Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi
Journal of Applied Sciences of Mehmet Akif Ersoy University

Comparison of Different Normalization Techniques on Speakers' Gender Detection

Serhat Celil İLERİ[1*], Armağan KARABİNA[2], Erdal KILIÇ[3]

[1]Ondokuz Mayıs Üniversitesi, Bilgisayar Mühendisliği Bölümü, orcid id: 0000-0002-0259-0791

[2]Öğr. Gör., Recep Tayyip Erdoğan Üniversitesi, Rektörlük, Enformatik Bölümü, orcid id: 0000-0002-4905-5761

[3]Doç. Dr., Ondokuz Mayıs Üniversitesi, Bilgisayar Mühendisliği Bölümü, orcid id: 0000-0003-1585-0991

Geliş Tarihi/Received: 28.03.2018
Kabul Tarihi/Accepted: 12.04.2018 Araştırma Makalesi/Research Article

ABSTRACT

In this study, the effect of Short-time Mean and Variance Normalization (STMVN), Short-time Cepstral Mean and Scale Normalization (STMSN), Min-Max Normalization, Z-Score Normalization and Standard Deviation Normalization techniques on the classification performance was investigated in determining speakers' gender. In the study, voice records which belongs to 192 male and 192 female speakers from TIMIT data set were used as data set. Features were extracted from Mel Frequency Cepstral Coefficients (MFCC) technique by using voice records and extracted features' dimension was reduced to Principal Component Analysis (PCA), then normalized with different techniques. Support Vector Machine (SVM) was used as classifier. As a result of study, it was observed that, the highest accuracy in speakers' gender estimation is obtained as 98.18% from features which were normalized with Standard Deviation Normalization technique and other normalization techniques were reduced accuracy.

Keywords: Max-Min Normalization, Z-Score Normalization, Standard Deviation Normalization, Short-time Mean and Variance Normalization, Short-time Cepstral Mean and Scale Normalization.

* Sorumlu yazar/Corresponding author
E-mail/e-ileti: celil.ileri@bil.omu.edu.tr

Konuşmacı Cinsiyetinin Tespitinde Değişik Normalizasyon Tekniklerinin Kıyaslanması

ÖZET

Bu çalışmada Kısa-zaman Ortalama ve Değişinti Normalizasyonu (Short-time Mean and Variance Normalization - STMVN), Kısa-zaman Sepstral Ortalama ve Ölçeklendirme Normalizasyonu (Short-time Cepstral Mean and Scale Normalization - STMSN), Asgari – Azami (Min-Max) Normalizasyonu, Z-Skor (Z-Score) Normalizasyonu ve Standart Sapma (Standard Deviation) Normalizasyon tekniklerinin, konuşmacı cinsiyetinin tespitinde sınıflandırma başarımına etkisi araştırılmıştır. Çalışmada veri seti olarak TIMIT veri setindeki 192 erkek ve 192 kadın konuşmacıya ait ses kayıtları kullanılmıştır. Ses kayıtlarından Mel Frekansı Sepstral Katsayısı (Mel Frequency Cepstral Coefficient – MFCC) tekniği ile öznitelik çıkarılmış ve çıkarılan özniteliklerin boyutu Temel Bileşen Analizi (Principal component analysis – PCA) ile indirgenerek, değişik teknikler ile normalize edilmiştir. Sınıflandırıcı olarak Destek Vektör Makinesi (Support Vector Machine – SVM) kullanılmıştır. Çalışma sonucunda konuşmacı cinsiyeti tahmininde en yüksek başarımın %98.18 ile Standart Sapma Normalizasyon Tekniği ile normalize edilmiş özniteliklerden elde edildiği gözlemlenmiş olup diğer tekniklerin başarımı düşürdüğü gözlemlenmiştir.

Anahtar kelimeler: Asgari – Azami Normalizasyonu, Z-Skor Normalizasyonu, Standart Sapma Normalizasyonu, Kısa-zaman Ortalama ve Değişinti Normalizasyonu, Kısa-zaman Sepstral Ortalama ve Ölçeklendirme Normalizasyonu.

1. INTRODUCTION

Speaking, the fundamental form of the communication among people, is transferred physical and emotional information not only to the voice but also to words. (Nabiyev, Yücesoy, 2009) In the previous studies, various physical and mental state estimations were made, including gender (Yücesoy, Nabiyev, 2009), age (Yücesoy, Nabiyev, 2016) and emotional (Durukal, Hocaoğlu, 2015) characteristics.

In paper (Nabiyev, Yücesoy, 2009), the authors aimed to estimate the gender of the speaker independently of the text. The feature vectors were extracted with MFCC by using TIMIT data set.

The obtained feature vectors were classified by VQ method and recorded in database. The test phase including 1680 sound data consisting of 10 sentences, was carried out. The result performance was 1646 correct and 34 incorrect estimation, which means 98.80% success. (Nabiyev, Yücesoy, 2009)

The study (Yücesoy, Nabiyev, 2009) aims automatically determining the sex of an individual from speaker independent voice signals. First, the feature vectors extracted from MFCC later removed the time shift occurring during the distance evaluation of the feature vectors with using DTW. One and multiple word data were recorded in different languages such as Turkish, English, and German. The tests' success rate were 100% with sound that consisting single word phrase and 98% with sound multi-word phrases in different languages. (Yücesoy, Nabiyev, 2009)

Heerden et al. were increased the performance of the proposed method from 45% to 50.7% by using the age regression and gender classifier together. (Heerden et. al. 2010)

In the study performed by Djemili et al. extracted the feature vectors with MFCC, and obtained the feature vectors to Learning Vector Quantization (LVQ), Multilayer Perceptron (MLP), Gaussian Mixture Model (GMM) and Vector Quantization (VQ), and achieved a 96.4% success in estimating gender of the speaker by classifying with four different classifiers. (Djemili et. al. 2012)

Oscal et al. conducted a study with 3564 voice data in 2015. The speakers were divided into 3 groups according to their age and 2 groups according to their gender. Different characteristics were selected for voice data divided into low and high according to stimulation densities. The performance rate, which was 71.6% and 97.2% in gender estimation, increased by 1.7% to 98.9%. (Chen, Gu, 2015)

In the study of Přibil and colleagues, age was determined in four age groups of children, young, adult and elderly from the audio signal using GMM in two levels in Czech and Slovak languages. As a result of this study, 90.4% in children, 95.1% in young men, 92.4% in adult men, 93.2% in elderly men, 100% in young women, 86.7% in adult women and 88.7% in elderly women. Using the GMM as a result of the study, the two-level speaker was convinced that gender / age determination architecture was comparable to standard listening-based methods.(Přibil et. al. 2016)

In Islam's work, Gammatone Frequency Cepstral Coefficient (GFCC) -based robust gender classification method is presented. The text was used to perform better in the gender behavioral model with the Gaussian mixture model (GMM) in the study where an independent speech data was used and only clean signals were used in the training set. The clean sound signal is converted into noisy sound signals ranging from 0dB to 10dB using five different noise samples. In the case of TSTs made with 0dB noise, 100% for females and 92% for males were obtained. According to the result of this study, this method is interpreted as a method which is not affected by noise. (Islam, 2016)

In the study (Khanum, Firos), the success rate in determining speaker gender was investigated not only in text but also in noisy environments such as train station, restaurant, airport. MFCC for attribute extraction, Artifical Neural Network (ANN) for classification, and 6 fuzzy clusters and 10 hidden neurons for optimum results, and a slight increase in noisy ambient noises. (Khaum, Firos, 2017)

In the study of Yusnita and colleagues, we used the Linear Prediction Coefficients (LPC) to derive the sesten feature of 93 speakers and used Artificial Neural Network (ANN) recognition engine through Multi-Layer Perceptron (MLP). Experimental results have yielded an average of up to 93.3%. (Yusnita et al., 2017)

In this study, the feature vectors were extracted using the Mel Frequency Separation Coefficient of the voice data. The dimensions of the feature vectors were reduced to PCA. The classification was performed with and without normalization and the changes in the performance was observed. After reducing of the obtained feature vectors, the effect of the normalized data on the classification performance by using five different normalization methods is investigated. The aim of the study is to obtain a higher success rate in estimating gender by normalizing the reduced data.

In Section 2, the used methods MFCC, STMVN, STMSN, Z-Score Normalization, Min-Max Normalization, Standard Deviation Normalization techniques and materials are explained. In Chapter 3, the effects of the applied normalization techniques on sex ratio, the effect on achievement rate, and the change in gender estimation of male and female speakers were investigated. In the results section, the effects of different normalization techniques on speaker gender determination performance and possible future research topics were investigated.

2. METHOD AND MATERIALS

In this study, voice files of 192 female and 192 male speakers were used from the TIMIT Acoustic-Phonetic Continuous Speech Corpus Dataset. Feature extraction was performed through Mel-frequency Cepstral coefficients of the sound data in the test set obtained by the MFCC method. The feature vectors obtained by MFCC method have different dimensions for each voice file, depending on the length of the audio file, the number of bits per second, and so on.

Since the dimensions of the data used in the classification stage must be equal, hence the feature vectors dimensions were reduced by PCA method to uniform the dimensions. The dimensional reduced feature vectors were prepared to classify. STMVN, STMSN, Z-Score Normalization, Min-Max Normalization and Standard Deviation Normalization techniques were separately applied to the obtained reduced feature vectors. They were recorded and compared.

Finally, the files containing normalized and non-normalized feature vectors were classified by Support Vector Machine. The obtained results were noted and compared.

2.1. Feature Extraction

While estimating speech from speech data, the speech signals are firstly converted to parametric values which give information about speech and speaker characteristics. It has lower discriminative properties, and less variables. (Yücesoy, Nabiyev, 2014) Feature extraction can be performed to use different methods. One of them is MFCC used in this study.

2.1.1. MFCC

The Mel in the MFCC method is a unit of measure generated by the difference of the frequency of a voice tone detected by the human ear with the actual frequency. The reason why the physical frequency and its value that perceived by the human ear is that the perceived sound waves of the human are not linear. The relationship between linear frequency and Mel frequency is given by (1) (Yücesoy, Nabiyev, 2014).

$$F_{mel} = 2595\, log_{10}\left(1 + \frac{\mathrm{f}}{700}\right) \tag{1}$$

Mel Frequency Cepstral Coefficients method is the most commonly used method of extracting features that gives high performance in speaker recognition applications. Pre-emphasis, framing, windowing, and Fast Fourier Transform is applied to voice signals.

Mel spectrum and Mel frequency cepstral coefficients are obtained by processing with a filter sequence generated by the Mel scale and by performing discrete cosine transform of the logarithm of the Mel spectrum respectively. (Kizrak, Bolat, 2014) The steps of the Mel frequency Cepstral Coefficients method are shown in Figure 1 as a block diagram.

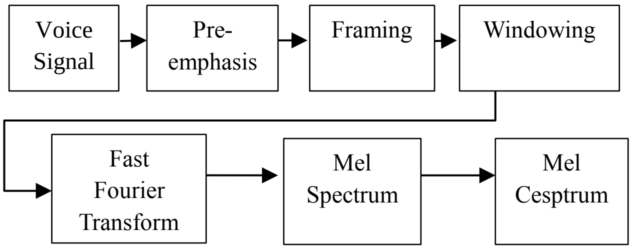

Figure 1. Feature vector extraction steps

2.2. Normalization Techniques

2.2.1. Short-time mean and variance normalization

Short-time Mean and Variance Normalization formula is shown in equation (2). Here, m and k represent frame and feature vector, respectively.

$$C_{STMVN}(m,k) = \frac{C(m,k) - \mu_{ST}(m,k)}{\sigma_{ST}(m,k)} \tag{2}$$

In the formula (2), $\mu_{ST}(m,k)$ obtained from formula (3) and $\sigma_{ST}(m,k)$ calculated with formula (4) are Short-time Mean and Short-time standard variation. (Alam et. al. 2011)

$$\mu_{ST}(m,k) = \frac{1}{L}\sum_{j=m-\frac{L}{2}}^{m+\frac{L}{2}} C(j,k) \tag{3}$$

$$\sigma_{ST}(m,k) = \frac{1}{L}\sum_{j=m-\frac{L}{2}}^{m+\frac{L}{2}} (C(j,k) - \mu(m,k))^2 \tag{4}$$

2.2.2. Short-time cepstral mean and scale normalization

Short-time Cepstral Mean and Scale Normalization is shown in equation (5). Here, m and k shows frame and feature vector.

$$C_{STMVN}(m,k) = \frac{C(m,k) - \mu_{ST}(m,k)}{d_{ST}(m,k)} \tag{5}$$

In formula (5), $d_{ST}(m,k)$ represents Short-time lower and upper bound difference and is calculated by using the equation (6). The range of the j parameter is given by equation (7). (Alam et. al. 2011)

$$d_{ST}(m,k) = \max\ C(j,k) - \min C(j,k) \tag{6}$$

$$\left(\mathrm{m} - \frac{\mathrm{L}}{2}\right) \leq \ \mathrm{j} \leq \ \left(\mathrm{m} + \frac{\mathrm{L}}{2}\right) \tag{7}$$

2.2.3. Max-min normalization

Max – Min Normalization is one of the most fundamental normalization techniques and shown in equation (8). Here, m and k shows frame and feature vector.

$$C_N(m,k) = \frac{C(m,k) - \min C(m,k)}{\max C(m,k) - \min C(m,k)} \tag{8}$$

2.2.4. Z-score normalization

Z-Score Normalization and Standard Deviation Normalization represented in equation (9) and (10), respectively. Here std C shows standard deviation of C and $\bar{C}$ Shows mean value of C.

$$C_Z(m,k) = \frac{C(m,k) - \bar{C}}{std\ C} \tag{9}$$

$$C_S(m,k) = \frac{C(m,k)}{std\ C} \tag{10}$$

3. EXPERIMENTAL RESULTS

By using the reduced feature vectors obtained from 192 female speakers without any normalization, the gender of 187 female speakers was correctly estimated. But five female speakers’ gender was estimated as male. The estimation success rate was 97.3958%. From the 192 male speakers, the gender of 188 male speakers was correctly estimated and four male speakers’ gender is estimated as female. The estimation success rate was 97.9166%.

In total 375 of the 384 speakers gender were correctly estimated and the gender of 9 speakers was incorrectly estimated. Classification success rate without applying any normalization is shown in Figure 2.

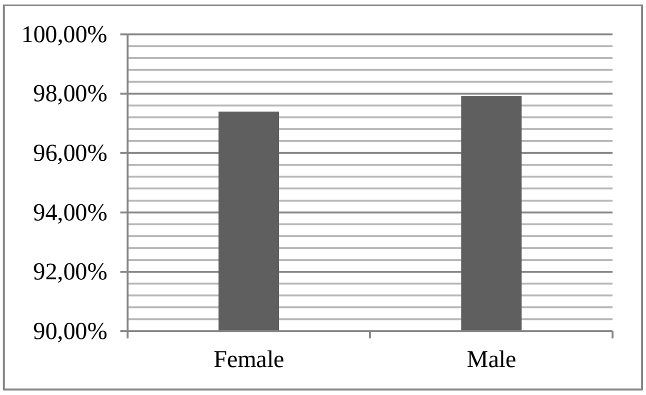

Figure 2. Classification success results without applying normalization

After applied Short-time Mean and Variance Normalization to the reduced feature vectors, the gender of 188 male speakers was correctly estimated. But four male speakers' gender was estimated as female. The estimation success rate was 97.9166%.

While the gender of 180 of 192 female speakers was correctly estimated, the gender of 12 speakers was incorrect. In total the gender of 368 of 384 speakers was successfully predicted, while the gender of 16 speakers was unsuccessfully estimated. Total success rate was %95.8333.

Results shows that estimation success for female speakers is decreased and estimation success for male speakers does not changed when features which normalized by Short-time Mean and Variance Normalization. So it could be said that Short-time Mean and Variance Normalization is not a good choice for speakers' gender estimation.

Comparison of success rates between before normalizing and after are given in Figure 3.

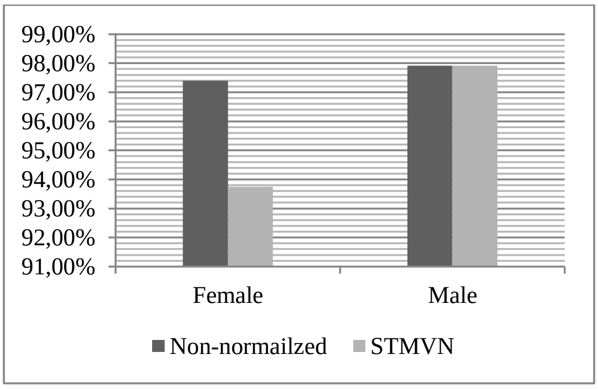

Figure 3. Classification success results for STMVN

After applied Short-Time Cepstral Mean and Scale Normalization the reduced feature vectors, the gender of 188 male speakers was correctly estimated. But 4 male speakers' gender was estimated as female. And the estimation rate was 97.9166%.

While the gender of 178 of 192 female speakers was correctly estimated and the gender of other 14 females speakers was incorrect. And success rate was 92.7083 for female speakers. In total the gender of 368 of 384 speakers was successfully predicted, while the gender of 16 speakers was unsuccessfully estimated.

Total success rate was %95.3125. According to this results, success rate of gender estimation for females is decreased and success rate of gender estimation for males is not changed. Similarly to short-time mean and variance normalize, short-time mean and scale normalization is also not a good normalizing technique to use for speakers' gender estimation.

Comparison of success rates between before normalizing and after are given in Figure 4.

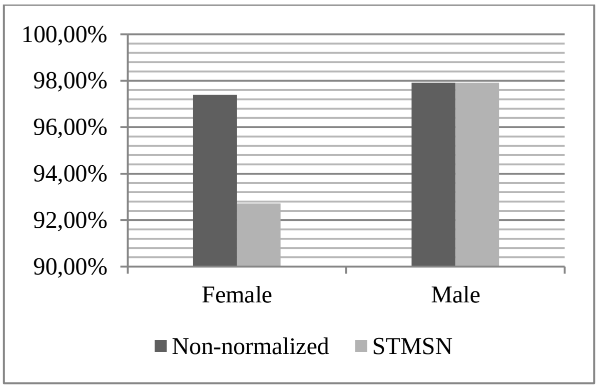

Figure 4. Classification success results for STMSM

For max-min normalized feature vectors, it is seen that gender of 181 of 192 female and 181 of 192 male speakers was correct from the test results. Success rate was %94.2708. This results shows us, using max-min normalized feature vectors to estimate speakers' gender decreases success rate for both male and female speakers. So max-min normalization is another bad choice to use for speakers gender estimation. Comparison of success rates between before normalizing and after are given in Figure 5.

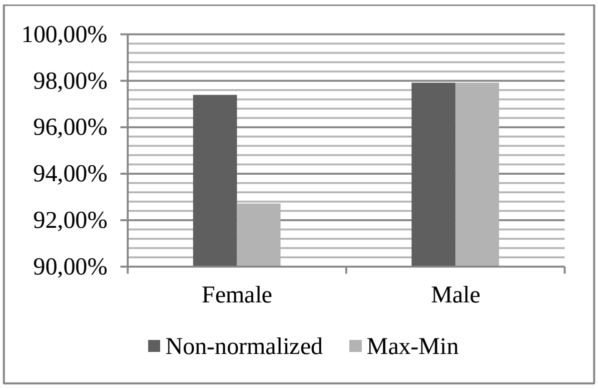

Figure 5. Classification success results for max-min normalization

After applied Z-Score Normalization to the reduced vectors, the gender of 183 of 192 female speakers with a success rate as 95.3125%. And the gender of 191 of 192 male speakers was estimated correctly with a success rate as 99.4791%. General success rate for both gender was 97.3958%. It shows that in spite of the increasing success rate of estimating male speakers, success rate of estimating gender for all speakers is decreased because of decreasing success rate of estimating female speakers. Comparison of success rates between before normalizing and after are given in Figure 6.

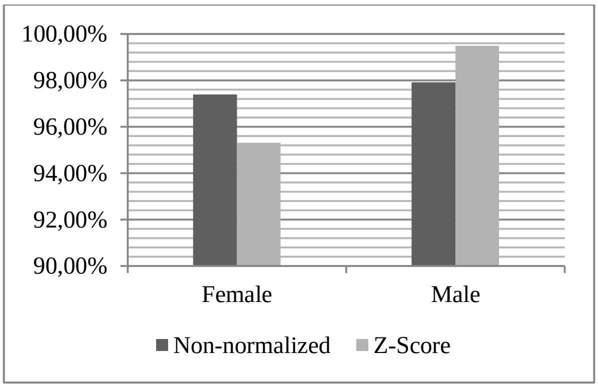

Figure 6. Classification success results for z-score normalization

In contrast to other normalizing techniques, standard deviation normalizing gives better results and improves success rate for speakers’ gender estimation with an average success rate as 98.1771%. When feature vectors that normalized by using standard deviation normalizing were used, the gender of 186 of 192 female speakers and gender of 191 of 192 males was predicted successfully.

And success rates were 96.8750% for females and 99.4791% for males.

In spite of decreasing success rate of estimating gender of female speakers, average success rate is increased because of significant increase of estimation success rate for male speakers. As a result, it could be said that standard deviation normalization may be a good choice for speakers gender estimation. Comparison of success rates between before normalizing and after are given in Figure 7.

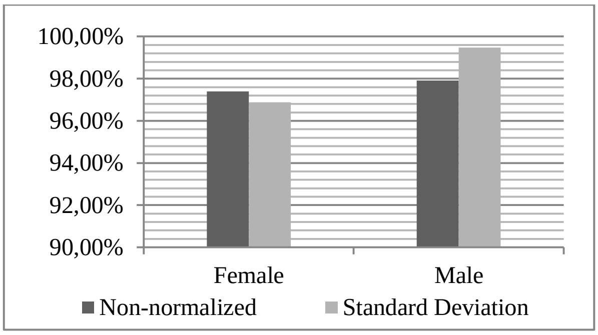

Figure 7. Classification success results for standard deviation normalization

Comparison of success rates for non-normalized features and normalized ones by each normalizing techniques are given in Figure 8.

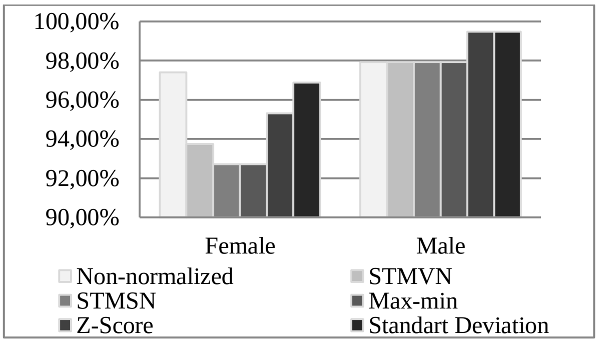

Figure 8. Compare of non-normalized and all normalization techniques

4. RESULTS AND DISCUSSION

When comparing the obtained results with and without any normalization techniques the below following conclusions can be reached: The highest success rate is taken by Standard Deviation Normalization technique within STMVN, STMSN, Z-Score Normalization, Min-Max Normalization and Standard Deviation Normalizations. The other normalization techniques decreases the total success rate in the test scenarios.

REFERENCES / KAYNAKLAR

Alam, M. J. vd. (2011) Comparative Evaluation of Feature Normalization Techniques for Speaker Verification. International Conference on Nonlinear Speech Processing, Springer Berlin Heidelberg.

Chen, O. T-C. & Gu, J. J. (2015) Improved Gender/Age Recognition System Using Arousal-Selection and Feature-Selection Schemes. Digital Signal Processing (DSP), 2015 IEEE International Conference on. IEEE

Djemili, R. vd. (2012)A Speech Signal Based Gender Identification System Using Four Classifiers. Multimedia Computing and Systems (ICMCS), 2012 International Conference on. IEEE

Durukal, M. & Hocaoğlu A. K. (2015) Performance Optimization on Emotion Recognition from Speech. Signal Processing and Communications Applications Conference (SIU), 2015 23th. IEEE

Heerden C. vd. (2010) Combining Regression and Classification Methods for Improving Automatic Speaker Age Recognition. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE

Islam, M. A. (2016). GFCC-Based Robust Gender Detection. In Innovations in Science, Engineering and Technology (ICISET), International Conference on. IEEE.

Khanum, S., & Firos, A. (2017). Text Independent Gender Identification In Noisy Environmental Conditions. In Computing, Communication and Automation (ICCCA), 2017 International Conference on. IEEE.

Kizrak, M. A. & Bolat, B. (2914) Klasik Türk Müziği Makamlarının Tanınması. Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu (ASYU), 2-6.

Nabiyev, V. V. & Yücesoy, E. (2009) VQ Yöntemiyle Konuşmacı Cinsiyetinin Belirlenmesi. Turkish Journal of Computer and Mathematics Education Vol 1.1, 35-47.

Přibil, J. vd. (2016) GMM-Based Speaker Gender and Age Classification After Voice Conversion. Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016 First International Workshop on. IEEE

Yücesoy, E. & Nabiyev, V. V. (2014) Comparison of MFCC, LPCC and PLP Features for The Determination of A Speaker's gender. Signal Processing and Communications Applications Conference (SIU), 2014 22nd. IEEE

Yusnita, M. A. vd. (2017) Automatic Gender Recognition Using Linear Prediction Coefficients and Artificial Neural Network on Speech Signal. In Control System, Computing and Engineering (ICCSCE), 2017 7th IEEE International Conference on. 2017

Yücesoy, E. & Nabiyev, V. V. (2009) Gender Identification of The Speaker Using DTW Method. Signal Processing and Communications Applications Conference, SIU 2009. IEEE 17th. IEEE

Yücesoy, E. & Nabiyev, V. V. (2016) Konuşmacı Yaş ve Cinsiyetinin Gkm Süpervektörlerine Dayalı Bir Dvm Sınıflandırıcısı ile Belirlenmesi. Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 31.3