Web Audio Classification Using Support Vector Machine Based On LPCC Feature

Aranga Arivarasan[1], Dr.M.Karthikeyan [2]
[1]*Research Scholar,* [2]*Associate Professor*
[12]*Division of Computer and Information Science / Annamalai University, India*

Abstract: *Now a days the web document classification has become a very important research area. The growth of the www also makes the web document classification very wide, important as well essential. Our aim is to give focus on web document classification based on audio signals present in the web documents. It is a challenge to extract the most common and salient themes from unstructured raw audio data. In this paper, we propose an effective algorithm to automatically classify audio content into different categories such as news and music. Support vector machines are applied to classify audio into news and music by learning from training data. The linear prediction cepstral coeffieients are extracted to characterize the audio data. The extracted features are used for training using support vector machines. Support Vector Machine shows better performance in audio analyzes than traditional Euclidean distance.*

Keywords: *SVM, LPC, LPCC, Classification*

I. Introduction

Digital audio is one of the most important data types distributed by the Internet. However, it is still difficult for a computer to automatically analyze audio content, especially to automatically classify and recognize audio content. In the age of digital information, audio data has become an important part in many web documents. A typical multimedia database often contains millions of audio clips, including environmental sounds, machine noise, music, animal sounds, speech sounds, and other non-speech utterances. The need to automatically recognize to which class an audio sound belongs makes web document classification and categorization an emerging and important research area. In general, audio categorization can be performed in two steps. In the first step, an audio sound is reduced to a small set of parameters using various feature extraction techniques, and in the second step, classification or categorization algorithms ranging from simple Euclidean distance methods to sophisticated statistical techniques are carried out over these parameters. The efficacy of an audio classification or categorization system depends on the ability to capture proper audio features and to accurately classify each feature set corresponding to its own class.

Our system classifies the audio signals into predefined categories such as news and music. Each and every audio signal extracted from a web document will have its own features. These features are extracted using linear prediction analysis. The linear prediction analysis will produce the linear prediction coefficients (LPC). The LPCC features are obtained from the linear prediction coefficients. The extracted LPCC features are given as input to the support vector machine (SVM). SVM finds support vectors for each category and it is used for classifying the given audio data. The experimental result shows that the proposed method gives an accuracy of about 96% for categorizing the audio signals.

II. Related Works

A number of methods have been proposed to discriminate music, speech, silence, and environmental sound. The most successful achievement in this area is speech/music discrimination, because speech and music are quite different in spectral distribution and temporal change pattern. The work by [1] proposes two approaches of spectral coefficient optimization. The two approaches are (1) optimized based on discrete spectral features and (1) combine spectral features. Experimental studies have been performed through the Berlin Emotional Database, using a support vector machine (SVM) classifier, and five spectral features including MFCC, LPC, LPCC, PLP and RASTA-PLP. The authors in [2] proposes a new method for speaker diarization using support vector machines (SVM) and auto associative neural network (AANN). The speaker diarization process consists of segmenting a conversation speech signal into homogeneous segments which are then clustered into speaker classes. The proposed method uses SVM and AANN models to capture the speaker specific information from Predictive Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC) In [3] average zero-crossing rate and the short time energy were used as features and simple thresholding method was applied to discriminate speech and music from the radio broadcast. The authors in [4] used thirteen features in time, frequency and cepstrum domains and different classification methods to achieve a robust performance. Both approaches reported accuracy rate for real-time classification over 95% if a window size of 2.4s was used. However, the performance will decrease for the above approaches if a small window size is used or other audio scenes such as environment sounds are taken into consideration.

Further research works have been done to segment audio data into more categories. In [5] a method was proposed to classify audio signals into speech, music, and others for the purpose of parsing of news story. In [6] an acoustic segmentation approach was proposed that mainly applied to the segmentation of discussion recordings in meetings. Audio recordings were segmented into speech, silence, laughter and non-speech sounds by using cepstral coefficients as features and a hidden Markov model (HMM) as the classifier. The accuracy rate depended on different types of recording. The authors in [7] proposed an approach to divide the generic audio data segmentation and classification task into two stages. In the first stage, audio signals were segmented and classified into speech, music, song, speech with music background, environmental sound with music background, six types of environmental sound, and silence. In the second stage, further classification was conducted within each basic type.

Speech was differentiated into the voice of man, woman and child. Music is classified into classics, blues, jazz, rock and roll, music with singing and the plain song, according to the instruments or types. Environmental sounds were classified into semantic classes such as applause, bell ring, footstep, wind-storm, laughter, bird’s cry, and so on. The accuracy rate was reported over 90%. Robust two-stage audio classification and segmentation method to segment an audio stream into speech, music, environment sound and silence was proposed in [8]. The first stage of classification was to separate speech from non-speech based on K-nearest-neighbor (KNN) and linear spectral pairs—vector quantization (LSP-VQ) classification scheme and simple features such as zero-crossing rate ratio, short time energy ratio, spectrum flux, and LSP distance. The second stage further segmented the non-speech class into music, environment sounds and silence with a rule-based classification scheme and two new features: noise frame ratio and band periodicity. The total accuracy rate was reported over 96%.

Our system classifies the audio signals into predefined categories such as news and music. Each and every audio signal will have its own features. These features are extracted using linear prediction analysis. The linear prediction analysis will produce the linear prediction coefficients (LPC). The LPCC features are obtained from the linear prediction coefficients. The extracted LPCC features are given as input to the support vector machine (SVM). SVM finds support vectors for each category and it is used for classifying the given audio data. The experimental result shows that the proposed method gives an accuracy of about 96% for categorizing the audio signals.

III. Feature Extraction

Feature selection is important for audio content analysis. As like the text and image we can extract features from audio signals. The selected features should reflect the significant characteristics of different kinds of audio signals extracted from web documents. Speech is defined as human made sound by vocal tract in an idea for communication. Recorded speech and computer-generated sound are also considered speech. We can split up the speech domain into many sub-categories. One way is to identify the spoken language. One can also use the speech to identify whom or what the speaker is, using the emotional content spoken by the speaker and the subject matter of the speech. Further, we can classify speech into different categories like news, sports commentary, and advertisement content so on. The Music can be human made sound by instruments, as well human vocal tract to reflect particular emotion or feeling. The natural sounds like waterfall, birds sound, animal sounds are also considered as music. All the above said audio signals are distant in some way. Thos hidden information is represented as features of audio signals and those features are to be used to train the machine learning algorithm. The exact difference between these different type audio signals is exactly to be extracted will lead to better results.

IV. Linear Prediction Analysis

The theory of linear prediction (LP) is closely linked to modeling of the vocal tract system, and relies upon the fact that a particular speech sample may be predicted by a linear weighted sum of the previous samples. The number of previous samples used for prediction is known as the order of predictions. The weights applied to each of the previous speech samples are known as linear prediction coefficients (LPC). They are calculated so as to minimize the prediction error.

In many applications, Euclidean distance is used as a measure of similarity or dissimilarity between feature vectors. The sharp peaks of the LP spectrum may produce large errors in a similarity test, even for a slight shift in the position of the peaks. Hence, linear prediction coefficients are converted in to cepstral coefficients using a recursive relation. Cepstral coefficients represent the log magnitude spectrum, and the first few coefficients model the smooth envelope of the log spectrum. These coefficients can be obtained either from linear prediction coefficients or from the inverse discrete fourier transform (IDFT) of log magnitude spectrum of the speech signal. In both cases, the process results in estimating vocal tract system characteristics from the speech signal.

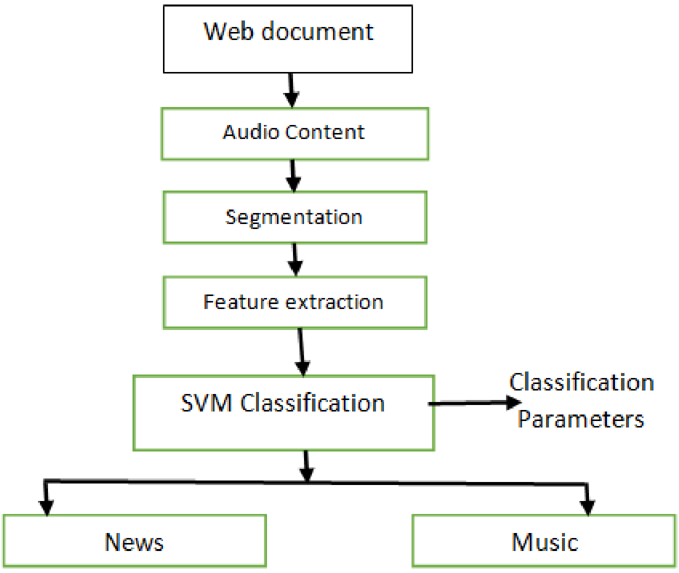


Fig 1 Web Audio Classification

V. Linear Prediction Cepstral Coefficients

The cepstral coefficients derived from either linear prediction (LP) analysis or a filter bank approach are almost treated as standard front end features. Speech systems developed based on these features have achieved a very high level of accuracy, for speech recorded in a clean environment. Basically, spectral features represent phonetic information, as they are derived directly from spectra. The features extracted from spectra, using the energy values of linearly arranged filter banks, equally emphasize the contribution of all frequency components of a speech signal. In this context, LPCCs are used to capture emotion-specific information manifested through vocal tract features. In this work, the 10th order LP analysis has been performed, on the speech signal, to obtain 13 LPCCs per speech frame of 20ms using a frame shift of 10 ms. the human way of emotion recognition depends equally on two factors, namely: its expression by the speaker as well as its perception by a listener. The purpose of using LPCCs is to consider vocal tract characteristics of the speaker, while performing automatic emotion recognition. Cepstrum may be obtained using linear prediction analysis of a speech signal. The basic idea behind linear predictive analysis is that the nth speech sample can be estimated by a linear combination of its previous p samples as shown in the following equation.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + a_3 s(n-3) + \cdots + a_p s(n-p)$$

where a1, a2, a3, . . . are assumed to be constants over a speech analysis frame. These are known as predictor coefficients or linear predictive coefficients. These coefficients are used to predict the speech samples. The difference of actual and predicted speech samples is known as an error. It is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

Where e(n) is the error in prediction, s(n) is the original speech signal, ˆs(n) is a predicted speech signal, aks are the predictor coefficients. To compute a unique set of predictor coefficients, the sum of squared differences between the actual and predicted speech samples has been minimized (error minimization) as shown in the equation below

$$E_n = \sum_m \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2$$

Where m is the number of samples in an analysis frame. To solve the above equation for LP coefficients, En has to be differentiated with respect to each ak and the result is equated to zero as shown below

$$\frac{\partial E_n}{\partial a_k} = 0$$

After finding the aks, one may find cepstral coefficients using the following recursion.

$$c_0 = \log_e p$$

$$C_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} C_k a_{m-k}, \quad for\ 1 < m < p\ and$$

VI. Audio Classification Using Support Vector Machines

A support vector machine (SVM) performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. SVM models are closely related to neural networks. In fact, a SVM model using a sigmoid kernel function is equivalent to a two-layer, feed-forward neural network. Using a kernel function, SVM's are an alternative training method for polynomial, radial basis function and multi-layer perceptron classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training. The goal of SVM modeling is to find the optimal hyperplane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyperplane are the support vectors.

SVM learning is a useful statistic machine learning technique that has been successfully applied in the pattern recognition area. If the data are linearly non-separable but nonlinearly separable, the nonlinear support vector classifier will be applied. The basic idea is to transform input vectors into a high-dimensional feature space using a nonlinear transformation, and then to do a linear separation in feature space. To construct a nonlinear support vector classifier, the inner product is replaced by a kernel function. The SVM has two layers. During the learning process, the first layer selects the basis, (as well as the number), from the given set of bases defined by the kernel the second layer constructs a linear function in this space. This is completely equivalent to constructing the optimal hyperplane in the corresponding feature space. The SVM algorithm can construct a variety of learning machines by use of different kernel functions. Three kinds of kernel functions are usually used. They are as follows.

- Polynomial kernel of degree d

  K(x,y)=((x,y)+1)d
- Radial basis function with Gaussian kernel of width c>0

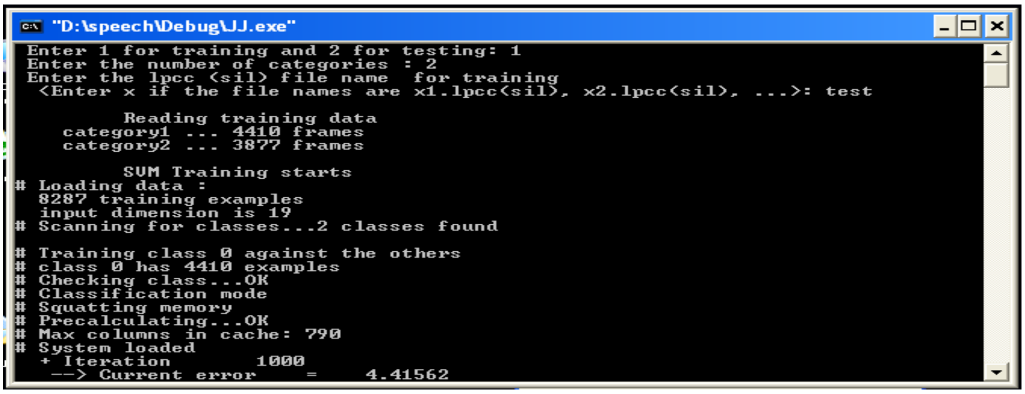
  K(x,y)=exp(-|x-y|2/c)
- Neural networks with tanh activation function

  K(x,y)=tanh(k(x,y)+ □).

Among the several kernel functions available, we chose to use the radial basis functions (RBF). The svm training process analyzes training data to find an optimal way to classify the audio data into sports, news, advertisement and music. The training data should be sufficient to be statistically significant. The training data is segmented into fixed-length and overlapping frames (in our experiments we used 20 msec frames with 10 msec overlapping). When neighboring frames are overlapped, the temporal characteristics of audio content can be taken into consideration in the training process. Features such as LPC and LPCC are calculated from each frame. The support vector machine learning algorithm is applied to produce the classification parameters. The training process needs to be performed only once. The derived classification parameters are used to classify the given audio into sports, news, advertisement and music.

VII. Experimental Results

The experimental studies are conducted for two different categories of audio data. The data has been collected from news web document and music web document. The output will be one among the two categories. The data were randomly divided in to two sets namely training sets and testing sets. For training the LPCC features are extracted for each audio data. The LPCC features are extracted for each category and the SVM is trained in multi-class mode. The class label 0 and 1 corresponds to the category news and music. The SVM training process creates two models, one model for each category. The snapshot of the SVM training is shown in figure 2. For testing, the LPCC features are extracted from the audio data and are given as input to the SVM. The SVM test the category for each audio frame and the majority rule is used to decide the category of the given audio data. The snapshot of the SVM testing is shown in Figure 3. The audio signals used to train and test the SVM with news class is shown in figure 4 and the audio signals used to train and test the SVM with music class is shown in figure 5.



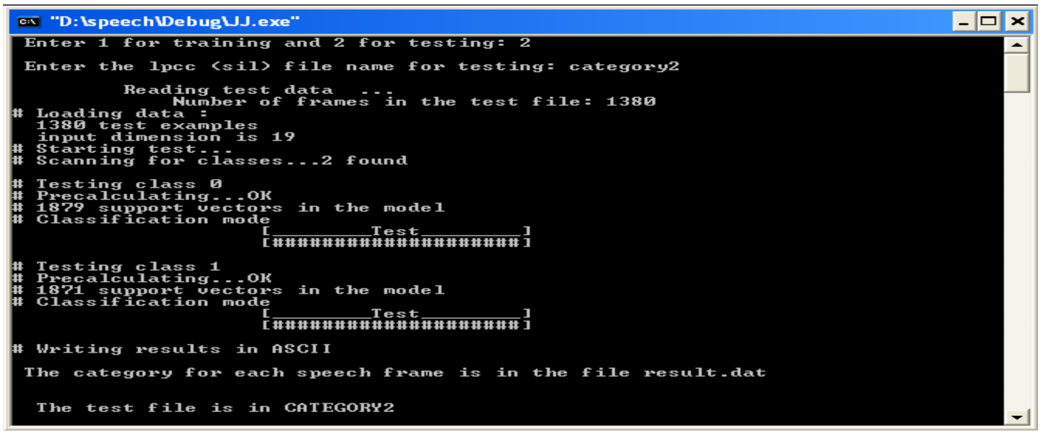
```
 Enter 1 for training and 2 for testing: 1
 Enter the number of categories : 2
 Enter the lpcc <sil> file name  for training
  <Enter x if the file names are x1.lpcc<sil>, x2.lpcc<sil>, ...>: test

         Reading training data
    category1 ... 4410 frames
    category2 ... 3877 frames

         SVM Training starts
# Loading data :
  8287 training examples
  input dimension is 19
# Scanning for classes...2 classes found

# Training class 0 against the others
# class 0 has 4410 examples
# Checking class...OK
# Classification mode
# Squatting memory
# Precalculating...OK
# Max columns in cache: 790
# System loaded
  + Iteration        1000
   --> Current error     =     4.41562
```

Fig 2: SVM training Process



```
 Enter 1 for training and 2 for testing: 2

 Enter the lpcc <sil> file name for testing: category2

         Reading test data   ...
             Number of frames in the test file: 1380
# Loading data :
  1380 test examples
  input dimension is 19
# Starting test...
# Scanning for classes...2 found

# Testing class 0
# Precalculating...OK
# 1879 support vectors in the model
# Classification mode
                   [_________Test__________]
                   [#######################]

# Testing class 1
# Precalculating...OK
# 1871 support vectors in the model
# Classification mode
                   [_________Test__________]
                   [#######################]

# Writing results in ASCII

 The category for each speech frame is in the file result.dat

  The test file is in CATEGORY2
```

Fig 3: SVM testing Process

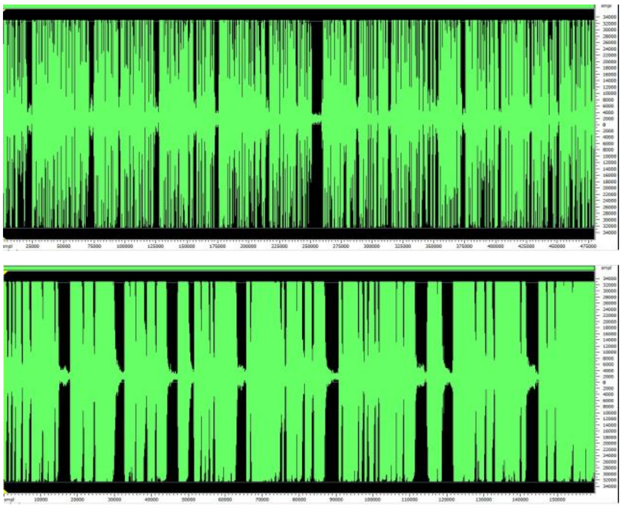
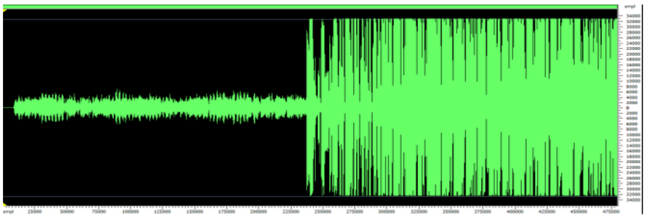


Fig 4. News Train and Test file



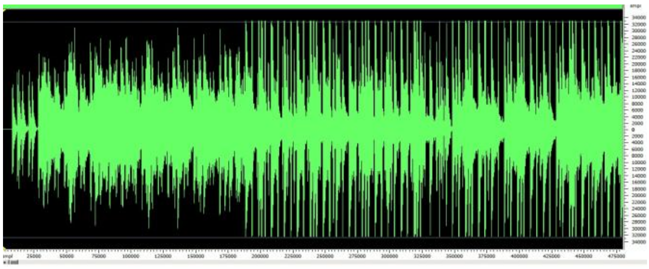


Fig 5. Music Train and Test file

The experimental results show that the method achieves an accuracy of about 96%.

VIII. Conclusion

In this work, we propose a method to classify the two type of different audio signasl news and music taken from the web document. The 14th order LPC and 19 dimensional LPCC features for these audio signals were extracted as features and given as input to train the SVM. The same is applied for testing also. The experimental results show that the proposed method achieves an accuracy of about 96%. In future the category may be extended and also other type of audio signal features were also be considered for trained and tested with SVM and also with other classification machine learning algorithms. The text and images present in the web documents will also be taken and their features can also be used to classify the type of web documents.

References

[1] Idris and Md S. Salam, “Improved Speech Emotion Classification from Spectral Coefficient Optimization”, Advances in Machine Learning and Signal Processing,Lecture Notes in Electrical Engineering , DOI 10.1007/978-3-319-32213-1_22, pp. 387, Springer International Publishing Switzerland 2016.

[2] J. Gladson 1 Maria Britto and 2S. Suresh Kumar, “Speaker diarization using Support Vector Machines and AutoAssociative Neural Network”, Middle-East Journal of Scientific Research, DOI: 10.5829/idosi.mejsr.2016.3858.3868, 24 (12), 3858-3868 2016.

[3] J. Sounders, “Real-time discrimination of broadcast speech/music”, in Proc. ICASSP96, vol. 2, Atlanta, GA, 1996, pp. 993–996.

[4] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature music/speech discriminator”, in Proc. ICASSP97, vol. 2, pp. 1331–1334, 1997.

[5] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal , “Speech/music discrimination for multimedia application”, in Proc. ICASSP00, Istanbul,Turkey, Jun. 2000.

[6] D. Kimber and L. Wilcox, “Acoustic segmentation for audio browsers”, in Proc. Interface Conf., Sydney, Australia, 1996.

[7] T. Zhang and C.-C. Kuo, “Video content parsing based on combined audio and visual information”, in Proc. SPIE 1999, vol. 4, San Jose, CA, pp. 78–89, 1999.

[8] Helge Homburg, Ingo Mierswa, B¨ulent M¨oller, KatharinaMorik and Michael Wurst ,“A Benchmark Dataset For Audio Classification And Clustering”, Queen Mary, University of London, 2005.

[9] Changsheng Xu, Senior Member, IEEE, Namunu C. Maddage, and Xi Shao , “Automatic Music Classification and Summarization”, Institute for InfocommResearch, IEEE, 2005.

[10] Jeffrey P. Woodard, “Modeling And Classification Of Natural Sounds By Product Code Hidden Markov Models”, Autonetics Strategic Systems Division, IEEE,1992.

[11] Sholom M. Weiss et al.: Text Mining Predictive Methods for Analyzing Unstructured Information, ISBN 0-387-95433-3, 2005

[12] N.T. Nguyen et al. (Eds.): “Study for Automatic Classification of Arabic Spoken Documents”., ICCCI 2017, Part II, LNAI 10449, pp. 459–468, 2017. DOI: 10.1007/978-3-319-67077-5_44

[13] Shweta C. Dharmadhikari et al.: “Empirical Studies on Machine Learning Based Text Classification Algorithms”., Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, 2011

[14] Bimbot F, Bonastre J, Fredouille C, Gravier G, Chagnolleau MI, Meignier S, Merlin T, Garcia OJ, Delacretaz P, Reynolds DA (1997) “A tutorial on text independent speaker verification”, EURASIP J Appl Sig Proc 2004(4):430–451

[15] Barigou, F.: “Improving K-nearest neighbor efficiency for text categorization”. Neural Netw. World 26(1), 45 (2016)

[16] Dai, P., et al.: “A novel feature combination approach for spoken document classification with support vector machines”. In: Proceedings of Multimedia Information Retrieval Workshop, pp. 1–5 (2003)

[17] Yu-Chuan Chang, S. Ming Chen, “Multi-Label Text Classification based on a new linear classifier learning method and a category sensitive refinement method”, Science Direct Expert Systems with Applications. Volume 34, Issue 3, April 2008. 1948-1953.

[18] Ooi Chia Ai et al.: “Classification of speech dysfluencies with MFCC and LPCC features”, Expert Systems with Applications 39 2012) 2157–2165.

[19] Cheng Deng et al.: “Multi-Class Support Vector Machine via Maximizing Multi-Class Margins”, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).