Özellik Vektörlerinde Enerji Türevleri ile Konuşmacı Bağımsız Türkçe Konuşma Tanıma İyileştirmesi

Speaker Independent Turkish Speech Recognition Optimization with Energy Derivates on Feature Vectors

Mert Yılmaz ÇAKIR
Bilgisayar Bilimleri ve Mühendisliği
İstanbul Sabahattin Zaim Üniversitesi
İstanbul, Türkiye
mert.cakir@std.izu.edu.tr

Yahya ŞİRİN
Bilgisayar Bilimleri ve Mühendisliği
İstanbul Sabahattin Zaim Üniversitesi
İstanbul, Türkiye
yahya.sirin@izu.edu.tr

***Özetçe*—Günümüzde akıllı cihazlarda kullanımı artan konuşma tanıma uygulamaları konuşmacı bağımsız sistemler üzerinde başarılı performans göstermesiyle önem kazanmaktadır. Bu çalışmada konuşma içerisinden geleneksel çalışmalar ile çıkartılan özellik vektörlerinde arttırım sağlanmıştır. Bu işlem için ses bantlarındaki saklı konuşmacı kimliklerini ölçmeye yarayan enerji ve delta türevleri uygulanmıştır. Elde edilen katsayılar akıllı cihazlara tahmin yeteneği kazandırabilmek için giriş değeri olarak verilmektedir. Çalışmanın verimliliğini belirtmek için geleneksel çalışmalar ile yapılan çalışmanın kıyaslanması ve değerlendirilmesi sunulmuştur.**

Anahtar Kelimeler — ses işleme, veri madenciliği, örüntü tanıma, konuşmacı bağımsız konuşma tanıma; özellik çıkarımı; mel frekanslı kepstral katsayılar; enerji ve delta türevleri.

***Abstract*— At the recent times, speech recognition applications, which are increasingly used in smart devices, are gaining importance as they perform well on speaker-independent systems. In this study, an increase is obtained in the feature vectors extracted from the speech by traditional studies. For this process, energy and delta derivatives were applied to measure hidden speaker identities in audio bands. The coefficients obtained are given as input values to gain estimation ability to intelligent devices. The comparison and evaluation of the work, that is done with the traditional studies, is presented to indicate the efficiency of the work.**

Keywords — sound processing, data mining, pattern recognition, speaker independent speech recognition; feature extraction; mel frequency cepstral coefficients; energy and delta derivates.

I. GİRİŞ

Bilgisayar bilimlerinin ilgilendiği disiplinler arasında doğal dil işleme gözde konulardan biri olmaktadır. Doğal dil işleme konularından ve insanlar arasındaki etkili iletişim araçlarından biri olan konuşma tanıma, günümüz ihtiyaçlarının kolaylaştırılması ve güvenlik sistemleri başta olmak üzere eğilimin arttığı bir alandır. Akıllı cihazların kullanım oranının artmasıyla konuşma tanıma önem kazanmıştır. İlgilendiği konular kapsamında konuşma tanıma sinyal işleme [1], örüntü tanıma [2] gibi farklı bilgisayar bilimlerinin kesişiminde yer almaktadır.

Konuşma tanıma sistemleri konuşmacı bağımlılığı temel alınarak iki gruba ayrılır; konuşmacı bağımlı sistemler ve konuşmacı bağımsız sistemler [3]. Konuşmacı bağımlı konuşma tanıma sistemleri belirli kişi / kişiler tarafından önceden sisteme tanıtılmış konuşmalar ile bu kişi / kişileri tanımaya imkân sağlar. Konuşmacı bağımsız konuşma tanıma sistemi ise herkesin konuşmasını tanımak için tasarlanır. Bu nedenle eğitim setinin birçok faklı kullanıcılar ile eğitilmesi gereklidir.

1950'li yıllardan konuşmacı bağımlılığı üzerine çalışılan konuşma tanıma sistemlerde başka bir konuşmacının sisteme tanıtılması istenildiğinde, sistem üzerinde kayıtlı olan ve konuşma tanıma için kaynak olarak alınan şablonların güncellenmesi gerekmektedir. Günümüzde konuşma tanıma alanında yapılan çalışmalar konuşmacı bağımsız konuşma tanıma sistemleri üzerinedir. Fakat bu sistemlerin bir dezavantajı, herhangi bir dil için bütün konuşmacı varyasyonlarını modellemenin olanaksız olmasıdır. Bu neden ile konuşmacı bağımsız sistemler performans açısından konuşmacı bağımlı sistemlere göre daha geride kalmaktadır. Fakat kullanım alanı göz önüne alındığında konuşmacı bağımsız sistemler, konuşmacı bağımlı sistemlere tercih edilmektedir.

Bir konuşma sinyalinin, zaman ekseninde dalga formu tüm işitsel bilgileri taşır. Akustik ve konuşma teknolojisindeki geçmiş araştırmalar ile bilgi olarak kabul edilebilecek verileri sabit matematiksel ifadelere dönüştürmek için birçok yöntem geliştirilmiştir. Bu bağlamda bir konuşma tanıma sisteminin tasarımı, konuşmanın alınması, özellik çıkarımı ve sınıflandırma aşamaları olmak üzere üç kısma ayrılır. Özellik çıkarımı yöntemi, ölçülen ilk verilerle başlayarak bilgilendirici ve gereksiz olmaması amaçlanan türetilen değerleri (özellikleri) oluşturur [4]. Konuşmanın elde edilen bilgilerle tanıma işlemlerine kaynak oluşturacak aşama sınıflandırmadır. Bu aşamalar sonucunda konuşma yazıya çevrilir.

Bu çalışmada Türkçe konuşmacı bağımsız konuşma tanıma için farklı konuşmacılardan konuşmalar alınarak eğitim için bir

978-1-5386-1501-0/18/$31.00 ©2018 IEEE

veri seti oluşturulmuştur. İkinci aşamada konuşmalardan bir sonraki sınıflandırma aşamasına verilecek sabit özellik katsayılarını çıkarmak için literatürde yaygın olarak kullanılan Mel Frekanslı Kepstral Katsayılar (Mel Frequency Cepstral Coefficients (MFCC)) [5] yöntemi kullanılmıştır. Geleneksel çalışmalarda MFCC ile 12 katsayı elde edilirken, bu çalışma ile elde edilen katsayılar üzerinden konuşmacı bağımlı sistemlerde kullanılan [6] enerji ve delta türevleriyle 39 katsayı elde edilerek kıyaslama yapılmıştır. Eğitim setinden farklı konuşmacıların sesleri ile test aşamasında vurgulamanın öneme sahip olduğu Türkçe dili için başarı sağlanmıştır.

Bu çalışma şu şekilde yapılandırılmıştır: İkinci bölümde geçmiş çalışmalar incelenmiş ve değerlendirme yapılmıştır. Üçüncü bölümde önerilen çalışmanın detayları anlatılmaktadır. Dördüncü bölümde veri setinden bahsedilmiş ve tasarlanan sistemin test aşamasına değinilmiştir. Beşinci bölümde deney sonucuyla elde edilen performans değerleri yorumlanarak gelecekte başarımın arttırılması adına kullanılması muhtemel veri setlerinden bahsedilmiştir.

II. ÖNCEKİ ÇALIŞMALAR

Geçtiğimiz yıllarda konuşmacı bağımlı ve bağımsız yöntemler üzerine birçok çalışma yapılmıştır. Joshi ve Cheeran [7], konuşma tanıma için özellik çıkarma tekniklerini tartışmışlar ve konuşma tanıma için genel bir bakış açısı sunmuşlardır. Çalışmalarını, özellik çıkarma ve sınıflandırma olmak üzere iki farklı kısma ayırmışlardır. Özellik çıkarmada MFCC tekniği ile sınıflandırmada Yapay Sinir Ağları (Artificial Neural Networks (ANN)) tekniğinde yaklaşık olarak %80 doğruluk elde etmişlerdir.

Cahavan ve Sable [8], konuşma tanıma için özellik çıkarma tekniklerini tartışmışlar ve konuşma tanıma için genel bir bakış açısı sunmuşlardır. Çalışmalarını, özellik çıkarma ve sınıflandırma olmak üzere iki farklı kısma ayırmışlardır. Özellik çıkarmada MFCC tekniği ile sınıflandırmada Dinamik Zaman Bükmesi (Dynamic Time Warping (DTW)) tekniğinde yaklaşık olarak %90, Saklı Markov Modeli (Hidden Markov Models (HMM)) tekniğinde yaklaşık olarak %96 oranında doğruluk elde etmişlerdir.

Ananthi ve Dhanalakshmi [9] konuşma tanıma yaklaşımını işitme engelli insanlar için daha yararlı olabilecek konuşma sözlüğünden gelen metni tanıma amacı ile kullanmışlardır. Sınıflandırmada, konuşma tanıma sistemi için yaygın olarak kullanılan teknikler olan Destek Vektör Makinesi (Support Vector Machines (SVM)) ve HMM'yi kullanmışlardır. Özellik çıkarmada MFCC ile akustik özellikleri çıkarmışlardır. Sistem, HMM için %98.92 SVM için %91.46'lık bir doğruluk sergilemiştir. Ayrıntılı analiz sonucu, MFCC ile HMM'nin SVM gibi diğer modelleme tekniklerinden daha iyi performans gösterdiğini belirtmişlerdir.

Ses işlemede, Mel Frekanslı Kepstrum (MFC), bir frekansın doğrusal olmayan Mel skalasında kısa süreli güç spektrumunun bir gösterimidir. Sesin bir tür kepstral gösteriminden türemiştir. MFC'de frekans bantları, insan ses sistemi tepkisini yaklaşık olarak Mel ölçeğinde eşit aralıklarla yerleştirilmiştir. Böylece ses sıkıştırmasında daha iyi bir ses temsili sağlanabilir. Bu çalışmada da geçmiş çalışmaların incelenmesi sonucu özellik çıkarımında MFCC ile sınıflandırma aşamasında istatistiki süreç verileri iyi biçimde tanımlayabilen HMM yöntemleri kullanılmıştır.

III. SUNULAN ÇALIŞMA

Bu çalışmada ses sinyali, pencereleme ile sabit sayıda değerlere dönüştürülerek özellik çıkarımı aşamasına verilmektedir. Konuşmacı bağımsız konuşma tanıma sistemlerinde geleneksel çalışmalarda MFCC ile elde edilen katsayılardan fazla olarak, pencerelenen ses sinyali numuneleri üzerinden enerji, MFCC katsayları üzerinden delta ve delta delta teknikleri uygulanmıştır. Enerji ve delta türevleri genellikle konuşmacı kimliklerinin tanınmasında kullanılmaktadır. Vurgulamanın önemli olduğu Türkçe dili için bu tekniklerle bir çalışma yapılmıştır.

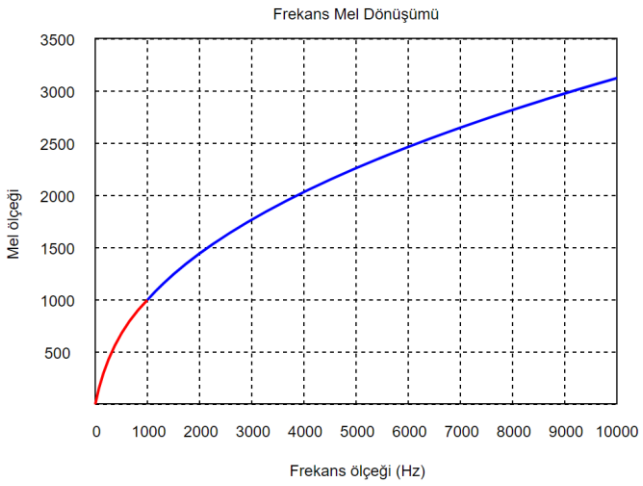
Konuşma sürekli değişen bir sinyaldir. Bu sebeple özellik çıkarımı ile konuşmanın temsilini en iyi şekilde sağlayan sabit sayılar çıkarılmalıdır. MFCC aşamasına giriş verisi olarak her konuşmanın, her biri anlık ses olan 512 örnekli çerçevelere ayrılmış şekilde verilmesiyle işleme başlanılmıştır. Çerçeveler üzerine Hamming Pencereleme ile her bir çerçeve sonlu hale getirilir. Fonemlerden elde edilen sabit sayılar MFCC aşamasına giriş değeri olarak verilmektedir.

Önerilen çalışmada kullanılan MFCC tekniğinde ilk olarak Ayrık Fourier Dönüşümü (Discrete Fourier Transform (DFT)) uygulanır. Geçerli çerçevenin frekans içeriğini (spektrumu) ayıklamak için pencereli sinyal üzerinde işlem yapılmıştır. Spektral bilginin çıkartılması için örnek olarak sinyalin, ayrık zaman (örneklenmiş) sinyali için ayrı frekans bantlarında ne kadar enerji içerdiği DFT ile ifade edilmiştir. DFT'ye giriş, pencereli bir sinyal x[n] ... x[m] olup, N ayrı frekans bandının her biri için çıktı, frekans bileşeninin büyüklüğünü ve fazını temsil eden bir karmaşık sayı orijinal sinyaldeki X[k]'dır. DFT, Fourier dönüşümünün eşit aralıklı frekanslardaki örneklerine özdeştir; Sonuç olarak N -noktalı bir DFT'nin hesaplanması Fourier dönüşümünün N örneğinin, N eşit aralıklı frekanslarla $(w_k = 2\pi k/N)$, z-düzlemindeki birim çember üzerinde N nokta ile hesaplanmasına karşılık gelir. Buradaki temel amaç N -noktalı DFT'nin hesaplanması için verimli algoritmaların kullanılmasıdır. "(1)" Numaralı formül DFT'nin hesaplanmasıdır. Çalışmada, w açısal frekansın bir fonksiyonu olmak üzere $S_w(n)$ ayrık işaretini, 2π periyodunda N tane eşit aralıklı örnek olarak alınırsa $(2\pi/N)$ formül "(1)" gibi olmaktadır.

$$bin_k = \left|\sum_{n=1}^{N} S_w(n) e^{-i(n-1)\frac{2\pi k}{N}}\right|, k = 0,1,2,\dots,N-1 \quad (1)$$

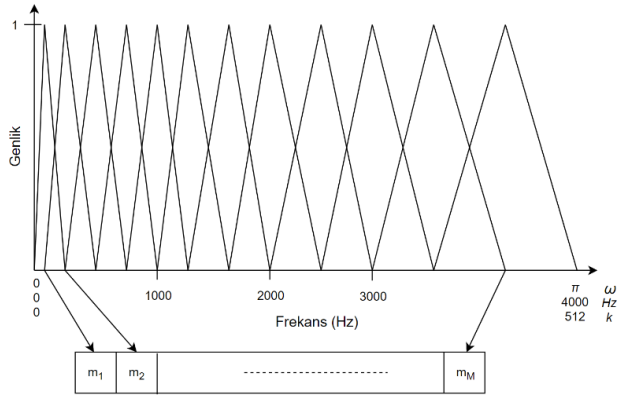
İkinci aşamada Stevens ve arkadaşları [10] tarafından ortaya atılan bir tondaki frekansı ifade eden ve insan tarafından algılanan ses sinyali frekansının ölçümü olan Mel ölçümü kullanılmıştır. Mel ile frekans arasındaki ilişkisel formül "(2)" deki gibi gösterilir [11].

$$F_{mel}(f_{Hertz}) = 1127In(1 + \frac{f_{Hertz}}{700}) \quad (2)$$



Şekil 1. Frekans ve Mel arasındaki ilişki

Şekle göre Mel filtre bankasını açıklamak gerekirse;



Şekil 2. Mel Filtre Bankası

Şekil 1 ve 2'de görüleceği üzere Mel filtreleme ile elde edilen değerlerde frekans – genlik grafiğinde 1000 Hz eşik değerdir. Grafiğin altında ise her bir frekans anında Mel katsayıları gösterilmektedir [12]. "(2)" Formülünün bir diğer gösterim şekli her bir frekans anı x doğrusal frekans olacak şekilde "(3)" formülü gibidir.

$$Mel(x) = 2595\log[1 + \frac{x}{700}] \quad (3)$$

Mel frekans çarpması, Mel frekanslarına göre merkezlenmiş filtreler içeren bir filtre bankası kullanılarak yapılmıştır. Üçgen filtrelerin genişliği, Mel ölçeğine göre değiştiğinden merkez frekans etrafındaki kritik banttaki toplam enerji dahil edilmiştir. Otuz Filtre kullandığımız filtre bankasında, Mel filtreleme sonucunda her bir Mel ölçek bandı enerji dağılımı hakkında bilgi vermektedir. "(4)" Formülünden x değeri elde edilmiştir.

$$x = 700(10^{\frac{mel}{2595}} - 1) \quad (4)$$

Üçüncü aşamada Mel filtreleme ile elde edilen katsayıların her birine log güç spektrumunun hesaplaması "(5)" formülüne göre yapılmıştır:

$$f_i = \ln(f_{bank_i}) \quad (5)$$

Dördüncü aşamada Ayrık Fourier Dönüşümünün Tersi (Inverse DFT (IDFT)) uygulanarak her çerçevenin MFCC özelliğini elde etmek için, kepstrum dönüşümü filtre çıkışlarına uygulanmıştır. Üçgen filtre çıktıları Y(i), i = 0,1, 2, ..., M logaritma kullanılarak sıkıştırılmıştır ve DFT uygulanmıştır. Burada M, filtre bankasındaki filtre sayısına, yani otuza eşittir. c[n], her çerçeve için MFCC vektörü olacak şekilde "(6)" formülü uygulanmıştır.

$$c[n] = \sum_{i=1}^{M} logY(i)\cos\left[\frac{\pi n}{M}(i - \frac{1}{2})\right] \quad (6)$$

IDFT sonrasında her konuşma çerçevesinden çıkarılan özellik vektörü MFC olarak adlandırılır ve tek tek bileşenler MFCC'yi oluşturur.

Bir konuşma sinyali, kaydedildiğinde kanal gürültüsüne maruz kalabilir ve belirli bir konuşmacı için eğitim verilerini kaydederken kanal etkisi, konuşmacı sistemi kullandığında sonraki kayıtlarda kanal etkisinden farklıysa, bir sorun ortaya çıkar. Sorun, eğitim verileri ile yeni kaydedilen veriler arasındaki yanlış uzaklığın, farklı kanal etkilerinden dolayıdır. Bu sebeple "(7)" formülündeki Kepstral Ortalama Çıkarma ile kanal efekti, Mel kepstrum katsayılarının ortalama Mel kepstrum katsayılarıyla çıkarılmasıyla ortadan kaldırılmıştır.

$$mc_j(q) = c_j(1) - \frac{1}{M}\sum_{i=1}^{M} c_i(q),\ q = 1,2,\ldots,12 \quad (7)$$

Geçmişte yapılan çalışmalarda özellik çıkarımı aşamasında önerilen MFCC ile 12 katsayılı özellik vektörleri [13] elde edilmiştir. Bu çalışmada genelde konuşmacı tanıma uygulamalarında kullanılan ses bantlarında saklı konuşmacı kimliklerini ölçmeye de yarayan enerji ve delta türevleri ile 39 katsayılı özellik vektörleri elde edilerek 12 katsayılı özellik vektörleri ile yapılan çalışma kıyaslanmış ve performans değerlendirmesi yapılmıştır.

Bir çerçevedeki enerji, çerçevedeki örneklerin gücü toplamıdır; böylece zaman örneği t_1 'den zaman örneği t_2 'ye kadar bir penceredeki bir sinyal x'in enerjisi "(8)" formülünden elde edilmiştir.

$$Energy = \sum_{t=t_1}^{t_2} x^2[t] \quad (8)$$

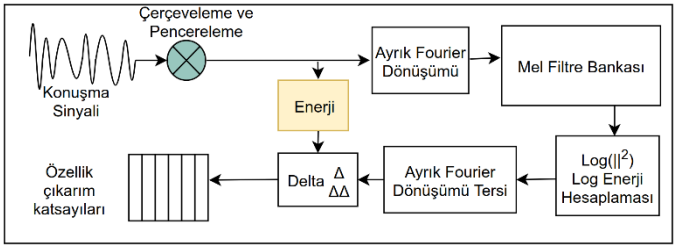
Konuşma sinyali, kareden kareye sabit olduğundan hız (delta) ve ivmelenme (delta delta) katsayıları genellikle statik pencere tabanlı bilgi ile elde edilmektedir. Bu delta ve delta delta katsayıları, bitişik pencereler arasındaki kepstral özellik vektörlerinin değişim hızını ve ivmesini modeler [6]. Deltalar, çerçeveler arasındaki farkın hesaplanması ile elde edilir; böylece çalışmada belli bir kepstral c(t) zamanı t için delta değeri d(t) "(9)" formülünden elde edilmiştir.

$$d(t) = \Delta f_k[i] = f_{k+M}[i] - f_{k-M}[i] \quad (9)$$

Farklılaştırma yöntemi basittir fakat parametre alanında yüksek geçiren bir filtreleme işlemi görevi görmesi nedeniyle, gürültüyü yükseltme eğilimi gösterir. Bu soruna çözüm olarak doğrusal regresyon, yani birinci derece polinom alınmıştır ve regresyon penceresi boyutu M=4 olacak şekilde en küçük kareler çözümü "(10)" formülü ile hesaplanmıştır.

$$\Delta f_k[i] = \frac{\sum_{m=-M}^{M} m f_{k+m}[i]}{\sum_{m=-M}^{M} m^2} \qquad (10)$$

Bu aşamalar neticesinde özellik katsayıları elde edilmektedir. Özellik çıkarımı aşamasında uygulanan teknikler detaylı olarak Şekil 3'te gösterilmektedir.



Şekil 3. Özellik çıkarımı katsayılarının elde edilmesi

Özellik çıkarımı ile elde edilen sabit katsayılar ile sınıflandırma aşamasına giriş değerleri verilmiştir.

IV. DENEYLER

Bu bölümde önerilen konuşma tanıma sisteminin testi için hazırlanan veri setinden ve geleneksel çalışmalar ile kıyaslamasından bahsedilmektedir. Önerilen çalışmada eğitim aşamasında farklı cinsiyet ve yaş gruplarından oluşan 15 konuşmacı 10 farklı kelimeyi üçer kere seslendirmiştir. Test aşamasında her kelime 3 farklı konuşmacı tarafından üçer kere test edilmiştir.

Özellik çıkarımı ile elde edilen katsayılar 12 MFC katsayısıdır. Önerilen enerji türevlerinin çerçevelere ve 12 MFC katsayısına uygulanması ile 12 MFCC, 12 Delta, 12 Delta Delta, 1 Enerji, 1 Delta Enerji, 1 Delta Delta Enerji olmak üzere toplam 39 özellik elde edilmiştir. Bu özellikler sınıflandırma aşamasına verilerek sistemin tahmin yapması sağlanmıştır. Bunların sonucunda 12 katsayılı özellik vektörleri ile %84,3 başarı, 39 katsayılı özellik vektörleri kullanılarak %86,6 başarı elde edilmiştir.

V. SONUÇ

Türkçe kelime içinde hecelerin, cümle içinde ise kelime / kelime gruplarının diğerlerine göre daha baskın ya da kuvvetli söylenmesine göre vurgunun öneme sahip olduğu bir dildir. Vurguların konuşmalarda öneme sahip olduğu düşünüldüğünde genellikle konuşmacı bağımlı konuşma tanıma sistemlerinde kullanılan ses bantlarında saklı konuşmacı kimliklerini ölçmeye yarayan enerji ve delta türevlerinin bu çalışma ile konuşmacı bağımsız konuşma tanıma sistemlerinde de etkili olduğu tespit edilmiştir. Bu çalışma ile konuşmacı bağımsız konuşma tanımada, sistemin tahmin performansını yükseltmek amacıyla, literatürdeki çalışmalara kıyasla özellik çıkarımında daha fazla bilgi çıkarılmıştır. Özellikle daha büyük veri setlerinde bu çalışma ile elde edilen başarı oranının daha da artacağı düşünülmektedir.

KAYNAKLAR

[1] Rabiner, L.R., Gold, B. (1975). Theory and application of digital signal processing. Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p.

[2] Juang, B. H., Hou, W., & Lee, C. H. (1997). Minimum classification error rate methods for speech recognition. IEEE Transactions on Speech and Audio processing, 5(3), 257-265.

[3] Çakır M.Y. (2017). Real-Time High-Quality Voice Recognition, Master Thesis, Istanbul Sabahattin Zaim University, Halkalı / İstanbul, December.

[4] Shrawankar, U., & Thakare, V. M. (2013). Techniques for Feature Extraction In Speech Recognition System: A Comparative Study. International Journal Of Computer Applications In Engineering, Technology and Sciences (IJCAETS), ISSN 0974-3596,2010, 6 May (s. 412-418). Cornell University Library.

[5] Müller, M. (2007). Information retrieval for music and motion (Vol. 2). Heidelberg: Springer.

[6] Kinnunen, T. (2003). Spectral Features for Automatic Text-Independent Speaker. Finland: University of Joensuu, Department of Computer.

[7] Joshi, S. C., & Cheeran, A. N. (2014). MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 3(7).

[8] Chavan, M. R. S., & Sable, G. S. (2013). An Implementation of Text Dependent Speaker Independent Isolated Word Speech Recognition Using HMM. International Journal Of Engineering Sciences & Research Technology, 2(9).

[9] Ananthi, S., & Dhanalakshmi, P. (2015). SVM and HMM modeling techniques for speech recognition using LPCC and MFCC features. In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014 (pp. 519-526). Springer, Cham.

[10] Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America, 8 (3), (s. 185–190).

[11] Uzunçarşılı, M. (2005). Vektör Nicemleme Tekniklerine Dayalı Konuşmacı Tanıma Algoritmalarının İncelenmesi. Ankara: Ankara Üniversitesi Fen Bilimleri Üniversitesi.

[12] Jurafsky, D., & Martin, J. H. (2006). Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.

[13] Rabiner, L., & Juang, B. (1993). Fundamental of Speech Recognition. Englewood Cliffs, New Jersey: PTR Prentice Hall.