Ses Etiketleme

Özkan YILMAZ(040080405), Yaşar KARTAL(040080513), Y.Buğra ÖZER(040080534)

Elektrik-Elektronik Fakültesi, Elektronik ve Haberleşme Mühendisliği Bölümü
Telekomünikasyon Mühendisliği Programı
İstanbul Teknik Üniversitesi, İstanbul
yilmazozkan@itu.edu.tr, yasar.kartal@itu.edu.tr, ozeryu@itu.edu.tr

Özet

Bu çalışmada belirli bir ses dosyasındaki insan sesi ve müziğin etiketlenmesi amaçlanmıştır. Bunun için çalışma öncesinde ses dosyasından alınan insan sesi ve müzik örnekleri, mel frekansı sepstral katsayıları ile çarpılarak öz vektörler elde edilmiştir. Elde edilen bu öz vektörler veri bankasına eklenmiş, eklenen bu öz vektörler karar kuralı oluşturmak için etiketleme aşamasında kullanılmıştır.

1. Giriş

Ses tanıma ve etiketleme sayısal işaret işlemenin önemli konularından bir tanesidir. Bu alanda yapılmış pek çok çalışma bulunmaktadır. Yapılan bu çalışmada bunlardan bir tanesidir. Bu tür çalışmalar, ses ayırt etme ve tanıma üzerine yapılan çalışmalar ile benzerlik gösterdiğinden, bu alanda kullanılan yöntemler ve geliştirilecek algoritmalar, ses tanıma işlemlerinin geliştirilmesinde ve karşılaşılan engellerin aşılmasında da büyük katkılar sağlayacaktır.

2. Ses Etiketleme Öncesinde Yapılan İşlemler

2.1 Tek Kanala İndirgeme

Ses etiketleme öncesinde işlemleri kolaylaştırmak için, iki kanallı (stereo) ses dosyası tek kanallı (mono) hale çevrilmiştir.

2.2 Normalizasyon

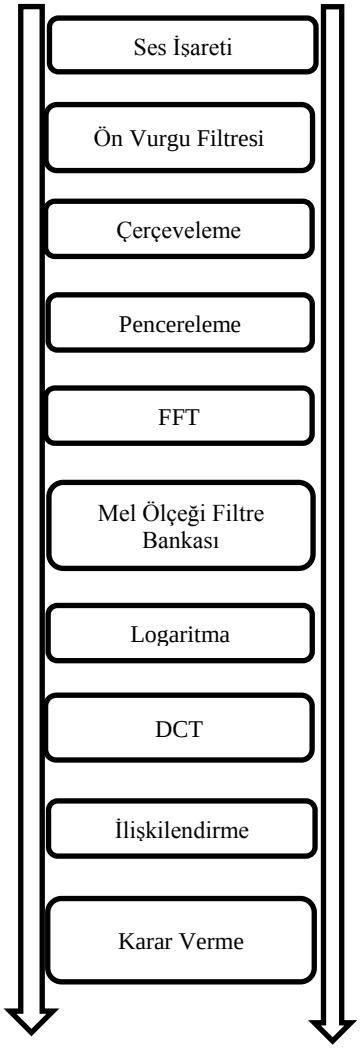
Ses dosyasının içindeki yüksek genlik değerlerinin öz vektör çıkarma sırasında olumsuz etkilerini kaldırmak için normalizasyon yapılmıştır. Normalizasyon işlemi, dosya içindeki örneklerin her birinin en yüksek genlik değerine bölünmesiyle yapılmıştır.[1]

$$x_{norm}[n] = \frac{x[n]}{\max\{|x[n]|\}} \qquad (2.2)$$

2.3 Sıfırların Elenmesi

Ses etiketleme işlemi sırasında sıfıra bölme hatasından kaçınmak için sıfır genlikli örnek değerleri çok küçük bir sayı olan 10^{-9} ile yer değiştirilmiştir.

3. Ses Etiketlemede Kullanılan İşlem Basamakları



Şekil 3.1 Ses Etiketleme Süreci

3.1 Ön Vurgu Filtresi

Ön vurgu filtresinin amacı sinyalin yüksek frekans spektrumuna ilişkin enerjisinin arttırılmasıdır. Ön vurgu filtresinin zaman bölgesindeki transfer fonksiyonu aşağıdaki eşitlikte gösterilen FIR (Sonlu Dürtü Tepkisi) filtresi ile yapılır.[7]

$$y(n) = x(n) - 0.95x(n-1) \qquad (3.1)$$

3.2 Çerçeveleme

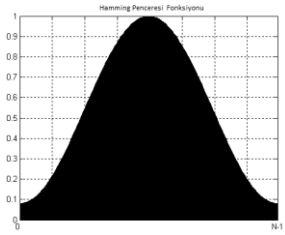
Bu aşamada ses sinyali, komşu çerçevelerin birbirinden M (M<N) sayıda örnekle ayrıldığı, N sayıda örnekten oluşan çerçevelere bölünür. İkinci çerçeve birinci çerçeveden M örnek sonra başlar ve birinci çerçeveyle N-M örnek kadar iç içe geçmiş durumdadır [2]. Bu çalışmada, kullanılan özelliklere bağlı olarak 512 örnekten oluşan %50 iç içe geçmiş çerçeveler kullanılmıştır. Çerçeve uzunluğu seçilirken öz vektörlerinin ses tanımada gösterdiği performans dikkate alınmıştır.

3.3 Pencereleme

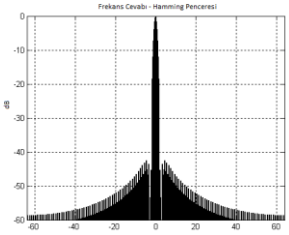
Sonsuz uzunlukta bir işaret dizisi ile çalışmak imkânsız olduğundan, tüm işaret analizinde pencelereme kaçınılmazdır. Analiz için işaretin bir bölümü seçilir seçilmez orijinal verinin pencerelendiği söylenebilir. En basit pencereleme tekniğinde verilen işaretin incelenecek bölümü 1 ile dışarda kalan gözlem dışı aralık ise 0 ile çarpılır. Örneğin bir sinüs veya kosinüs işaretinin sadece bir bölümünü makasla almak bu türden bir pencereleme işlemidir. Bu işlem sinüs dalgasının sonlu genişlikte birim pencere ile çarpımına eş değerdir. Frekans aralığında bu işlemin karşılığı konvolüsyondur. Yani orijinal işaretin Fourier dönüşümü olan impuls ile pencerenin spektrumunun konvolüsyonu söz konusudur. Pencerenin Fourier dönüşümünde ki yan lopları nedeniyle yan bantlarda bir spekturum sızıntısı vardır.
Dikdörtgen pencere fonksiyonundaki uçlardaki süreksizliklerin oluşturduğu spektrum dağılımından dolayı, diğer pencere fonksiyonları kullanma yoluna gidilir.[3]
Literatürde çok sayıda pencere fonksiyonu mevcuttur. Yapılan 'Ses Etiketleme' projesinde hamming pencere fonksiyonu kullanılmıştır.

$$W_n = 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right) \quad n = 0,1,2,\dots,N-1$$

(3.3)



Şekil 3.3 a



Şekil 3.3 b

Şekil 3.3 a,b Hamming Penceresi Fonksiyonu, Frekans Cevabı

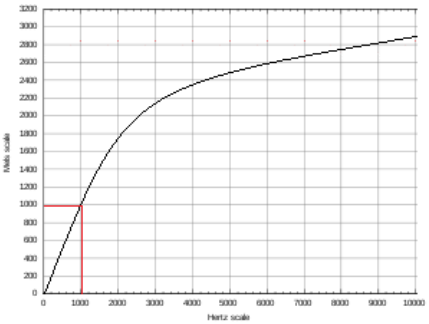
3.4 FFT

Ayrık Fourier dönüşümünün (DFT) doğrudan hesaplanmasında her bir X_K değeri için N karmaşık çarpma ve N-1 karmaşık toplama işlemi kullanılmaktadır. Bu neden ile N adet DFT değeri bulurken N^2 çarpma ve N.(N-1) toplama işlemi gereklidir. Ayrıca her karmaşık çarpma işlemi dört gerçel çarpma ve iki gerçel toplama işlemi ve her bir karmaşık toplama iki gerçel toplama işlemi ile gerçeklenmektedir. Sonuç olarak, dizi uzunluğu olan N'nin 1000'nin üzerinde olması durumunda direkt DFT'nin bulunması çok fazla miktarda işlem gerektirmektedir. Yani, N sayısı artarken gereken işlem sayısı çok hızlı artmaktadır. Bu noktada hızlı Fourier dönüşümü (FFT) algoritmaları DFT'nin hesaplanmasında hızlı ve daha ekonomik bir yöntem olarak karşımıza çıkmaktadır.[3]

$$X(m) = \sum_{n=0}^{N-1}\left(\mathrm{x(n)}.\,\mathrm{e}^{\frac{-j\pi m}{N}}\right) \qquad (3.4)$$

3.5 Mel Ölçeği Filtre Bankası

Mel frekansı kepstral katsayıları, konuşma tanıma sistemlerinde en çok kullanılan özelliklerden biri haline gelmiştir. Kepstral katsayılar doğrusal ölçekli olmakla beraber insan kulağı 1kHz'in altındaki frekansları doğrusal ölçekli, 1kHz'inüstündeki frekansları logaritmik ölçekli olarak duymaktadır. MFCC'nin kullanım amacı kepstral katsayıların, insan işitme sistemiyle uyumlandırılmasıdır. Aşağıdaki şekilde, 1,000Hz'in altında doğrusal aralıklı, 1,000Hz'in üstünde logaritmik aralıklı frekans bantlarından oluşan mel ölçeği yer almaktadır.

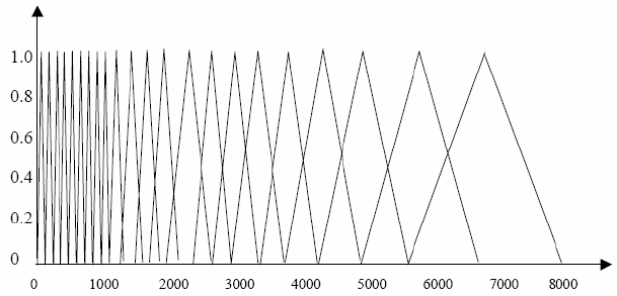


Şekil 3.5 Doğrusal Frekansı- Mel Frekansı Grafiği

Referans noktası olarak seçilen 1kHz'lık ses 1,000 mel olarak ifade edilir. Hz cinsinden verilen frekansı mel olarak ifade etmek için (3.5) formülü kullanılır[6].

$$mel(f) = 2595.log_{10}\,(1 + \frac{f}{700}) \qquad (3.5)$$

Mel spektrumunu ifade etmek için kullanılan yaklaşımlardan biri, her bir mel frekans bileşeni için bir filtre kullanmaktır.(şekil 3.51)



Şekil 3.51 Mel Ölçeği

Bu filtre bankası üçgen bant geçiren frekans karakteristiğine sahiptir ve bant genişliği sabit mel frekansı aralıklarıyla belirlenir. İç içe geçmiş üçgen filtrelerden her birinin kesim frekansı, komşu iki filtre tarafından belirlenir. Filtre bankası 13 doğrusal aralıklı (133Hz ile 1kHz arası), 27 logaritmik aralıklı (frekanslar 1.0711703 çarpanı ile aralıklandırılır) filtreden oluşur [5]. Kulağın duyma hassasiyetini modellemek amacıyla bu yapı kullanılmıştır. Mel filtre bankasında istenilen frekans aralığındaki (örnekleme frekansına bağlıdır) frekans değerlerine karşılık gelen ağırlık değerlerinin hesaplanmasında üçgen benzerliğinden yararlanılır.

3.6 Ayrık Kosinüs Dönüşümü (DCT)

'Ses Etiketleme' projesinin en önemli basamağı olan Mel Frekansı Sepstral Katsayıları (MFCC) bulunması işleminin son aşamasında, mel spektrumunun logaritması zaman bölgesine çevrilir. Bu zaman bölgesine çevirme işlemi ayrık kosinüs dönüşümü (DCT) yardımıyla yapılmaktadır.[5] Son aşamanın sonucu olan mel güç spektrumu katsayılarını Yk, k=1,2,…,K ile gösterirsek, MFCC'ler (c_y) aşağıdaki şekilde hesaplanabilir[4].

$$c_y(n) = u_n \sum_{k=0}^{K-1} (logY_K)\cos(\frac{(2k+1)n\pi}{2K})$$

$$u_n = \sqrt{\frac{1}{K}} \;\; ; \;\; n = 0 \qquad u_n = \sqrt{\frac{2}{K}} \;\; ; \;\; n > 0$$

n = mel frekansı kepstral katsayısı indeksi
k = mel filtresi indeksi
K = toplam mel filtresi sayısı

'Ses Etiketleme' projesi kapsamında öncelikle müzik ve insan sesinden oluşan 1'er saniyelik sekiz örnek alınarak bu örneklerin MFCC'leri hesaplanmış ve veri bankasına kayıt edilmiştir. Bu işlemden sonra karar kuralı oluşturmak için elde edilen dört insan sesi MFCC'sinin ortalaması alınmış aynı işlem müzik örneklerine de uygulanmıştır.

3.7 İlişkilendirme

Bu aşamada etiketlenmesi amaçlanan ses dosyası, veri bankasında bulunan örneklerle karşılaştırılarak dosyanın her saniyesi için veri bankasındaki müzik ve insan sesi dosyalarıyla olan ilişkisi hesaplanmıştır. Bu ilişkiyi hesaplamak için MATLAB programında corrcoef(X,Y) fonksiyonu kullanılmıştır.

```
S =

    1.0000    0.6181
    0.6181    1.0000


M =

    1.0000    0.0168
    0.0168    1.0000
```

Şekil 3.7 Örnek İlişkilendirme Değerleri

Karar vermesinde kullanılacak olan örnek çıktı şekil 3.7 deki gibidir.

3.8 Karar Verme

Karar verme işleminde, etiketlenmek istenen dosya saniye saniye taranarak MFCC'leri hesaplanıp, her saniye için insan sesi ve müzik etiketlemesi yapılması amaçlanmıştır. Bu doğrultuda ilişkilendirme sonuçlarından yararlanan ayırma fonksiyonu elde edilmiştir.

Bu fonksiyon kullanılarak müzik ve insan sesi etiketleri 0 ve 1'ler olarak kodlanmış sonrasında ise. txt dosyasına şekil 3.8 de görüldüğü gibi kaydedilmiştir.

```
[0 - 1 sn. ] -  Müzik

[1 - 2 sn. ] -  Ses

[2 - 3 sn. ] -  Ses

[3 - 4 sn. ] -  Müzik
```

Şekil 3.8 Örnek .txt Çıktısı

4. Sonuç

Yapılan bu çalışmada belirli bir ses dosyasındaki insan sesi ve müziğin etiketlenmesi amacı doğrultusunda ses dosyasından alınan insan sesi ve müzik örnekleri, mel frekansı sepstral katsayıları ile çarpılarak öz vektörler elde edilmiştir. Elde edilen bu öz vektörler veri bankasına eklenmiş, eklenen bu öz vektörler karar kuralı oluşturmak için etiketleme aşamasında kullanılmıştır.

5. Kaynaklar

[1] Ericsson, L., 2009, Automatic speech/music discrimination in audio files, Master's thesis in Music Acoustics, School of Media Tecnology, Royal Institute of Tecnology, Sweden.

[2] Do, M.N., An Automatic Speaker Recognition System, Audio Visual Communications Laboratory, Swiss Federal Institute of Tecnology, Lausanne, Switzerland.

[3] Kayran, A., 1990, Sayısal İşaret İşleme, Elektrik – Elektronik Fakültesi, İstanbul Teknik Üniversitesi, İstanbul, p161-163.

[4] Milner, B., Shao, X., Speech Reconstruction from Mel-Frequency Cepstralm Coefficients Using a Source-Filter Model, University of East Anglia,Norwich, UK.

[5] Slaney, M.,1998, Auditory Toolbox, Interval Research Corporation, http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/

[6] Zhang, Yu., 2003, Mel-spectrum computation. Seminar Speech Recognition. http://www.liacs.nl/~erwin/SR2003/Students/04_Melspectrum%20Computation.ppt

[7] Wildermoth, B.R., 2001, Text Independent Speaker Recognition Using Source Based Features. Yüksek Mühendislik Tezi, Griffith University, Australia.