Automatic Speech Segmentation and Recognition using Class-Specific Features

J. Ujwala Rekha
Dept. of Computer Sci. & Engg.
JNTUH College of Engineering
Hyderabad, Telangana, India

K. Shahu Chatrapati
Dept. of Computer Sci. & Engg.
JNTUH College of Engineering
Manthani, Telangana, India

A Vinaya Babu
Dept. of Computer Sci. & Engg.
JNTUH College of Engineering
Hyderabad, Telangana, India

ABSTRACT
The class-specific automatic speech recognition systems construct an individual classifier for each class based on its own feature set, wherein the feature set for each class is selected such that it distinguishes that class from the other classes most accurately. Consequently, different feature set sequences must be fed into each of the classifiers, and the output of each of the classifiers must be combined to predict the actual class of the observation sequences. However, speech is continuous, and to be able to apply class-specific features, speech should be segmented and fed to the classifiers, which requires the identification of segmentation cues. This paper proposes a framework that jointly segments, and combines the output of the class-specific classifiers in the absence of any segmentation cues using a recursive formulation.

Keywords
Class-specific feature set, speech segmentation, speech recognition

1. INTRODUCTION
Generally, the relevant features for recognition are not known in advance, and a high dimensional feature set is incorporated. Unfortunately, the high dimensional feature set can consist of some redundant and irrelevant features leading to over-fitting and low efficacy of the recognition system. Therefore, feature selection can be employed to eliminate the redundant and irrelevant features present in a high dimensional feature set. Traditionally, a single feature subset is obtained for distinguishing each class from the other classes [1]. On the contrary, class-specific feature sets consisting of different feature subsets for each class can also be obtained [2]. Classification by means of class-specific feature sets involves generating individual classifiers for each class using the corresponding feature subset. Conventionally, a Bayesian classifier requires the probability density functions of the feature set to be known for all classes. However, by including a noise-only class, and applying Neyman-Fisher factorization theorem, the conventional Bayesian classification can be reformulated to include multiple feature sets, each specific to a class [3].

To employ class-specific feature sets in continuous speech recognition systems, the speech signal must be segmented and fed to the individual classifiers. Consequently, the accuracy of the recognition system depends partly on the segmentation process. The problem of speech segmentation and recognition reflect the chicken or the egg causality dilemma [4]. Discovering particular section of a speech as a meaningful unit presumes recognition of that unit; on the contrary, the recognition of the unit is possible only after segmentation. While the former approach is called top-down segmentation [5], the latter approach is called the bottom-up segmentation [6]. Generally, top-down and bottom-up approaches are integrated to harness the strengths of both approaches [7], thereby increasing the performance of the system.

From the procedural perspective, the segmentation and classification of a continuous speech can be done either jointly or sequentially. In a joint segmentation and classification approach, the system iteratively and systematically speculates the possible segments and the labels, and reinforces those speculations that fit together appropriately. Thus, it employs the information from both segmentation and classification; and eventually generates appropriate segments and labels of the continuous speech signal [8]. Alternatively, sequential segmentation and recognition approach, generates segments from the acoustic cues independent of the labels, which are then fed to the classifier to identify the labels ([6],[9]).

From the optimization perspective, there are optimal and sub-optimal approaches. The optimal criterion can either be the maximum likelihood or the least squares estimate of the change in the mean of the observed data [10]. Since each segment is assumed to be independent of the other segments, optimal search can be performed by applying dynamic programming rather than an exhaustive search [8]. However, the dynamic programming' implementation of the segmentation and the classification problem is computationally expensive in practice [11], if the number of segments is high. Consequently, sub-optimal approaches based on heuristics [12] and sequential estimation are proposed ([13],[14]). Though, computationally inexpensive, the suboptimal solutions may result in over-segmentation [13].

In this paper, a recursive formulation that computes an optimal solution for the joint segmentation and classification of continuous speech is presented. Initially, a model for joint segmentation and classification as described in [8] is presented in Section 2, then the model for classification using multiple classifiers that uses class-specific feature sets, is presented in Section 3. Subsequently, a recursive formulation for obtaining an optimal solution is provided in Section 4. Finally, the experiments and the results are reported in Section 5, and the conclusions are presented in Section 6.

2. JOINT SEGMENTATION AND RECOGNITION
The joint segmentation and classification model described in [8] is employed in this paper. Let the sample sequence of a speech signal $X = \{x_1, x_2, ..., x_N\}$ contain K segments defined by $K-1$ boundaries as $T = \{t_1, t_2, ..., t_{K-1}\}$ and K labels as $L = \{l_1, l_2, ..., l_K\}$, then each segment i denoted as $X[t_{i-1}, t_i - 1]$ consists of the observations $[x_{t_{i-1}}, ..., x_{t_i} - 1]$,

where $t_o = 1$ and $t_K - 1 = N$. Therefore, a joint segmentation and classification problem constitutes of determining unknown K, T, and L such that the likelihood of sample sequence X is maximized. Assuming that the K segments are statistically independent of each other, the likelihood of the sequence X can be computed as follows

$$P(X) = \prod_{i=1}^{K} P(X[t_{i-1}, t_i - 1] \mid H_{l_i}) \quad (1)$$

where $P(X \mid H_{l_i})$ is an acoustic model defined for label l_i.

3. CLASSIFICATION USING CLASS-SPECIFIC FEATURES

Classification of a sample defined by the feature set X into one of the M classes can be derived from the Bayesian classifier as follows:

$$\arg\max_{j=1}^{M} P(H_j \mid X) = \arg\max_{j=1}^{M} P(X \mid H_j) P(H_j) \quad (2)$$

With no loss of generality, assuming $P(H_j)$ is identical for all j, it can be ignored. Furthermore, let z_j be the class-specific feature set corresponding to the class H_j, then define $Z = \bigcup_{i=1}^{M} z_i$. If $P(Z \mid H_j)$ exists for all j, then classifier based on Z becomes

$$\arg\max_{j=1}^{M} P(H_j \mid Z) = \arg\max_{j=1}^{M} P(Z \mid H_j) \quad (3)$$

Although, class-specific feature sets z_j exist for each of the classes H_j, application of (3) requires that feature sets be joined together into a super-set Z.

According to [3], construction of a speech recognition system using the class-specific feature sets is possible with the inclusion of common null hypothesis and the application of Neyman-Fisher factorization theorem. Let $H_0 \in H_j$ for all $j = 1,2,..., M$ be a common sub-class of all classes, then the parameters of each class must include H_0 as a special case. For instance, H_0 can represent samples of *iid* (independent and identically distributed) Gaussian noise.

According to Neyman-Fisher factorization theorem [3], if x is a random variable whose probability distribution function $f_\theta(x)$ is dependent on θ, then $T(x)$ is a sufficient statistic for estimating parameter θ if and only if, non-negative real functions g and h exist such that f can be factored as follows:

$$f_\theta(x) = h(x) * g_\theta(T(x)) \quad (4)$$

where h does not depend on θ, and g depends on θ, which in turn depends on x only through $T(x)$.

Therefore, let $z_j (j = 0,..., M)$ be chosen such that it is a sufficient statistic in the Neyman-Fisher sense, then, according to Neyman-Fisher factorization theorem, $P(Z \mid H_j)$ can be factored as follows

$$\begin{aligned} P(Z \mid H_j) &= g(T(Z) \mid H_j) h(Z) \\ &= \mathrm{P}(\mathrm{z_j} \mid H_j) h(Z) \end{aligned} \quad (5)$$

The logical corollary of the Neyman-Fisherman theorem is that, any likelihood ratios are constant, if they are expressed in terms of a sufficient statistic. Thus, if z_j is a sufficient statistic, then

$$\frac{P(Z \mid H_j)}{P(Z \mid H_k)} = \frac{P(z_j \mid H_j)}{P(z_j \mid H_k)} \quad (6)$$

Therefore, replacing H_k with H_0 a common null (for instance, noise-only) hypothesis that is a sub-class of all classes, Eqn (3) can be written as follows

$$\arg\max_{j=1}^{M} P(H_j \mid Z) = P(Z \mid H_0) \arg\max_{j=1}^{M} \frac{P(z_j \mid H_j)}{P(z_j \mid H_0)} \quad (7)$$

It can be seen that $P(Z \mid H_0)$ is not dependent on j, and can therefore be ignored. Consequently, Eqn (7) can be reduced as follows:

$$\arg\max_{j=1}^{M} P(H_j \mid Z) = \arg\max_{j=1}^{M} P(Z \mid H_j) = \arg\max_{j=1}^{M} \frac{P(z_j \mid H_j)}{P(z_j \mid H_0)} \quad (8)$$

4. OPTIMAL JOINT SEGMENTATION AND RECOGNITION

For the classifier based on class-specific feature sets z_j, the likelihood of a given sample sequence X according to (1), can be rewritten as follows:

$$\begin{aligned} P(X) &= \prod_{i=1}^{K} P(Z[t_{i-1}, t_i - 1] \mid H_{l_i}) \\ &\prod_{\mathrm{i}=1}^{\mathrm{K}} \frac{\mathrm{P}(\mathrm{z}_{l_i}[t_{i-1}, t_i - 1] \mid H_{l_i})}{P(z_{l_i}[t_{i-1}, t_i - 1] \mid H_0)} \end{aligned} \quad (9)$$

Consequently, joint segmentation and classification constitutes choosing K, T and L that maximizes (9), which can formally be stated as follows:

$$\{\hat{K}, \hat{T}, \hat{L}\} = \arg\max_{T,K,L} P(X) \quad (10)$$

Similarly, the maximum likelihood of a sequence X is the maximizer of (9) defined as follows:

$$\hat{P}(X) = \max_{T,K,L} P(X) \quad (11)$$

Let us denote the maximum likelihood of a subsequence $X[p:r]$ containing samples $x_p,..., x_r$ as $\Delta(X[p:r])$, then it can be represented as follows:

$$\Delta(X[p:r]) = \hat{P}(X[p:r]) = \max_{T',K',L'} P(X[p:r]) \quad (12)$$

Furthermore, $\Delta(X[p:r])$ can be calculated recursively as follows:

$$\Delta(X[p:r]) = \begin{cases} \Delta(X[p:r]) & if\ L_{\min} \le r\text{-}p+1 \le L_{\max} \\ \max_{q} \Delta(X[p:q]) + \Delta(X[q+1,r]) & otherwise \end{cases} \quad (13)$$

where q is restricted such that

$$(p + (L_{\min} - 1)) \le q \le (r - L_{\min}) \quad (14)$$

Consequently, the length of the segments is also restricted such that

$$L_{\min} \le (t_i - t_{i-1}) \le L_{\max} \quad (15)$$

The solution for the original problem is $\Delta(x[1:N]$ and the recursive algorithm for obtaining optimal segmentation and classification is given in Algorithm 1.

5. EXPERIMENTS AND RESULTS

A description of the speech corpus employed, baseline system, and the experimental setup will be presented before discussing the results. Furthermore, the process of feature selection for obtaining class-specific feature subsets and the evaluation metrics considered in the experimentation along with the details of the implementation of the proposed framework are described in this section.

5.1 Speech Corpus

TIMIT speech corpus is used in the experiments to evaluate the efficacy of the proposed framework against the baseline system. TIMIT [15] is an acoustic-phonetic database that contains manually-labeled and segmented data. In the experiments, phonemes are considered as one segmentation unit and the 61 original TIMIT phonemes are replaced with a smaller set consisting of 39 phonemes as in [16].

5.2 Evaluation Metrics

The metrics defined in [17] are employed to evaluate the efficacy of the proposed framework against the baseline system and are listed below.

i. Correct Detection Rate: It is defined as the fraction of the correct boundaries detected, which is calculated as follows:

$$CDR = \left(\frac{\text{Total Number of Correct Boundaries Detected}}{\text{Total Number of True Boundaries}} \right) \times 100 \quad (16)$$

ii. Miss Rate: It indicates the fraction of true boundaries not detected and is calculated as follows:

$$MR = 1 - CDR \quad (17)$$

iii. Over-Segmentation: It specifies the segments hypothesized in contrast to the actual number of segments and is given as follows:

$$OS = \left(\frac{\text{Total Number of Boundaries Found}}{\text{Total Number of True Boundaries}} - 1 \right) \times 100 \quad (18)$$

iv. False Alarm Rate: It is a measure that indicates the fraction of boundaries that are not correctly detected and is calculated as follows

$$FAR = \left(1 - \frac{\text{Total Number of True Boundaries Detected}}{\text{Total Number of Boundaries Detected}} \right) \times 100 \quad (19)$$

5.3 Baseline System

Speech is sampled using frames of size *25ms* at the rate of *10ms*, and parameterized into 12-Mel Frequency Cepstral Coefficients (MFCCs). A context-independent five-state left-to-right Hidden Markov Model (HMM) is generated from the TIMIT training set for each of the phonemes. Continuous speech segmentation and recognition are done by connecting the HMMs corresponding to each phoneme in a sequence and running the Viterbi algorithm. To analyze the system in the presence of known and unknown data, it is tested on both TIMIT training set and test set.

5.4 Proposed Framework for Segmentation and Recognition using Class-Specific Features

In the proposed framework for segmentation and recognition using class-specific features, feature subsets that distinguish a phoneme from all other phonemes are to be determined. Therefore, we describe the feature selection process, before discussing about the implementation of the proposed framework.

Algorithm 1:Recursive algorithm for generating optimal segmentation and label sequence

function: $\Delta(X[p:r])$

Input $\{x_p, x_{p+1}, ..., x_r\}$

Output Number of Segments $\hat{K}$

Transition Time $\hat{T}$

Labels of the Segments $\hat{L}$

if $L_{\min} \le r\text{-}p+1 \le L_{\max}$

{

$\Delta(X[p:r]) = \max_j \frac{P(z_j[x_p,..,x_r] \mid H_j)}{P(z_j[x_p,..,x_r] \mid H_0)}$;

$\hat{L}[p:r] = \arg\max_j \frac{P(z_j[x_p,..,x_r] \mid H_j)}{P(z_j[x_p,..,x_r] \mid H_0)}$;

$\hat{K}[p:r] = 1$;

if $x_r \ne N$ **then** $\hat{T}[p:r] = \{x_p, x_r\}$;

else $\hat{T}[p:r] = \{x_p\}$;

}

else

{ **for** $q = (p + L_{\min} - 1)$ **to** $(r - L_{\min})$

{ $c[p,q] = \Delta(x[p:q])$; $c[q+1,r] = \Delta(x[q+1,r]$;

}

$\Delta(x[p:r] = \max_q \{c[p,q] + c[q+1,r]\}$;

$q' = \arg\max_q \{c[p,q] + c[q+1,r]\}$;

$\hat{T}[p:r] = \hat{T}[p,q']) \cup \hat{T}[q'+1,r]$;

$\hat{L}[p:r] = \hat{L}[p,q']) \cup \hat{L}[q'+1,r]$;

$\hat{K}[p:r] = \hat{K}[p,q']) + \hat{K}[q'+1,r]$;

}

5.4.1. Feature Selection

Let Z represent a feature vector and $z_m \subseteq Z$ a feature subset; M a set of classes, and $m \in M$ a particular class; then $P(z_m \mid H_m)$ is an acoustic model obtained using the feature set z_m that distinguishes class m from all other classes; and $acc_m(z_m)$ is the accuracy of the classifier, which is defined as the number of times m is accurately classified divided by the total number of occurrences of m in the given data set. Therefore, feature subset z_m must be chosen such that it is the maximizer of

$$\arg \max_{z_m \in 2^z} acc_m(z_m) \qquad (20)$$

However, in order to determine the subset z_m that maximizes (20), one needs to run experiments on all subsets of the feature set, which is not practical. Hence, feature selection algorithm of ([18],[19]) is used in the experiments to generate an acoustic model $P(z_m \mid H_m)$ for all $m \in M$.

5.4.2. Implementation

Similar to the baseline system, speech is sampled at the rate of 10ms using frames of size 25ms, and parameterized into 12 MFCCs. Individually, for each phoneme the MFCCs that arerelevant in distinguishing that phoneme from other phonemes are obtained using feature selection. Later, elemental classifiers, one for each phoneme are generated from the corresponding feature subset using a five-state context-independent left-to-right HMMs. The TIMIT training set is employed for both feature selection and training the HMMs. Then, the input patterns $X = \{x_1, x_2,..., x_N\}$ belonging to both TIMIT training set and test set are run through the recursive algorithm presented in Algorithm 1. The output is the number of segments K, the transition times of the segments T, and the labels of each of the segments L. The average minimum phoneme length of 17ms and the average maximum phoneme length of 185ms is reported in [20] regarding the TIMIT corpus. Therefore, in our experiments, we set $L_{\min} = 1\,frame$ and $L_{\max} = 30\,frames$.

Table 1. Correct Detection Rates of Viterbi Algorithm and the Recursive Algorithm on the TIMIT Training and Test Sets

Algorithm/Data Set/Feature Set	Distance between true boundary & boundary detected				
	≤ 30ms (in %)	30ms-50ms (in %)	50ms-80ms (in %)	>80ms (in %)	Total CDR
Viterbi Algorithm/Training Set/Single Feature Set	5.4	11.2	16.3	12.03	44.93
Recursive Algorithm /Training Set/Class-Specific Feature Sets	11.1	24.6	19.5	10.03	65.23
Viterbi Algorithm/Test Set/Single Feature Set	4.5	9.4	15.6	11.74	41.24
Recursive Algorithm/Test Set/Class-Specific Feature Sets	8.7	21.7	21.2	12.15	63.75

Table 2. False Alarm Rates and Miss Rates of Viterbi Algorithm and the Recursive Algorithm on the TIMIT Training and Test Sets

Algorithm/Dataset/Feature Set	False Alarm Rate (in %)	Miss Rate (in %)
Viterbi Algorithm/Training Set/Single Feature Set	57	55.07
Recursive Algorithm /Training Set/Class-Specific Feature Sets	39.75	34.77
Viterbi Algorithm/Test Set/Single Feature Set	55	58.76
Recursive Algorithm/Test Set/Class-Specific Feature Sets	41.24	36.25

Table 3. Over Segmentation Rates of Viterbi Algorithm and the Recursive Algorithm on TIMIT Training and Test Set

Algorithm/Dataset/Feature Set	Over Segmentation Rate (in %)
Viterbi Algorithm/Training Set/Single Feature Set	2
Recursive Algorithm /Training Set/Class-Specific Feature Sets	2
Viterbi Algorithm/Test Set/Single Feature Set	11
Recursive Algorithm/Test Set/Class-Specific Feature Sets	5

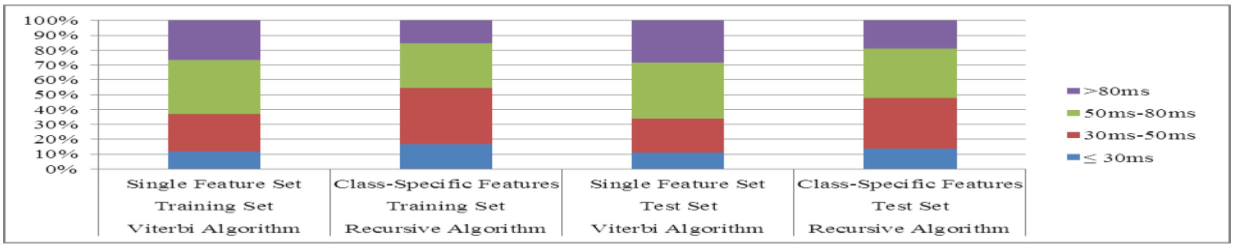

Figure 1. Percentage of Correct Boundaries Detected with Boundary Errors ≤30ms, 30ms-50ms, 50ms-80ms, >80ms

5.5. Results

The transcriptions obtained using the baseline system and the proposed framework, are compared against the hand-labeled TIMIT transcriptions. Generally, the boundaries obtained may not match exactly with hand-labeled boundaries, therefore, a boundary is assumed to be correct, if the detected and the true boundaries are t units apart from each other and contain the same label. The metrics are obtained when true and detected boundaries are within *30ms*, *50ms*, and *80ms* from each other. The quantities for the metrics CDR, FAR and OS are presented in Table 1, Table 2, and Table 3, respectively.

It can be seen that, the proposed framework outperformed the Viterbi algorithm that uses a single feature set in detecting the boundaries and the labels more accurately. On an average it increased the accuracy of the system by more than 20% without over segmenting the speech. Furthermore, it can be observed from the Figure 1 that the error in detecting boundaries is less with the proposed framework rather than the baseline system.

6. CONCLUSIONS

The traditional M-ary classifier is replaced with M binary classifiers that use class-specific feature sets, and a recursive formulation for obtaining optimal segmentation is presented. Experiments are run on the TIMIT corpus for evaluating the baseline system that uses a single feature set against the proposed system. The proposed system outperformed the baseline system in accurately detecting the boundaries as well as labels of the speech signal. In general, M classifiers increase the recognition cost by a factor of M. The methods must be explored to realize improvement in performance without evaluating every classifier for every segment of observed data.

7. REFERENCES

[1] Kalamani, M.; Valarmathy, S.; Poonkuzhali, C.; Catherine, J.N., "Feature selection algorithms for automatic speech recognition," Computer Communication and Informatics (ICCCI), 2014 International Conference on , vol., no., pp.1,7, 3-5 Jan. 2014.

[2] Altun, Halis, and Gökhan Polat. "Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection." Expert Systems with Applications 36.4 (2009): 8197-8203.

[3] Baggenstoss, Paul M. "Class-specific feature sets in classification." Intelligent Control (ISIC), 1998. Held jointly with IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA), Intelligent Systems and Semiotics (ISAS), Proceedings. IEEE, 1998.

[4] Cairns, Paul, et al. "Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation." Cognitive Psychology 33.2 (1997): 111-153.

[5] Bozonnet, Simon, Nicholas WD Evans, and Corinne Fredouille. "The LIA-EURECOM RT'09 speaker diarization system: enhancements in speaker modelling and cluster purification." Acoustics, Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010.

[6] Juneja, Amit, and Carol Espy-Wilson. "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines." Neural Networks, 2003. Proceedings of the International Joint Conference on. Vol. 1. IEEE, 2003.

[7] Matsunaga, Sho-Ichi, and Kiyohiro Shikano. "Speech recognition based on top-down and bottom-up phoneme recognition." Systems and Computers in Japan 17.7 (1986): 95-106.

[8] Wang, Z. Jane, and Peter Willett. "Joint segmentation and classification of time series using class-specific features." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 34.2 (2004): 1056-1067.

[9] Siegler, Matthew A., et al. "Automatic segmentation, classification and clustering of broadcast news audio." Proc. DARPA speech recognition workshop. Vol. 1997. 1997.

[10] Bridle, J., and N. Sedgwick. "A method for segmenting acoustic patterns, with applications to automatic speech recognition." Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'77.. Vol. 2. IEEE, 1977.

[11] Martens, Jean-Pierre, and Lieven Depuydt. "Broad phonetic classification and segmentation of continuous speech by means of neural networks and dynamic programming." Speech communication 10.1 (1991): 81-90.

[12] Zimmermann, Matthias, et al. "A* based joint segmentation and classification of dialog acts in multiparty meetings." Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on. IEEE, 2005.

[13] Andre-Obrecht, Regine. "A new statistical approach for the automatic segmentation of continuous speech signals." Acoustics, Speech and Signal Processing, IEEE Transactions on 36.1 (1988): 29-40.

[14] Rekha, J. Ujwala, K. Shahu Chatrapati, and A. Vinaya Babu. "Game theoretic approach for automatic speech segmentation and recognition." Electrical & Electronics Engineers in Israel (IEEEI), 2014 IEEE 28th Convention of. IEEE, 2014.

[15] Lamel, Lori F., Robert H. Kassel, and Stephanie Seneff. "Speech database development: Design and analysis of the acoustic-phonetic corpus." Speech Input/Output Assessment and Speech Databases. 1989.

[16] Lee, K-F., and H-W. Hon. "Speaker-independent phone recognition using hidden Markov models." Acoustics, Speech and Signal Processing, IEEE Transactions on 37.11 (1989): 1641-1648.

[17] Estevan, Y. P., Wan, V., & Scharenborg, O. (2007, April). Finding maximum margin segments in speech. In Proc. ICASSP (Vol. 4).

[18] Rekha, J. Ujwala, K. Shahu Chatrapati, and A. Vinaya Babu. "Feature selection using game theory for phoneme based speech recognition." Contemporary Computing and Informatics (IC3I), 2014 International Conference on. IEEE, 2014.

[19] Rekha, J. Ujwala, K. Shahu Chatrapati, and A. Vinaya Babu. "Feature Selection for Phoneme Recognition Using a Cooperative Game Theory Based Framework." Proceedings of the International Conference on Multimedia, Communication and Computing Application (MCCA 2014).

[20] Kettner, Andreas, and Lothar Thiele. "Speech Features for Optimal Discrimination of Phonemes." (2012).