IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE)
e-ISSN: 2278-1684,p-ISSN: 2320-334X
PP. 20-24
www.iosrjournals.org

Robust Command And Control System For Military Weapons

Akella Amarendra Babu, , Siva Prasad Nadakuduru
St. Martin's Engineering College, Hyderabad, Telangana, IndiaICFAI Business SchoolHyderabad, Telangana, India

ABSTRACT:- *Military Weapons Operate In Harsh Battlefield Noisy Environment And The Voice Command And Control Need Robust Man Machine Interface (Mmi). In This Paper, We Present Two Methods To Achieve Robustness. Firstly, Incorporate Error Tolerance Techniques And Secondly, Adapt Pronunciation Of The Speaker. Phonemic Distance Measurement Algorithms Are Used For Measuring The Phonetic Distance Between Various Commands And Their Pronunciation Variations. The Algorithms Are Implemented Using Java And The Data Sets Are Taken From Timit Database And Cmu Pronunciation Dictionary. The Performance Of The New System Is More Robust When Compared With Existing Systems.*

Nomenclature
Machine learning, automatic speech recognition systems, man machine interface, pronunciation adaptation.

I. Introduction

In the battle, command and control messages are given by various commanders to the troops under their command to achieve operational fighting movements. These commands are given to the troops through the voice commands. As the troops move forward, the enemy fires artillery guns, mortars which drop the shells on the advancing troops. The shells are burst after landing on the ground and make high decibel sounds. The fighting tanks of our military advancing alongside produce severe sounds as they run past the troops. Besides, the fighter aircrafts of our air force and that of enemy produce high volumes of sounds. Under the above battle condition, the commands given by the commanders to their troops get corrupted.

Various artillery weapons are controlled by the man machine interface by the artillery commanders while neutralizing the enemy positions. After spotting the enemy aircrafts, the anti-aircraft missiles are fired through missile command and controls systems. The command and control is through the voice commands through MMI. Forward Error Correction (FEC) using Hamming codes are used to achieve error correction.

The nature of the speech signal is unique. Firstly, there is a lack of invariance among various phonemes due to co-articulation effect. The articulators move early in anticipation of the subsequent phonemes. The acoustic waveform generated by a given phoneme depends on the context. It results in a big difference in acoustic waveform for the same phoneme and a very little difference between some phonemes. Secondly, the length, size and shape of the vocal tract differ from speaker to speaker. It results in generating different formant frequencies for the same phoneme. Therefore, the phoneme sequences generated for a word will vary and depend on the speaker's accent, mood and the context [1]

ASR systems are trained using training speech corpus and tested with "everyday speech". The training speech waveforms are labeled manually. Labeling "everyday speech" is time consuming, manpower intensive and extremely expensive. Therefore, it is impossible to label "everyday speech" corpus. The human speech recognition system has the inherent ability to learn from the "everyday speech" without labeling [2]. Adaptation of the process followed by the human speech recognition system will help incorporating this ability in Automatic Speech Recognition (ASR) systems.

The human speech recognition system follows a process to learn from the conversation among human beings. It hears a sentence, compares each word with the words in its memory. It hypothesizes a word which has maximum similarity, checks the context and accepts the same. The process is simple if the pronunciation already exists in the memory. In case, the pronunciation doesn't exist in the memory, it enrolls the new pronunciation in its memory and uses the same for future references [3]. This process is known as unsupervised adaptation to the environment.

The above process may be incorporated in ASR systems to achieve unsupervised learning and adaptation to the environment. The significant step in the above process is to compare the analysis phoneme sequence with the phoneme sequences corresponding to the words existing in the ASR system memory. The word with maximum similarity is hypothesized and checked for the context. The critical step in the process is to

find the similarity between two phoneme sequences or in other words, finding the phonetic distance between two phoneme sequences.

In this paper, we present an algorithm which uses Dynamic Phone Warping (DPW) as a measure of finding the distance between two phoneme sequences [4]. We developed critical distance criteria to conclude whether the analysis phoneme sequence corresponds to the pronunciation variation of a word existing in the memory or the analysis word corresponds to a new word which is not in the vocabulary of the ASR system memory. The critical distance parameter is determined using data driven approach. An application is developed to implement the DPW algorithm and tested using different inputs.

This paper is organized as follows. Related work is covered in section 2. Forward error correction is covered in section 3. Adaption of pronunciations is covered in section 4. Implementation and results are covered in section 5. Interpretation of results and conclusions are covered in section 6. Section 7 covers the future enhancements.

II. Related Work

The baseform pronunciation of a word is obtained using different algorithms. Firstly, the orthographic spelling of the word is used to derive the sequence of phonemes. Secondly, when an ASR system encounters Out of Vocabulary (OOV) word, the correct word is supplied to the system. The ASR system correlates the phoneme sequence to the new word and enrolls the new word into its memory. Thirdly, the phone transition costs are used to calculate the transition penalty. The above methods are combined and the combined score is calculated through formulation. The highest scoring phoneme sequence is enrolled as the baseform pronunciation of the word in the pronunciation dictionary [5].

Another way to build the pronunciation dictionary is suggested by Trym Holter et al [6-9]. It is based on maximum likelihood criteria. The pronunciation of a word is speaker dependent. The input spoken word is converted to a sequence of phonemes and compared with the baseform pronunciations available in the dictionary. The Maximum likelihood criteria are used to select the baseform pronunciation and pronunciation variations. In such cases, there will be more than one pronunciation corresponding to a single word. The different pronunciation variations represent different accents of speakers. An unsupervised algorithm for speech recognition was suggested by A. S. Park, *et al.* The algorithm extracts the patterns from the raw speech data input waveform. The similar patterns are grouped together and labeled [2, 10]. The process of speech recognition is carried out off-line.

III. Error Protection

In 1950, Hamming introduced the (7,4) code. It encodes 4 data bits into 7 bits by adding three parity bits. Hamming (7,4) can detect and correct single-bit errors. With the addition of an overall parity bit, it can also detect (but not correct) double-bit errors. In MELP algorithm, Forward Error Correction (FEC) is implemented in the unvoiced mode only. The parameters that are not transmitted in the unvoiced mode are the Fourier magnitudes, band pass voicing and the aperiodic flag. FEC replaces these 13 bits with parity bits from three Hamming (7,4) codes and one Hamming (8,4) code. However, no error correction is provided for the voiced mode MELP coder. The DES/3DES encryption algorithms process input data in 64-bit blocks. 54 bits are allocation for MELP encoded speech frame. 1-bit per frame is added for authentication. Remaining 9 bits are utilised for FEC parity bits for voiced mode from three Hamming (7, 4) codes [21 – 24].

IV. Adaptation Of Pronunciations

The Standard English language has 39 phonemes. They are listed in Annexure A. A set of articulators are used to generate a phonic sound. When human being speaks a word, the articulators change their positions temporally to generate a sequence of phonic 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. The phonetic distances from the front vowel /e/ to all other phonemes are given in Annexure B.

Dynamic Phone Warping

The dynamic phone warping is a variant of dynamic programming technique. It estimates the minimum distance between two sequences of phonemes ***There are two steps in calculating the phoneme distance between the above two sequences. The first step is the declaration of a matrix D with m rows and n columns and its initialization. The first column and the first row are initialized.***

The second step is to fill the remaining entries of the matrix table. It is done using the following formula:

The value at the bottom right hand corner of the matrix table gives the distance between SeqA and SeqB. This distance is normalized by the length of the maximum of the two sequences.
The above procedure is illustrated in Figure 1. The two phoneme sequences correspond to two different pronunciations of the word 'TOMATO'.

Alignment:
SeqA = T AH0 M **EY1** T OW2
SeqB = T AH0 M **AA1** T OW2

- **Figure 1(a).** Alignment of phoneme sequences.

0.00	0.18	0.36	0.54	0.72	0.90	1.08
0.18	**0.00**	0.18	0.36	0.54	0.72	0.90
0.36	0.18	**0.00**	0.18	0.36	0.54	0.72
0.54	0.36	0.18	**0.00**	0.18	0.36	0.54
0.72	0.54	0.36	0.18	**0.36**	0.54	0.72
0.90	0.72	0.54	0.36	0.54	**0.36**	0.54
1.08	0.90	0.72	0.54	0.72	0.54	**0.36**

Figure 1(b). Matrix table.

	Word	Phoneme String
Analysis Word	TOMATO	T AH0 M EY1 T OW2
Comparison word	TOMATO(1)	T AH0 M AA1 T OW2
Maximum phoneme string length = 6		
Normalised Phonetic distance = 0.06		

Figure 1(c). Normalized distance between the two pronunciations.

Figure 1 shows the process of calculating the distance between two phoneme sequences. Figure 1(a) gives the alignment of the phoneme sequences corresponding to two pronunciations. Figure 1(b) shows the matrix table filled with the values calculated during the alignment using DPW algorithm. Figure 1(c) shows the calculations for obtaining the normalized phoneme distance.

V. Implementation And Results

CMU pronunciation dictionary CMUDICT is used to extract the words and pronunciation phoneme sequences. The CMU pronunciation dictionary has 130984 orthographic words followed by its phoneme sequences, out of which 8513 words have multiple pronunciation phoneme sequences.
In test case 1, ten words are selected randomly from the CMUDICT which have two or more different pronunciations. The baseform pronunciation phoneme sequences are listed in the test file1 and the pronunciation variation phoneme sequences are listed in the test file2. The total number of pairs compared is 100. The results obtained for test case 1 with $\gamma = 0.5$ are given in Figure 2.

EXPERIMENTAL RESULTS
Total Number of comparisons = 100
Total Number of errors = 2
Word Error Rate (WER) = 2.00%

Figure 2: Results of test case 1

The test case 2 is conducted with 400 pairs of comparisons. The results obtained for the test case 2 are given in the Table 1.

Table 1. Results of test case 2

Gamma Value	No of comparisons	Errors	WER %
0.20	400	12	3.00
0.25	400	8	2.00
0.30	400	7	1.75
0.35	400	5	1.25
0.40	400	4	1.00

0.45	400	5	1.25
0.50	400	4	1.00
0.55	400	5	1.25
0.60	400	7	1.75
0.65	400	7	1.75
0.70	400	6	1.50
0.75	400	7	1.75
0.80	400	14	3.50
0.85	400	19	4.75

The value of γ is varied between 0.2 to 0.85 for all the three test cases and the summary of the results obtained for all the three test cases, for various values of γ is given Table 2 and the same is presented graphically in figure 4.

Table 2. Summary of Results

Gamma Value	100 comparisons	400 comparisons	3025 Comparisons
0.20	7.00	3.00	1.42
0.25	4.00	2.00	1.26
0.30	4.00	1.75	1.09
0.35	3.00	1.25	0.86
0.40	2.00	1.00	0.79
0.45	3.00	1.25	0.89
0.50	2.00	1.00	0.76
0.55	3.00	1.25	0.86
0.60	3.00	1.75	0.99
0.65	3.00	1.75	1.09
0.70	4.00	1.50	1.16
0.75	4.00	1.75	1.32
0.80	5.00	3.50	1.88
0.85	6.00	4.75	2.41

The results of test cases 1, 2 and 3 with 100, 400 and 3025 pairs of comparisons respectively are shown graphically in Figure 4. There is a dip in WER at two points of Gamma – 0.4 and 0.5.

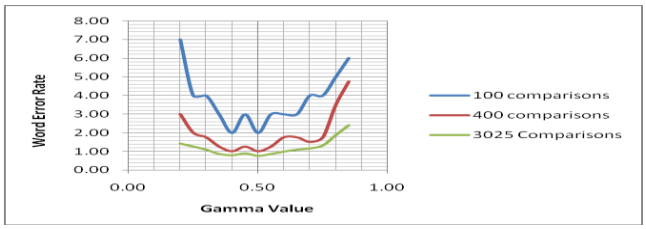

Figure 3. Graphical representation of the results

The results of test cases 1, 2 and 3 with 100, 400 and 3025 pairs of comparisons respectively are shown in Figure 4. The WER is decreasing with the increase in the number of word comparisons.

VI. Summary And Conclusions

The critical distance criterion is developed to differentiate accent variations from the new words. A given sequence of phonemes can be classified as an accent variation or an OOV word with 99.69% accuracy. Confidence interval tests validated the results at 1 per cent level of significance. The above technique is used in the Adaptation of ASR systems to "Everyday Speech". Data sets are selected from daily newspapers and TIMIT databases. The experimental results showed there is an improvement by 13.3% when adaptation of pronunciations and error protection are applied to the verbal command and control systems.

VII. Future Enhancements

The cost matrix used to find the phonetic distance is based on the feature set of the respective phonemes. Further variations in the algorithm may be used and the performance of the MMI model may be observed.
Different domains of the environment may be experimented to estimate the optimum size of the vocabulary in the memory

Acknowledgments

We acknowledge the contributions of R&D laboratory programmers who helped us in coding and experimenting with various datasets.

References

[1] 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.

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

[3] Gopala Krishna Anumanchipalli, Mosur Ravishankar and Raj Reddy, 2007. Improving Pronunciation Inference using N-Best list, Acoustics and Orthography. In Proceedings of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, USA, 2007.

[4] Rabiner, L., Juang, B. and Yegnanarayana B. 2009. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, N.J.

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

[6] Anand Venkataraman. 2001. A Statistical Model for Word Discovery in Transcribed Speech. Association for Computational Linguistics, Volume 27, Number 3, pp 351 -372.

[7] Ben Hixon, Eric Schneider, Susan L. Epstein. Phonemic Similarity Metrics to Compare Pronunciation Methods. In INTERSPEECH 2011, 28-31 August 2011, Florence, Italy.

[8] Chomsky, N. A. 1986. Knowledge of Language: Is Nature, Origin, and Use. Praeger. New York, NY, 1986.

[9] Holter, T. and Svendsen, T. 1999. Maximum likelihood modelling of pronunciation variation. In Speech Communication. vol. 29, no. 2-4, pp. 177–191.

[10] Huang, Acero, Hon. 2001. Spoken Language Processing Guide to Algorithms and System Development. PH.

[11] Wai C. Chu, (2003), Speech Coding Algorithms, Wiley Interscience.

[12] John S. Collura, Diane F. Brandt, Douglas J. Rahikka (2002), The 1.2Kbps/2.4Kbps MELP

[13] Speech Coding Suite with Integrated Noise Pre-Processing, National Security Agency.

[14] William Stallings, (2009), Network Security Essentials Applications and Standards, Pearson Education

[15] Lann M Supplee, Ronald P Cohn, John S. Collura from US DOD and Alan V McCree from Corporate R&D , TI, Dallas, MELP: The New Federal Standard at 2400 bps.

[16] PENG Tan,CUI Huijuan,TANG Kun (2010); Speech coding and transmission algorithm based on multi folded barrel shifting majority judgment; Journal of Tsinghua University (Science and Technology)

[17] JI Zhe,LI Ye,CUI Huijuan,TANG Kun (2009); Leaping frame detection and processing with a 2.4 kb/s SELP vocoder; Journal of Tsinghua University (Science and Technology