ISSN 2348 – 9928
IJAICT Volume 4, Issue 7, July 2017 Doi:01.0401/ijaict.2014.08.05 Published on 05 (7) 2017

A STUDY ON TRAINING, COMPARISON OF WORDS IN AUTOMATIC SPEECH RECOGNITION USED FOR FLUENCY DISORDER THERAPY – STUTTERING

Ms. M. A. Josephine Sathya
Research Scholar,
Mother Teresa Women's University,
Kodaikanal, Tamil Nadu, India

Dr. E. Chandra
Professor & Director,
Dr.SNS.Rajalakshmi College of Arts and Science,
Coimbatore, Tamil Nadu, India

Abstract — Speech is considered to be a very effective and developed means of communication. Speech disorders or speech impediments are a type of disorders, where 'normal' speech is disturbed. Someone who is unable to speak due to a speech disorder is considered mute. All children seem slow in the early stages of learning language, but some children continued with the some problems. Speech disorders describe children whose speech and language is developing abnormally. Sometimes, however, the message might be affected by physiological or psychological factors, becoming incoherent for the listener. This paper presents the types of disorders, especially focused on the fluency disorders where artificial recognition and disfluency identification are considered to be complicated and complex, and also discussed about the training methods used in the pattern – matching approach to automatic speech recognition which can be used for fluency disorder therapy especially for the stuttering people.

Keywords— Fluency disorder, Speech, Speech disorders, Stuttering; Training, Word detection, Word Comparison.

I. INTRODUCTION

A communication disorder is the destruction in the ability of receiving, sending, processing, and to realize the concepts or spoken, unspoken and graphic symbol systems. A communication disorder may be obvious in the processes of investigation, verbal communication, and/or speech. A communication disorder may range in harshness from mild to deep. It may be developmental or acquired. Individuals may demonstrate one or any combination of communication disorders. A communication disorder may result in a primary disability or it may be secondary to other disabilities [9].

1.1 Types of disorders

A speech disorder is the destruction of the articulation of speech sounds, fluency and/or voice.

- An articulation disorder is the typical production of speech sounds characterized by substituting, omitting, adding or distortions that may get in the way with clearness.
- A fluency disorder is a disruption in the flow of speaking characterized by a typical rate, rhythm, and repetitions in sounds, rules, terms, and phrases. This may be exposed by excessive tension, abnormal behavior, and secondary characteristics [6].
- A voice disorder is characterized by the abnormal production and/or absences of spoken quality, pitch, intensity, quality, and/or duration, which is unsuitable for an individual's age and/or sex [7].

1.2 Challenges

At an abstract level the Speech or communication disorders are facing some below challenges which is taken for further research.

- Issues in fluency disorders (stuttering) [12].
- Identification of Stuttering [7].
- Training Methods [10].
- Comparison for word Detection [18], [17], [5], [16], [14].

II. FLUENCY DISORDER - STUTTERING

Stuttering is considered a neurological speech disorder that affects the speaker's ability to physically produce smooth speech. The specified interruptions, repetitions, pauses/blocks, prolongations, and interjections was made the flow of speech to be episodic or disfluent. Sometimes the term stuttering is used interchangeably with stammering. As per the explanation from many people who stutter was that they know what they want to say, but are unable to produce it in a smooth manner. This results in great disappointment and embarrassment on the part of the speaker. There is a strong genetic correlation of adults who stutter, approximately 60 to 70% of them have a family history of stuttering. In addition, children with other speech and language problems or developmental delays are

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.08.05 Published on 05 (7) 2017

more likely to stutter. Males are four times more likely than females to stutter. The belief of some people is that stuttering can be caused by a psychologically or neurologically traumatic event, but this is a controversy in the profession. If this were going to be the case then treatment would fall under the expertise of mental health experts rather than a speech therapist.

During stuttering, the absolute mean onsets of young stutterers' various speech production events were typically earlier than those during normal's' fluent utterances; however, the relative temporal sequence of these same events during stuttering was comparable to that associated with normally fluent children's fluent productions. These findings suggest that young stutterers are grossly within normal limits with regard to selected temporal aspects of coordination for speech production, a finding in contrast to previous reports of adult stutterers' apparent difficulties in coordinating multiple components for speech production [2].

2.1 Stuttering Identification - Hierarchical ANN System

This system is mainly applied for stuttering identification [7]. This paper covers the issue of applying neural networks to the recognition and categorization of non – fluent and fluent utterance records. Three types of stuttering episodes were applied. They are

- Blocks before words starting with plosives,
- Syllable repetitions, and
- Sound – initial Prolongations.

It built with hierarchical neural network framework, was used and then evaluated with respect to its ability to recognize and categorize disfluency types in stuttered speech to reduce the dimension of vector describing the input signals was the main purpose of the first network. After the analysis, the output matrix which consist of neurons winning in a particular time frame. This output matrix was taken as an input for the next network. Various types of MLP networks were examined with respect to their capability to categorize utterances correctly into two non – fluent and fluent groups. The results were accomplished and classified correctness exceeded 84 – 100% [15] depending on the disfluency type.

2.2 Training Methods

One of the most fundamental modules in the pattern – matching approach to automatic speech recognition is the method for construction of the reference pattern. This is so-called training problem [10]. Some of the training methods are:

- Casual Training,
- Robust Training,
- Discriminative Training, and
- Noise Adaptive Training.

2.3 Casual Training

A simple template training procedure that is used often is to use each token during the training session, as a reference pattern ,usually each utterance class is represented by a multiplicity of spoken tokens , hence a multiplied of reference is often called casual training [10].

For a simple vocabulary, a speaker must be able to produce a consistent set of reference patterns for the system to be useful. However since the method makes no attempt to estimate or predict the pattern variability, it could easily fail in systems in which the words in the vocabulary are sometimes confusable. Furthermore, because of the desire to maintain its inherent simplicity, errors committed in training, such as improper articulation, mispronunciation, mishandling of handset or microphone, acoustic noise, or other problems are often accepted as valid reference patterns without any possibility of correction. This obviously leads to poor performance under some conditions.

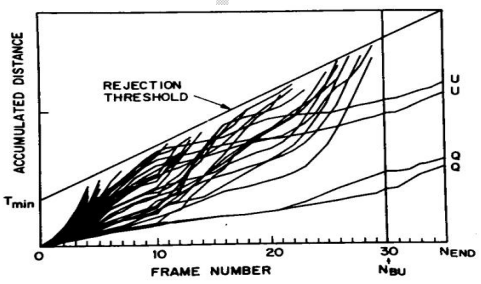

Fig 1 : Accumulated DTW distortions score versus test frame based on casual training with two reference patterns per word.

As an example of the use of casual training in a recognition system, Fig.1. Shows a plot of accumulated Dynamic Time Wrapping (DTW) distortion scores versus test frame for a recognition system with a 39 – word vocabulary (26 letters of the alphabet, 10 digits, 3 command words) based on using two casually trained templates per word [8].

The actual spoken word (the test utterance) was the word "Q". For speed of implementation, a rejection threshold curve on accumulated DTW path was calculated. Both casual templates of the spoken word "Q" provided the two lowest accumulated scores as anticipated. However it can also be seen that the reference templates for the word "U", which is phonetically (and acoustically) quite similar to the word "Q accumulated DTW scores that were the closest of all other words to the spoken "Q". For the last two – thirds of the frames of the DTW match, the slope of the accumulated distance curves for both the "Q". templates and the "U" templates are quite similar; indicating the region of difference is at the beginning of the word.

2.4 Robust Training

It is a sequential training method in which each utterance class is spoken multiply often until a consistent pair of tokens is obtained. The resulting reference pattern is calculating as the average (along the Dynamic Time Wrapping (DTW) path) of the pair of consistent tokens. The averaging of token is normally performed in the spectral domain [10].

The training procedure works as follows. We consider only training for a particular utterance class. Let $X_1 = (x_{11}, x_{12}, x_{13}\ldots x_{1T})$ be the first spoken token. When another process, resulting in a DTW distortion scored(X_1, X_2) is smaller than a prescribed threshold, say €, the pair of tokens are considered to be consistent. The dynamic Time Warping problem is formulated for a pair of Patterns (X_1, X_2), as minimization of the accumulated distortion.

$$d(X_1, X_2) \triangleq d_\phi(X_1, X_2) = \min d_\phi'(X_1, X_2) \quad (1)$$

Where

$$d_\phi'(X_1, X_2) \triangleq \sum_{k=1}^{T_y} d(\phi'_1(k), \phi'_2(k)) m(k)/M_{\phi'} \quad (2)$$

is based on a spectral distortion measure and a set of warping functions.

$$i_1 = \phi'_1(k) \text{ and } i_2 = \phi'_2(k) \quad (3)$$

The pair $\phi = (\phi_1, \phi_2)$ is the optimal warping function for X1, X_2. The index k is the "normal" time. Based on the optimal path

$$y_k = \left(X_{1\phi_1(k)} + X_{2\phi_2(k)}\right), \qquad K = 1,2,\ldots,T_y \quad (4)$$

T_y is the maximum of the normal time k for the optimal warping path. The vector x and y in the above are as usual, the short time spectra or spectral models.

For a speaker trained system, usually about 60% of the vocabulary words require two tokens, 35% of the words require three or four tokens, and only 5% of the words would need more than five tokens for a consistent pair of tokens to be obtained. The reference pattern so generated is usually robust and relatively insensitive to botched tokens [10].

2.5 Discriminative Training

Discriminative training attempts to optimize the correctness of a model by formulating an objective function that in some way penalizes parameter sets that are liable to confuse correct and incorrect answers. The training data from which speaker independent recognizers are trained is usually collected from a wide range of speakers. The inter-speaker variability is a possible source of recognizer error. In Speaker Adaptive Training (SAT) [1], these differences are minimized by the application of transforms calculated on the training speakers. In Discriminative Speaker Adaptive Training (DSAT), similar to the speaker adaptation case, DLTs are computed for the normalization of the training set speakers. Results in [19] show that using DSAT and MMI-trained models provides a 0.6% absolute increase over normal MMI on conversational telephone speech. Using MPE, a larger increase of 0.8% absolute is attained [11].

2.6 Noise Adaptive Training

In traditional methods for noise robust automatic speech recognition, the acoustic models are typically trained using clean speech or using multi-condition data that is processed by the same feature enhancement algorithm expected to be used in decoding. The author proposed [13] a Noise Adaptive Training (NAT) algorithm that can be applied to all training data that normalizes the environmental distortion as part of the model training. In contrast to feature enhancement methods, NAT estimates the underlying "pseudo-clean" model parameters directly without relying on point estimates of the clean speech features as an intermediate step. The pseudo-clean model parameters learned with NAT are later used with Vector Taylor Series (VTS) model adaptation for decoding noisy utterances at test time. Experiments performed on the Aurora 2 and Aurora 3 tasks demonstrate that the proposed NAT method obtain relative improvements of 18.83% and 32.02%, respectively, over VTS model adaptation[13].

2.7 Dynamic Comparison for word Detection

In a system of speech recognition containing words, the recognition requires the comparison between the entry signal of the word and the various words of the dictionary. The problem can be solved efficiently by a dynamic comparison algorithm whose goal is to put in optimal correspondence the temporal scales of the two words. An algorithm of this type is Dynamic Time Warping. The paper presents two alternatives for implementation of the algorithm designed for recognition of the isolated words [9]. Today's vocal recognition systems are based on the general principles of forms recognition [5]. The methods and algorithms that have been used so far can be divided into four large classes:

- Discriminant Analysis Methods based on Bayesian discrimination,
- Hidden Markov Models,
- Dynamic Programming – Dynamic Time Warping algorithm (DTW) [16], and
- Neural Networks.

2.8 Using DTW Algorithm in Speech Recognition

Vocal Signal Analysis, sound travels through the environment as a longitudinal wave with a speed that depends on the environment density. The easiest way to represent sounds is a

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.08.05 Published on 05 (7) 2017

sinusoidal graphic. The graphic represents variation of air pressure depending on time [5]. The shape of the sound wave depends on three factors:

- *amplitude* ,
- *frequency, and*
- *phase.*

The amplitude is the displacement of the sinusoidal graph above and below temporal axis ($y = 0$) and it corresponds to the energy the sound wave is loaded with. Amplitude measurement can be performed using pressure units' decibels (DB), which measure the amplitude following a logarithmic function as regards a standard sound. Measuring amplitude using decibels is important in practice because it is a direct representation of how the sound volume is perceived by people. The frequency is the number of cycles the sinusoid makes every second. A cycle consists of an oscillation starting with the medium line, then it reaches the maximum, then it reaches the minimum and then back to medium line. The frequency is measured in cycles per second or Hertz (Hz). The reverse of frequency is called the period. It is the time needed for the sound wave to complete a cycle. The phase measures the position from the beginning of the sinusoidal curve. The phase cannot be perceived by human senses, but humans can detect the relative phase changes between the two signals. In fact, this is the way human sensorial system perceives a sound location, counting on different phases perceived by the ears [16].

Assume that an incoming speech pattern and a template pattern are to be compared, having n and N frames respectively. Some metric has been used to calculate the distance (i) between frame i of the incoming speech and frame j of the template. The aim was to find a path from (1,1) to (n,N) such that the sum of the total distances between frames is minimum .One approach is to evaluate every possible path between both points, and select the path with the lowest overall distance which only work for very small values of n and N. To solve this problem efficiently, we use a mathematical technique known as Dynamic Programming (DP). In order for DP to be applicable, the problem must exhibit two properties:

- Overlapping sub problems: the problem can be broken down into sub problems whose solution method can be reused over and over
- Optimal substructure: the solution can be obtained by the combination of optimal solutions to its sub problems
- Consider the problem illustrated in the Fig.2, and assume that
- he path always goes forward in time (i.e., has a non-negative slope)
- We cannot skip individual frames from each pattern (i.e., jumps are not allowed)

Consider a point (i , j) in the middle of both patterns . Let (i, j) denote the cumulative distance along the optimum path from 1,1 to (i,j)

$$D(i,j) = \sum_{x=1}^{i,j} d(x,y) \quad (5)$$

$(x,y) \in optimal\ path$ .If (i,) is on the optimal path, then the optimal path must also pass through one of its three neighboring cells: $(i,-1),(i,j-1),(i-1,j-1)$. Therefore, the cumulative distance [14] can be computed as

$$D(i,j)=\min[D(i,j-1),D(i-1,j),D(i-1,j-1)+d(i,j) \quad (6)$$

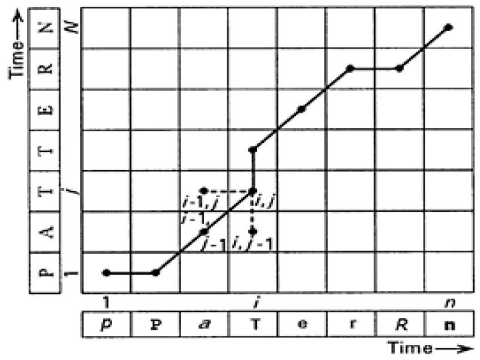

Fig 2 : The path always goes forward in time (i.e., has a non-negative slope), we cannot skip individual frames from each pattern (i.e., jumps are not allowed).

In other words, the best way to get to (i,j) is to get to one of its immediately preceding points by the best way, and then take the appropriate step to (i,j) .Thus, a simple procedure may be used to fill in matrix D .

Initialization: (1,1)=d(1,1)

Cells along the left-hand side can only follow one direction (vertical), so starting with (1,1), values for $D(1,j)$ can be calculated for increasing values of j . Once the left column is completed, the second column can be computed from

$$D(i,j)=\min[Di,j-1,Di-1,j,Di-1,j-1]+d(i,j) \quad (7)$$

and so forth ,the value obtained for Dn,N is the score for the best way of matching the two words .If you also are interested in finding the optimal path itself, then additional book-keeping is necessary to backtrack from $D(n,m)$ to $D(1,1)$ [14].

2.9 Word Detection

Today's detection techniques can accurately identify the starting and ending point of a spoken word within an audio stream, based on processing signals varying with time. They evaluate the energy and average magnitude in a short time unit, and also calculate the average zero-crossing rate. Establishing

the starting and ending point is a simple problem if the audio recording is performed in ideal conditions. In this case the ratio signal-noise is large because it's easy to determine the location within the stream that contains a valid signal by analyzing the samples. In real conditions things are not so simple; the background noise has a significant intensity and can disturb the isolation process of the word within the stream [16].

2.10Words Identification

Words identification can be performed by straight comparison of the numeric forms of the signals or by signals spectrogram comparison. The comparison process in both cases must compensate for both the different length of the sequences and non-linear nature of the sound. The DTW Algorithm succeeds in sorting out these problems by finding the warp path corresponding to the optimal distances between two series of different lengths [16].

III. DISCUSSION

Automatic Speech Recognition (ASR) is the process by which a machine is able to recognize and act upon spoken language or utterances. One of the most important aims of speech recognition systems is to maintain human – computer communication through voice communication which is used by the users widely. Speaking is the most natural aspect of human communication. It is important to learn speech recognition systems and to contribute to the improvement of individuals who suffer from speech disorders especially who suffer from fluency disorders like stuttering, stammering. The studies reviewed showed that ASR performance could be improved to a certain extent with increased user and system training. The language and speech delays are treated early and appropriately will improve over time, but it is not a simple process depending on the reason for delay, interventions need to help the child communication, make speech sounds if possible by employing some techniques. However, research findings suggest that the Hidden Markov Model (HMM) method provides a natural and highly reliable approach of recognizing speech for a broad range of applications. Further it is noted that the training methods available for speech recognition. Different training methods which were the major role in the training of data especially for the stuttered and non stuttered were found in this paper.

IV. CONCLUSION

The study of this paper presented the system to support the stuttered speech recognition process with the use of artificial neural networks. The proposed model, based on the hierarchical ANN structure, allows for the identification of the three most common disfluency types. Self-organizing networks are capable of distinguishing the features which are important in disfluency identification, preserving the utterance structure and reproducing silence moments. The networks are able to learn both language structure and idiosyncratic time-dependent relations that became the indicators of disfluency occurrence. The described procedures could be applied to continuous speech without the necessity of using parameterization methods. The obtained results show that the ANN has become an important tool to assist automatic non-fluent speech recognition systems, allowing objective disfluency recognition. Automatic speech pathology recognition is important for proper diagnosis and selection, as well as therapy monitoring. The method of disfluency recognition presented in this paper will constitute a significant part of the computer diagnosis system. The future work was planned to include cognitive analysis techniques adopted for the needs of speech disorder diagnosis into the system. As well as the system to be created by which the training should be given for the stuttered people.

References

[1] Anastasakos. J, McDonough. J, Schwarz. R and Makhoul J, "A Compact Model for Speaker-Adaptive Training", Proceedings ICSLP, pp. 1137-1140, 1996.

[2] Anthony J. Caruso, Edward G. Conture and Raymond H. Colton, "Selected temporal parameters of coordination associated with stuttering in children", Journal of Fluency Disorders, vol.13, issue. 1, pp. 57–8, 1988.

[3] Asgari. M and Shafran I "A Predicting severity of Parkinson's disease from speech". In: 32nd Ann International Conference of the IEEE Engineering in Medicine & Biology Society, Buenos Aires, Argentina, pp. 5201–5204, 2010.

[4] Awad. S.S. Instrumentation and Measurement Technology Conference, 1997. IMTC/97. Proceedings. Sensing, Processing, Networking. IEEE, vol. 2, Pp.1252 - 1256.

[5] Benoit Legrand. C.S. Chang, S.H. Ong, Soek-Ying Neo and Nallasivam Palanisamy, "Chromosome classification using dynamic time warping", Science Direct Pattern Recognition Letters, vol. 29, pp. 215–222, 2008.

[6] http://www.asha.org/policy/RP199300208/#sthash.9TAe8Irr.dpuf.

[7] Izabela Swietlicka, Wiesława Kuniszyk-Jó´zkowiak, and El˙zbieta Smołka, "Hierarchical ANN system for stuttering identification", Elsevier - Computer Speech and Language, Vol. 27, pp.228–242, 2013.

[8] Julian D. Arias-Londono, Juan.I Godino – Llorente Nicolar Saene – Lechon, Victor Osma – Ruiz and Germam castellanos–Domingnee, "Automatic Detection of Pathological Voices using Complexity Measures , Noise Parameters and Mel–Cepstral Coefficients", IEEE Transactions on Biomedical Engineering , vol.58, no.2, Feb 2011.

[9] Kennison and Shelia M, "Introduction to language development", Los Angeles: SAGE. ISBN: 978-1-4129-9606-8, 2014.

[10] Lawrence Rabiner, Biing–Hwang Juang, and B. Yegnanarayana, "Fundamentals of Speech Recognition", pearson Education, ISBN: 9780130151575, Ch.5, Pp.150 – 153,2009.

[11] McDonough. J, Schaaf, and T.A. Waibel, "On Maximum Mutual Information Speaker-Adapted Training", Proceedings of the IEEE, ICASSP02,Vol.1,SP-P08- Pp.601-604, 2002.

[12] Ooi Chia Ai, M. Hariharan, Sazali Yaacob, and Lim Sin Chee "Classification of speech dysfluencies with MFCC and LPCC features", Expert Systems with Applications,vol.39, pp. 2157–2165, 2012.
[13] Ozlem Kalinli, Michael L. Seltzer, Jasha Droppo, and Alex Acero, "Noise Adaptive Training for Robust Automatic Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, Pp. 8, 2010.
[14] Ricardo Gutierrez-Osuna, Introduction to Speech Processing, CSE@TAMU.
[15] Roman Cmejla, Jan Rusz, Petr Bergl, and Jan Vokral, "Bayesian changepoint detection for the automatic assessment of fluency and articulatory disorders", Elsevier - Speech Communication, vol.55, pp. 178–189, 2013.
[16] Sakoe. H and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE, Trans. Acoustics, Speech, and Signal Proc, Vol.26, 1978.
[17] Stan Salvador and Chan, "Fast DTW: Toward Accurate Dynamic Time Warping in Linear Time and Space", IEEE Transactions on Biomedical. Engineering, vol. 43, no. 4, 1978.
[18] Titus Felix and FURTUNĂ, "Dynamic Programming Algorithms in Speech Recognition", Revista Informatica Economicănr. Vol. 2(46), 2008.
[19] Wang, L and Woodland P.C. "Discriminative Adaptation and Adaptive Training", EARS STT Workshop, 2003.