking style of the speakers [9]. M. Wester suggested pronunciation models which use both based on knowledge and data-driven.

Knowledge based methods use phonological linguistic rules which generally cannot capture the irregularities in the spontaneous speech. There is a gap between the linguistic knowledge found in the literature and the variations generated in the spontaneous speech. H. Strik and C. Cucchiarini surveyed the literature covering various methods for modeling pronunciation variation [10].

3. THEORETICAL BACKGROUND

Figure 1 shows the schematic view of the vocal mechanism in humans. Vocal tract is one of the main articulators. It connects vocal cords to lips and consists of pharynx and mouth. The pharynx is the connection between esophagus to the mouth. The total length of the vocal tract in a male is 17 cm. The cross sectional area varies from zero when is completely closed to a maximum 20 square cm. Tract between velum and nostrils is called nasal tract. It produces the nasal sounds when the velum is lowered and the nasal cavity is acoustically connected to the vocal tract.

When the human takes breath, air enters the lungs. When the air escapes, the vocal cords are caused to vibrate. Articulators like jaws, longue, mouth, lips and velum adjust their positions and produce the desired sounds.

Linguistically distinct speech sounds are called phonemes. A set of articulators are used to generate a phonetic sound. When the human being speaks a word, the articulators change their positions temporally to generate a sequence of phonetic sounds. The articulators are the vocal cords, pharyngeal cavity, velum, tongue, teeth, mouth, nostrils, etc. The articulators and the positions they assume while generating a phoneme are called features corresponding to that phoneme.

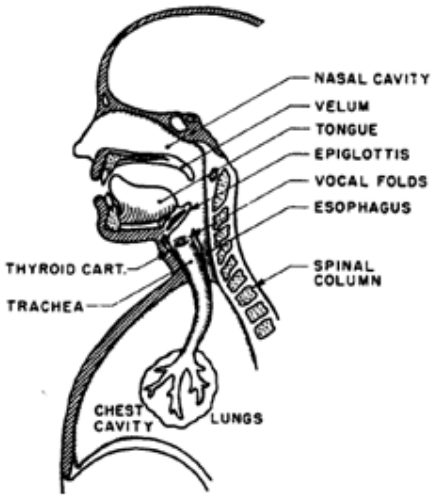


Figure1: The schematic diagram of vocal mechanism

The Standard English language has thirty-nine phonemes as shown in Figure 3.2. It consists of eleven vowel sounds, four diphthongs, four semi vowels, four nasal sounds, six stops, eight fricatives, two affricates and one whisper.

4. DISTANCE BETWEEN VARIOUS PHONEMES

Distance between one phoneme to another is termed as phonetic distance between them. It is measured using the configuration of the articulators while generating the two phonemes [12, 13]. The positions assumed by the articulators while generating the phoneme sound are called its feature set.

The methodology followed for computation of distances between various pairs of phonemes is described in this section.

Table 1: Weightage assigned to features at various levels

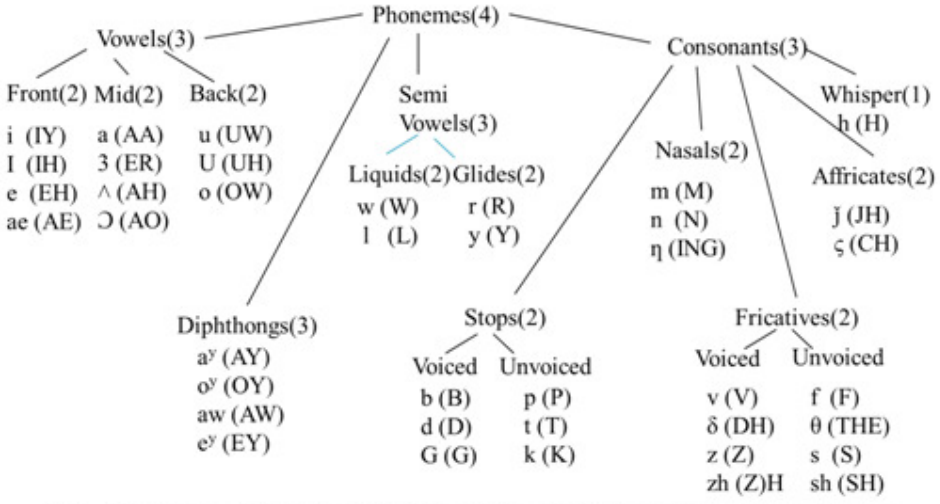
Level No.	Features	Weightage
1	Phoneme (root level)	4
2	Vowel, diphthong, semi-vowel, consonant	3
3	Front, mid, back, liquids, glides, nasals, stops, fricatives, affricates	2
4	All other features	1

Articulatory feature sets for various phonemes are extracted from the classification chart shown in figure 2. Features at various levels in the chart are assigned weightages as shown in table 1. The weights are extrapolated on the classification chart as shown in figure 2. The feature sets and their weightages for various phonemes are worked out.

The distance between two phonemes Pa and Pb is given by the Jaccard Coefficient

$$JC(Pa, Pb) = [1 - k * (Fa \cap Fb) / (Fa \cup Fb)] \quad (1)$$

K is a constant which is calculated experimentally.



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.

Figure 2: Classification of Standard English phonemes with weights assigned to various features

The computations are as under.

Example: 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.

3.2 Phoneme Substitution Cost Matrix

The substitution cost is the cost for replacing one phoneme with the other. The phonetic distances between 1521 pairs of phonemes are estimated.

3.3 Edit Operations

The three edit operations are substitution, insertion and deletion operations. Half of the substitution cost is taken as one Indel.

5. DPW ALGORITHM

DPW algorithm uses dynamic programming for global alignment. Needleman-Wunsch algorithm is modified to suit the usage of the algorithm in DPW algorithm. The phoneme cost matrix is used in place of similarity matrix and the Indel is used in place of the gap penalty. All the cells in the similarity matrix are filled using the substitution, and indel values. Bottom right hand corner cell value is phonetic distance between the given sequences.
Flow chart for DPW algorithm is given in figure 3.

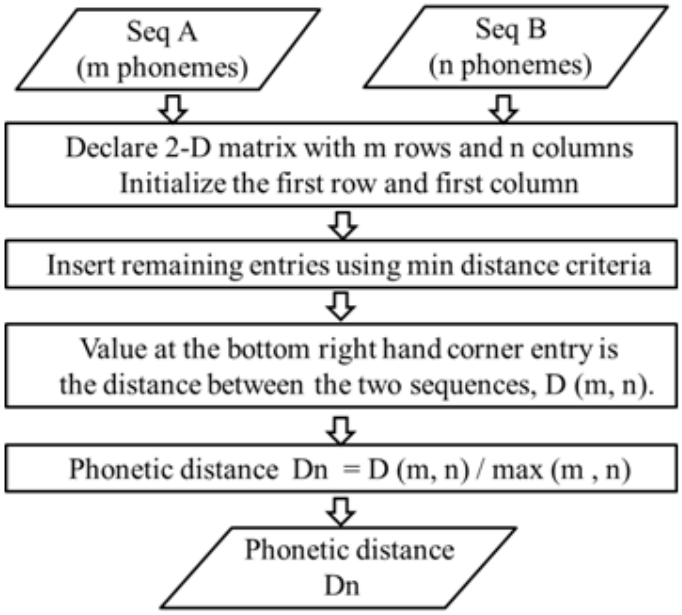


Figure 3: Flow chart for DPW algorithm

6. EXPERIMENTATION AND RESULT ANALYSIS

In this section, computation of phonetic distances using DPW algorithm is illustrated. The pronunciations and words are represented by their phoneme sequences.

Data Source

Datasets are drawn from CMU pronunciation dictionary (CMUDICT). The CMUDICT has 130984 orthographic words followed by its phoneme sequences, out of which 8513 words have multiple pronunciation phoneme sequences.

Experimental Setup

The experimental setup to measure the phonetic distances using DPW algorithm is shown in figure 4.

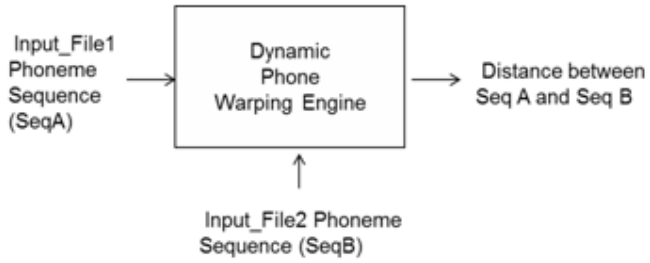


Figure 4: Test setup to compute phonetic distance using DPW algorithm

One sequence of phonemes are listed in file1 and the another sequence is taken File2. The algorithm described in flow chart is used to compute distance between the two sequences.

6.1 RESULT ANALYSIS

Experiment 1: Computation of phonetic distance in five different test cases is experimented and the results are recorded table 4.1.

The summary of normalized phonetic distances for test cases 1 to 5 is shown in table 2.

Table 2: Summary of normalized phonetic distances for test cases 1 to 5

Test Case No.	Details	Normalized Phonetic Distance
1	Same word, Same pronunciation compared with itself	0
2	Same word with different pronunciations	0.062
3	Same word with different pronunciations of unequal length	0.14
4	Different words with an unequal number of phoneme sequences	0.485
5	Different words with an equal length of phoneme sequences	0.358

In table 4.1, inter-pronunciation distances are computed in test cases 2 and 3 and inter-word distances are computed in test cases 4 and 5. It may be noted that the inter-word distances are greater than half of the Indel value and inter-pronunciation distances are less than half of the Indel value.

The results from the above five test cases reveal that the inter-pronunciation phonetic distance is less than inter-word phonetic distance.

The hypothesis resulted from the experimental results is that the distance between two pronunciations (D_A) of a word is significantly less than phonetic distance between any two words (D_W). It is possible to classify that a given sequence of phonemes is pronunciation variant or a new word itself.

$$D_A < D_W \quad (2)$$

Statistical z statistic tests have validated and confirmed the above hypothesis.

Experiment 2: 109 pairs of phoneme sequences and average phonetic distance calculations.

Criterion for Error Count Let the threshold phonetic distance for classification is half the value of one Indel. Let us call it as Critical Distance (Dc). The criterion for counting errors is as follows:

Let Wa and Wb be serial numbers of the words in test_file1 and test_file2 respectively. An error is counted in case the normalized phonetic distance (Dn) between a pair of pronunciations (Word A (Wa) = Word B (Wb)) is greater than Dc or the Dn for a pair of different words is less than or equal to Dc.

If {((Dn > Dc) && (Wa = Wb))} || {(Dn <= Dc) && (Dn != Wb)}

Increment Error_count; (3)

Results

Result summary is shown in Table 3.

Table 3: Results of comparison of 109 pairs of words

Total Number of input word pairs analyzed	=	109
Total Number of errors	=	12
Classification Error Rate	=	11.01%

Analysis

Experiment 2 gives the DPW results of 109 pairs of phoneme sequences corresponding to 55 different words with different lengths of phoneme sequences. The results show that the average normalized distance between the any two pronunciations is 0.069 and the average phonetic distance between the any two different words is 0.247. These results support that the inter-pronunciation phonetic distance is less than inter-word phonetic distance. Further experiments are carried out with larger datasets and hypothesis testing is carried out using z test statistic.

7. APPLICATION OF THE MODEL

The DPW algorithm is generic and can be used to classify a given sequence of phonemes corresponds to a new word or a pronunciation variant of an existing word in the dictionary. A critical distance threshold criterion can be developed to classify the given utterance into pronunciation variant or new words.

The phonemic distance measurements using DPW algorithm is independent of any particular language. Basically, it is using the phoneme set of that particular language. It can be used for any language generically. For instance, to utilise this algorithm for an Indian language, say Telugu, the phoneme set of Telugu language will be used in place the phoneme of English language.

Well-known languages like English have well-developed pronunciation dictionaries. But there are many sparse languages which are less known and do not have the readily available pronunciation dictionaries. The DPW algorithm can be used to build the pronunciation dictionaries for the sparse languages.

Pronunciation of a word differs from person to person. The DPW technology can be used to identify a speaker based on the pronunciation style. Therefore, the DPW algorithm can be used for speaker recognition.

A human being can understand and register pronunciation variability. But the machines like Interactive Voice Response (IVR) systems need supervised training to do so. The DPW technology can be used to build the online pronunciation capability in IVR systems. This the DPW technology has speech recognition applications.

The Information Technology (IT) companies have customers from all over the globe. The employees of the foreign companies will have the accent of their native language. The employees of Indian IT companies are educated to understand the accent of their native language. There is a challenge for the Indian employees understand the accent of foreign customers during initial stages. A pronunciation translator can be built to help the Indian employees to get over the above problem.

8. CONCLUSIONS

In this paper, the generation and perception of human speech is described. The phonemes are classified based on the articulatory features. The articulatory feature sets for generation of phonetic sounds are worked out. Weightage is assigned to each feature in the feature set and the total weightage of the feature set for each phoneme is computed. Phonetic distances between various pairs of the phonemes of the Standard English language are computed.

DPW algorithm is described with the help of a flow chart and is illustrated with the help of test cases. The analysis of the results led to the formulation of a hypothesis which gives the relationship between the inter-pronunciation distance and the inter-word distance. The hypothesis is tested at 1% significance level using z-test statistic.

REFERENCES

[1] Samuel Silva, António Teixeira, Unsupervised segmentation of the vocal tract from real-time MRI sequences, Computer Speech & Language, Vol 31, Volume 33, Issue 1, Pages 25-46, 2015.

[2] S.-A. Selouani, Y. Alotaibi, W. Cichocki, S. Gharsellaoui, K. Kadi, Native and non-native class discrimination using speech rhythm- and auditory-based cues, Computer Speech & Language Volume 31, Issue 1, Pages 28-48, 2015.

[3] Mahesh Kumar Nandwana, Ali Ziaei, John H. L. Hansen, Robust unsupervised detection of human screams in noisy acoustic environments, IEEE Proceedings on Audio, Speech and Signal Processing, ICASSP 2015 , 161 – 165, 2015.

[4] B. Yegnanarayana, S. Rajendran, Hussien Seid Worku and N. Dhananjaya, Analysis of glottal stops in speech signals, IEEE Proceedings, INTERSPEECH 2008, Brisbane, Australia, pp. 1481-1484, Sep. 22-26, 2008.

[5] 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.

[6] Baker, J. M., Li Deng, Sanjeev Khudanpur, Chin-Hui Lee, James Glass and Nelson Morgan. 2009. Historical Developments and future directions speech recognition and understanding. IEEE Signal Processing Magazine, Vol 26, no. 4 78-85, Jul 2009.

[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, 29(4) (1999) 225–246.

[12] L. Rabiner, B. Juang and B Yegnanarayana, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, N.J., 2010.

[13] Amos Tversky, Features of Similarity, Psychological Review. Vol 84, Number 4, July 1977.

AUTHORS

Akella Amarendra Babu received B. Tech (ECE) degree from JNU, M. Tech (CSE) degree from IIT Madras, Chennai and Ph. D. degree in Computer Science and Engineering from JNTUA, Ananthapuramu. He served Indian Army for 23 years as Lt Colonel in Corps of Signals and has 12 years of senior project management experience in corporate IT industry. He has two and half years research experience on mega defense projects in DLRL, DRDO and is working as Professor of CSE department in Engineering Colleges at Hyderabad. He published more than 20 research papers in various national and international conferences and journals. He published a book chapter and has a patent. His research interests include speech processing, computer networking, information security and telecommunications. He is a Fellow of IETE, life member of CSI and IAENG.

Y Rama Devi received B.E. from Osmania University in 1991 and M. Tech (CSE) degree from JNT University in 1997. She received her Ph. D. degree from Central University, Hyderabad in 2009. She is Professor, Chaitanya Bharathi Institute of Technology, Hyderabad. Her research interests include Speech and Image Processing, Soft Computing, Data Mining, and Bio-Informatics. She is a member for IEEE, ISTE, IETE, IAENG and IE. She has published more than 50 research publications in various national, international conferences, proceedings and journals.