İÇERİK TABANLI SORGU VE TARAMA İÇİN YAPISAL VE ANLAMSAL SES İÇERİK ANALİZİ

Mustafa SERT

DOKTORA TEZİ
ELEKTRONİK BİLGİSAYAR EĞİTİMİ

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TEMMUZ 2006
ANKARA

Mustafa SERT tarafından hazırlanan İÇERİK TABANLI SORGU VE TARAMA İÇİN YAPISAL VE ANLAMSAL SES İÇERİK ANALİZİ adlı bu tezin Doktora tezi olarak uygun olduğunu onaylarım.

Prof. Dr. Buyurman BAYKAL
Tez Yöneticisi

Bu çalışma, jürimiz tarafından Elektronik Bilgisayar Eğitimi Anabilim Dalında Doktora tezi olarak kabul edilmiştir.

Başkan: : Prof. Dr. İnan GÜLER

Üye : Prof. Dr. Buyurman BAYKAL

Üye : Prof. Dr. Adnan YAZICI

Üye : Prof. Dr. Ömer Faruk BAY

Üye : Doç. Dr. Gözde Bozdağı AKAR

Tarih : 05/07/2006

Bu tez, Gazi Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygundur.

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada orijinal olmayan her türlü kaynağa eksiksiz atıf yapıldığını bildiririm.

Mustafa SERT

İÇERİK TABANLI SORGU VE TARAMA İÇİN YAPISAL VE ANLAMSAL SES İÇERİK ANALİZİ

(Doktora Tezi)

Mustafa SERT

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Temmuz 2006

ÖZET

Otomatik ses içerik analizi, bilgisayar sistemlerinin çeşitli kullanımları için sayısal ses sinyallerinin içeriklerini anlamaya yönelik algoritmaların geliştirildiği genel bir araştırma alanıdır. Otomatik ses analizinin ana fonksiyonu, sayısal koleksiyonlarda gün geçtikçe sayısı artan ses verisinin daha iyi yönetilebilmesi için, ses sinyallerinden bazı önemli bilgileri çıkartmaktır. Ses verilerinin içerik tabanlı gösterimi, dizinlenmesi, taranması ve otomatik olarak etiketlenmesi gibi birçok uygulama bu bilgilerden yararlanmaktadır.

Bu çalışmada, içerik tabanlı ses yönetim sistemlerinde büyük öneme sahip olan iki konu araştırılmıştır. İlk olarak, ses verileri içerisinde tekrar eden desenlerin tespit edilebilmesi için yeni bir yapısal analiz yöntemi önerilmiştir. Bu desenler, ses verilerinin içerikleri hakkında önemli bilgiler vermektedir. Bu bilgilere örnek olarak, bir müzik içerisindeki nakarat bölümleri ya da konuşma verisinin konusu hakkında anlamsal bilgiler gösterilebilir. Önerilen yöntem, bir ses sinyalinin içerisindeki en önemli bilgiyi çıkarabilmek için, MFCC özniteliğini ve MPEG-7 standardında tanımlı olan spektral düzlük özniteliğini kullanarak sinyal içerisindeki yapısal değişiklikleri tespit etmeyi amaçlamaktadır. Benzer yöntemlerden farklı olarak, resim işleme tekniklerinin ses içerik analizine uygulanabilirliği de araştırılmıştır. Önerilen yöntemin test edilebilmesi için

yaklaşık 5 saatlik ses kliplerinden oluşan bir veritabanı hazırlanmıştır. Deneysel sonuçlara göre, müzik ve konuşma verileri içerisindeki tekrar eden desenlerin sırasıyla %86 ve %87 doğruluk oranı ile tespit edilebildiği gözlenmiştir.

Bu tezde ayrıca, elde edilen yapısal analiz sonuçlarını kullanan ve ses verisinin çok farklı şekillerde sorgulanmasını ve taranmasını sağlayan bir model önerilmiştir. Bu sorgulara nakarat sorguları, ses efektlerinin sorgulanması, örnek vererek sorgulama (ÖVS) gibi sorgu biçimleri örnek olarak verilebilir. Son olarak, çokluortam bilgilerinin sorgulanması esnasında çok önemli bir yere sahip olan nokta, aralık ve en yakın k-komşuluk sorgu biçimleri de desteklenmektedir.

Bilim Kodu : **702.3.006**
Anahtar Kelimeler : **Ses bölütleme ve analiz, ses özeti, ses bilgi getirimi, ses spektrum düzlüğü, MPEG-7**
Sayfa Adedi : **64**
Tez Yöneticisi : **Prof. Dr. Buyurman BAYKAL**

STRUCTURAL AND SEMANTIC ANALYSIS OF AUDIO CONTENT FOR CONTENT-BASED QUERYING AND BROWSING

(Ph.D. Thesis)

Mustafa SERT

GAZİ UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY
July 2006

ABSTRACT

Automatic audio content analysis is a general research area in which algorithms are developed to allow computer systems to understand the content of digital audio information for further exploitations. The main task of automatic audio analysis is to discover some valuable structures of audio signals in order to facilitate a better handling of the current explosively expanding amounts of audio data available in digital collections. The practical applications such as automatic labeling, efficient indexing, browsing, or content-based retrieval of audio data benefit from these structures.

In this research work, our investigation relies on two areas that are particularly important for content-based audio management systems. Firstly, we propose a new method for structural analysis of audio signals in order to detect repetitive patterns that are suitable for content-based audio information retrieval systems. These patterns provide valuable information about the content of audio, such as a chorus or a key concept for music and speech, respectively. The proposed method aims to detect the structural changes in music and speech based on the Audio Spectrum Flatness (ASF) and the MFCC feature sets in order to provide a way to extract the most salient information of an audio signal. Contrary to existing approaches, we

consider the applicability of image processing techniques in audio content analysis. A database of approximately 5-hours of audio clip is prepared for the evaluation of the proposed approach. The experimental results demonstrate that, all the repetitive patterns and their locations are obtained with the highest recognition rates of 86% and 87% for music and speech, respectively.

In this thesis, we also present a framework for flexible querying and browsing of audio data, which benefits from the structural analysis results. The proposed framework provides a wide range of opportunities to query and browse an audio data by content, such as querying and browsing for a chorus section, querying by sound effects, and query-by-example (QBE). In addition, the clients can express their queries in the form of point, range, or k-nearest neighbor, which are particularly significant in the multimedia domain.

Science Code : 702.3.006
Key Words : Audio segmentation and analysis, audio summary, audio information retrieval, audio spectrum flatness, MPEG-7
Page Number: 64
Adviser : Prof. Dr. Buyurman BAYKAL

TEŞEKKÜR

Gazi Üniversitesi F.B.E. yönetmeliği gereği, resmi olarak ikinci danışmanım olamamasına rağmen, doktora çalışmam boyunca bana danışmanlık yapan, değerli yardım ve tecrübeleriyle bana sürekli yol gösteren Prof. Dr. Adnan YAZICI'ya sonsuz teşekkürler ederim. Sizin yardımlarınız olmadan bu tezi tamamlayamazdım.

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren ve kendisiyle çalışma fırsatı veren danışmanım Prof. Dr. Buyurman BAYKAL'a sonsuz teşekkürler ederim. Akademik etik ve çalışma disipliniyle ilgili olarak sizden çok şey öğrendim.

Doktora çalışmamı enstitü dışından bir öğretim üyesi ile birlikte yürütebilmem için bana her türlü kolaylığı gösteren ve tez izleme toplantılarında değerli görüş ve önerileriyle beni yönlendiren Prof. Dr. İnan GÜLER'e teşekkürü bir borç bilirim.

Doktora eğitimim boyunca ders ve tez çalışmalarımı yürütebilmem için bana izin veren ve gerekli kolaylığı gösteren Başkent Üniversitesi Mühendislik Fakültesi dekanımız Prof. Dr. İmdat KARA'ya teşekkürü bir borç bilirim.

Manevi destekleriyle beni hiçbir zaman yalnız bırakmayan anneme, babama, sevgili ağabeyim Ümit Sert'e ve kardeşim Mesut Sert'e sonsuz teşekkürler ederim. Sizin desteğiniz herzaman benimleydi. Oğlunuz ve kardeşiniz olmaktan onur duyuyorum.

Çalışmalarımın matematiksel gösterimlerinde bana yardımcı olan, umutsuzluğa düştüğüm anlarda hep yanımda olarak bana destek veren, çay ve kahve ikramları sayesinde güzel sohbetler etme fırsatı bulduğum çok değerli arkadaşım Dr. Belgin Özkul'a anlayış ve yardımlarından dolayı sonsuz teşekkürler ederim. İyi ki varsın.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

RESİMLERİN LİSTESİ

Resim **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklama**
sa	Saat
dk	Dakika
sn	Saniye
ms	Mili saniye
σ	Başlangıç benzerlik matrisi
$\bar{\sigma}$	Elemanları normalize edilmiş benzerlik matrisi
$\hat{\sigma}$	Köşegen çizgileri güçlendirilmiş benzerlik matrisi
$\breve{\sigma}$	Yatay - dikey çizgilerden arındırılmış benzerlik matrisi
lThr	Alt eşik değeri
uThr	Üst eşik değeri
f_s	Örneklem hızı
loFreq	MPEG-7 ASF analiz çerçevesinin alt frekans sınırı
hiFreq	MPEG-7 ASF analiz çerçevesinin üst frekans sınırı

Kısaltmalar	**Açıklama**
ASF	Audio Spectrum Flatness
DDL	Description Definition Language
MFCC	Mel Frequency Cepstral Coefficient
MPEG	Moving Pictures Expert Group
ÖVS	Örnek Vererek Sorgu
OST	Otomatik Ses Tanıma
XML	Extendible Markup Language

1. GİRİŞ

Bilgi teknolojilerindeki ilerlemeyle birlikte veritabanlarındaki, arşivlerdeki ve sayısal kütüphanelerdeki çokluortam verilerinin miktarı büyük bir hızla artmaktadır. Bunun sonucu olarak, istenilen veri içeriklerinin verimli şekilde getirimi ve yönetimi beraberinde birçok sorunu getirmektedir. Bu koşullar altında, çokluortam verilerinin işlenmesi ve otomatik içerik analizi yöntemleri büyük önem arzetmektedir. Bu noktada, içerik analizi; özellikle içerik anlama ve anlamsal bilgi çıkarma, çokluortam içeriklerinin verimli şekilde getirimi ve işlenmesini gerektiren uygulamalarda büyük öneme sahiptir.

İnsanoğlu, ses verilerinin algılanması için üstün bir yeteneğe sahiptir. Öyle ki, bir ses kaydının sadece bir bölümünü dinlemek suretiyle ses kaydının tipini ve içeriğini tanımlayabilmektedir. Bu özellik, insanoğlunun üstün algılama yeteneğini göstermektedir. Müzik verisi sözkonusu olduğunda, bu konuda uzman olmayan bir kişi bile ilgili müziği tanımlayabilmekte ya da bu müziği daha önceden işitip işitmediğini algılayabilmektedir. Konuşma verisi için de durum pek farklı değildir. Bir konuşma kaydını dinleyen bir kişi, bu konuşmanın içeriğini (konusunu) tanımlayabilmektedir. Diğer yandan, bir ses kaydı bilgisayar sistemlerinde bir dizi sayısal örnek değeri olarak saklanmaktadır. Bilgisayar sistemlerinde, ses verilerine erişim için kullanılan en ilkel yöntem, ses başlıkları ya da dosya isimlerini kullanmaktadır. Ancak, ses başlıklarının ve dosya isimlerinin genellikle eksik ya da öznel olmasından dolayı, bu yaklaşım uygulamaların erişim gereksinimlerini herzaman karşılayamamaktadır. Ayrıca, bu yaklaşım "*Verilen ses kaydına benzeyen ses kayıtlarını bul*" şeklindeki sorguları desteklememektedir. Çünkü bu tip sorgular tam uyuşmayı gerektirmeyen, yaklaşık (benzerlik) sorgu özelliklerini taşımaktadır.

Bu problemleri çözmek için, içerik-tabanlı ses getirim yöntemlerine ihtiyaç duyulmaktadır. İçerik tabanlı sorguların en basit biçimi, sorgu elemanı ile veritabanı elemanlarının örnek örnek karşılaştırılması prensibine dayanır. Ancak bu yöntem aynı ses kayıtları için bile doğru sonuçlar üretmeyebilir. Çünkü, ses kayıtları farklı üretim ortamlarında, farklı örneklem hızlarında ve farklı bit çözünürlüklerinde

üretilebilmektedir. Bu nedenle, içerik-tabanlı ses getirim sistemlerinde yaygın olarak kullanılan yöntem, ses verisinin tamamını ifade eden ve ses verisinin kendisinden elde edilen öznitelik bilgilerini kullanmaktır. Bu yöntemde, ses verisi öncelikle genel ses tiplerine (konuşma, müzik, gürültü) ayrıştırılmakta, daha sonra ayrıştırılan her bölüm farklı şekillerde dizinlenmekte ve işlenmektedir. Örneğin konuşma verisi için, çeşitli konuşma tanıma teknikleri kullanılarak konuşma içerisinde geçen kelimeler tespit edilmekte ve bu kelimelere göre dizinlenmektedir. Bu işlem, sorgu verisi için de uygulanmak suretiyle dizinlenen veriler üzerinden sorgular sonuçlandırılmaktadır. Benzer sorguların müzik verileri üzerinden de gerçekleştirilebilmesi için, müzik verilerine ait ayırtedici bilgilerin çıkarılarak dizinlenmesi gerekmektedir.

Bu tez çalışmasında esas olarak, içerik-tabanlı analiz çerçevesinde, şarkılar içerisindeki anlamsal bölümlerin tespit edilebilmesi amacıyla müzik yapısının analizi ele alınmıştır. Buna ek olarak, geliştirilen analiz yöntemi konuşma verilerine de uygulanmak suretiyle, konuşma verisinin içeriği (konusu) hakkında bilgi edinilmesi amaçlanmıştır. Müzik ve konuşma verisinin yapısındaki tekrarlar ve dönüşümler, ilgili verinin kimliği hakkında çok önemli tekil bilgiler üretebilmektedir. Bu bilgilerin tespit edilerek müzik ve konuşma verilerinin anlamsal seviyede açıklanması, ses verilerinin daha iyi ele alınmasına önemli katkıda bulunmaktadır. Bu tezde özellikle pop ve rak müzik türleri ele alınmıştır, diğer müzik türleri (etnik, klasik, caz, vb.) araştırma dışında tutulmuştur. Benzer çalışmalardan farklı olarak, notalı müzik verisi (MIDI) yerine ham ses verileri üzerinde çalışılmıştır. Bu tezde gerçekleştirilen çalışmanın sonuçları birçok uygulamada kullanılabilmektedir:

1. En önemli uygulama alanlarından birisi içerik tabanlı ses-getirim sistemleridir. Örneğin, müzik ya da konuşma verisi çıkarılan anlamsal bilgiler üzerinden sınıflandırılabilir ve yapısal açıklamalarına göre benzerlik sorguları gerçekleştirilebilir.
2. Tarama uygulamalarında ses verisi içerisindeki çeşitli bölümlere erişim. Çıkarılan üst seviyeli anlamsal bilgilerin ses verisi içerisinde ifade ettiği bölümlere (nakarat, kavram) hızlı biçimde ulaşılması. Örnek olarak bir ses çalıcı uygulamasında bir bölümden diğer bölüme hızlı biçimde geçiş gösterilebilir.

3. Orijinal ses verisinden çıkarılan tekrar eden bölüm bilgisi, bu ses verisine ait kısa bir özet olarak açıklanabilir ve ses parmakizi uygulamalarında kullanılabilir [1]. Bu özelliği ile ses verilerinin saklama karmaşıklığının azaltılmasına ve tekil açıklamalar oluşturulmasına katkıda bulunur.
4. Ses işaretleme araçları. Ses yapısal analiz yönteminin, ses işaretleme araçları ile uygun biçimde entegrasyonu ile kullanıcı için başlangıç işaretlemelerinin yapılabilmesini mümkün kılar.
5. Ses ipuçlarının kullanımı ile video dizinleme ve sınıflandırma.

Bu tezde, içerik-tabanlı ses getirim sistemlerinde çok önemli bir yere sahip olan müzik ve konuşma verisi içerisindeki tekrar eden desenlerin tespit edilebilmesi için yeni bir yöntem geliştirilmiştir. Çalışmanın ana katkıları aşağıdaki gibi sıralanabilir:

1. Konuşma ve müzik verileri içerisindeki anlamlı bilgilerin otomatik olarak tespit edilebilmesi için yeni bir algoritma önerilmiştir. Anlamsal bilgiler, konuşma verileri için *konuşmanın anahtar kavramı*, müzik verileri içinse *şarkının nakarat bölümü* olarak kabül edilmiştir. Geliştirilen algoritma üç aşamalı olup birinci aşamada, verilen bir benzerlik matrisi için alt ve üst eşik değerleri hesaplanmaktadır. İkinci aşamada, hesaplanan eşik değerleri kullanılarak benzerlik matrisindeki tekrar eden desenler yorumlanmaktadır. Son aşamada ise, k-means gruplama algoritması [2] kullanılarak yorumlanan desenler içerisinden en anlamlı olan grup seçilmektedir. Kullanılan boyut ve yer tabanlı gruplama, ses verisinin arama ve tarama verimliliğini önemli ölçüde artırmaktadır.
2. Benzer yaklaşımlardan farklı olarak, çeşitli imge işleme ve boyut indirgeme teknikleri ses içerik analizine uygulanmıştır. Böylece, hem tanıma doğruluğu artırılmış hem de hesaplamalardaki bellek karmaşıklığı azaltılmıştır.
3. Geliştirilen yöntem iki güçlü öznitelik (MPEG-7 Audio Spectrum Flatness – ASF ve Mel Frequency Cepstral Coefficients – MFCCs) ailesi için uygulanmak suretiyle, özniteliklerin konuşma ve müzik verisindeki tanıma oranları değerlendirilmiştir.
4. Geliştirilen yöntemin ürettiği sonuçlar, içerik-tabanlı bir ses getirim çatısı altında birleştirilerek, ses verilerinin esnek biçimde sorgulanmasını ve taranmasını sağlayan bir model önerilmiştir. Modelin desteklediği sorgu tiplerine *Örnek Vererek Sorgu*

(ÖVS), *nakarat* ve *ses efekt* sorguları örnek olarak gösterilebilir. Önerilen model, çokluortam alanında çok önemli bir yere sahip olan *nokta*, *aralık* ve *en yakın k-komşuluk* sorgu biçimlerini desteklemektedir.

Bu tezin diğer bölümleri şu şekilde düzenlenmiştir. Bölüm 2'de mevcut otomatik ses içerik analizi ve içerik-tabanlı ses getirim sistemleri incelenmiştir. Müzik ve konuşma verilerine ait anlamlı bölümlerin otomatik olarak çıkarılması için geliştirilen algoritma Bölüm 3'de açıklanmıştır. Bölüm 4'de ses verilerinin içerik-tabanlı sorgulanması ve taranması için önerilen model tanıtılmıştır. Son olarak, elde edilen sonuçlar ve gelecek çalışma planı Bölüm 5'de gösterilmiştir.

2. LİTERATÜR TARAMASI

Bu bölümde, tezin çalışma alanı üzerinde önceden yapılan araştırmalar incelenmiştir. İlk olarak, ses içerik analiziyle ilgili olan çokluortam içerik analizine genel bir bakış yapılmıştır. Daha sonra, ses içerik analizi konusunda gerçekleştirilen çalışmalar özetlenmiştir.

2.1 Çokluortam İçerik Analizi

Büyük miktarlardaki çokluortam verilerine sahip sayısal veritabanlarının verimli şekilde yönetimi ve bu verilerin içerik-tabanlı getirimi önemli bir araştırma konusudur. Geleneksel arama araçları genellikle metin tabanlıdır. Bu araçlar çokluortam verilerinin dosya ismi ya da el ile girilen açıklamaları üzerinden arama işlemlerini gerçekleştirmektedir. Ancak, bu açıklamalar anlamlı girilmediği takdirde arama araçları kullanışsız olmaktadır [3]. Bununla birlikte, çok büyük miktardaki çokluortam verisinin tamamı için ayrı ayrı metinsel açıklama girişi çok fazla emek gerektirmektedir. Bu kapsamda, çokluortam içerik analizinin amacı, çokluortam verilerinin içerikleri hakkında anlamlı ve ayırtedici bilgilerin otomatik olarak çıkarılması şeklinde açıklanabilir.

Çokluortam kavramı, sayısal içeriğin video, ses ve resim olmak üzere bütün türlerini kapsamaktadır. Karmaşık bir yapıya sahip olan bu veri türü, birçok alanda yaygın olarak kullanılmaktadır. Sayısal video verisi metin, resim ve ses bileşenlerinin tamamından meydana gelmektedir [4]. Günümüzde, içerik-tabanlı video ve resim analizi çokluortam verileri arasında en hızlı gelişen araştırma alanlarından birisi olmuştur. Analiz işlemlerinde renk, şekil, doku ve hareket gibi alt seviyeli öznitelikler kullanılarak video çerçeveleri tanımlanmış ve içeriği anlaşılmaya çalışılmıştır. Günümüzde video analiz araştırmalarında ses bileşeninden de yararlanılmaktadır. Böylece daha güçlü içerik açıklamalarının elde edilmesi amaçlanmaktadır. Ses verisinin video içerik analizindeki tipik uygulama alanlarından birisi, ses içerik bilgilerinin video sahneleri ile ilişkilendirilerek video bölütleme ve sınıflandırma aşamalarında kullanılmasıdır [4-7].

2.2 Ses İçerik Analizi

Çokluortam içerik analiz çalışmalarının büyük bir kısmını resim ve video verileri oluşturmaktadır [8-11]. Çokluortam veri tiplerinden birisi olan ses verisi, video ve resim bileşenleri kadar araştırılmamıştır. Önceki bölümde değinilen nedenlerden ötürü ses verisi önemli bir çokluortam bileşenidir ve verimli şekilde yönetilmesi ve işlenmesi için uygun yöntemlere ihtiyaç duyulmaktadır. Ses içerik analizi, içerik-tabanlı ses yönetim sistemlerinin bir alt dalıdır. Tipik bir içerik-tabanlı ses yönetim sistemi üç grupta incelenebilir: (1) Ses verisinin bölütlenmesi ve sınıflandırılması; (2) ses içerik analizi; (3) içerik-tabanlı ses getirimi.

Birinci gruptaki ses bölütleme çalışmalarının büyük kısmı öznitelik çıkarımı ile ilgilenir. Bu çalışmalar [12-14] üç alt grupta incelenebilir. Bunlardan birincisi, ses verisini önceden belirlenmiş sınıflara (müzik, konuşma, gürültü) göre bölütlere ayırır [12]. İkinci bölütleme yaklaşımında, bir ses verisinin ardışık çerçeveleri için çıkarılan öznitelik vektörleri arasındaki ani değişimler tespit edilmeye çalışılır ve ses verisi ani değişimlerin görüldüğü kısımlardan bölütlere ayrılır [13]. Son yaklaşımda ise, ses verisi içerisindeki sessizlik ve duraklama bilgileri üzerinden bölütleme işlemini gerçekleştirir [14]. Bu tezde kullanılan bölütleme yaklaşımı ikinci gruba girmektedir, yani, bir ses sinyalinin belirli uzunluktaki analiz çerçeveleri için hesaplanan öznitelik vektörleri arasındaki ani değişimler kullanılmaktadır. Bu yaklaşım, önerilen yapısal benzerlik analiz yönteminin ilk adımını oluşturmaktadır.

İkinci gruptaki ses içerik analizinin ana amacı, verimli dizinleme ve getirim uygulamaları için ses verisinin yapısal açıklamalarını elde etmekdir. Mevcut yapısal müzik analizi araştırmalarında iki farklı yaklaşım kullanmaktadır. Bu yaklaşımlar ses-sinyal ve sembolik-gösterim yaklaşımları olarak adlandırılmaktadır. Ses-sinyal yaklaşımı gerçek ham verisini girdi olarak kullanırken, sembolik-gösterim yaklaşımı müzik verisinin nota bilgisini (MIDI) girdi olarak kullanmaktadır. Anlamsal ses bölütleme [15], müzik pulu oluşturma [11, 17-19], müzik özet çıkarımı [20], nakarat tespiti [21-22] ve tekrar eden desen tespiti [23-26] gibi ses analiz çalışmalarının tamamının farklı isimleri olmasına rağmen, bu çalışmaların ortak amacı ses verisinin

verimli biçimde taranması ve getirilmesidir. Bu işlem, ses verisinden elde edilen birtakım bilgiler kullanılarak gerçekleştirilmektedir. Chai ve Vercoe, müzik verileri içerisinde tekrar eden desenleri otomatik olarak tespit edebilmek amacıyla bir yöntem önermiştir [17]. Bu yöntem, ses verisine ait yapısal bilgiyi "AABABA" biçiminde üretmektedir. Foote, ses sinyallerinin yapısal özelliklerini gösterebilmek için analitik bir yaklaşım kullanmıştır [35]. Bu çalışmada, ses verisinin genel yapısını gösteren ve "benzerlik matrisi" adı verilen bir gösterim önerilmiştir. Ancak her iki çalışma da [17, 35], müzik verisi içerisindeki nakarat bölümlerinin tespiti ele alınmamıştır. Bu motivasyonu temel alan bir çalışma Goto tarafından gerçekleştirilmiştir [21]. Goto, akustik öznitelikler kullanarak bütün nakarat bölümlerini ve yerlerini tespit etmeyi amaçlamıştır. Ancak, önerilen yöntem müzik verisi içerisindeki nakarat bölümlerinin akustik özellikleri hakkında ön bilgiye ihtiyaç duymakla birlikte, elde edilen doğruluk oranı %80 olarak rapor edilmiştir. Diğer yandan, yukarıda bahsedilen çalışmalardan hiçbirisinde konuşma verisi ele alınmamıştır. Bu tez çalışmasında, çeşitli imge işleme teknikleri ses içerik analizine uygulanmak suretiyle tanıma doğruluk oranı artırılmış ve konuşma verisi için kullanılan test verileri üzerinde anlamlı sonuçlar elde edilmiştir.

Son gruptaki içerik-tabanlı ses getirim çalışmalarında yaygın olarak kullanılan bir yaklaşım sorgu ve veritabanı elemanları için bir dizi akustik özniteliklerin çıkarılması prensibine dayanır. Bu motivasyonu temel alan çalışmalardan en önemlisi Muscle Fish çalışmasıdır [27]. Bu çalışmada, ses verisinden çıkarılan beş özniteliğe (gürlük, perde, parlaklık, band genişliği ve harmoniklik) ait istatiksel veriler (ortalama ve değişinti) kullanılmıştır. Ancak sadece istatiksel bilgiler kullanıldığından, önerilen yöntemin çoklu tınıya sahip ses verileri için uygun sonuçlar üretmediği bildirilmiştir [28].

Toparlamak gerekirse, Foote [35] 'nun önermiş olduğu benzerlik matrisi gösterimi, ses verisinin yapısına genel bir bakış sağlamaktadır. Bu özelliği sebebiyle, benzerlik matrisi gösterimi bu tezde temel alınmıştır. Ancak benzerlik matrisi gösterimi, ses verilerinden çeşitli anlamsal bilgilerin çıkarılabilmesi için tek başına yeterli değildir. Bu nedenle, çıkarılacak anlamsal bilgilerin özelliklerine uygun yöntemlerin benzerlik

matrisine uygulanması gerekmektedir. Bu motivasyon doğrultusunda Cooper ve Goto'nun gerçekleştirdiği ses özet çıkarımı ve nakarat tespiti çalışmalarında benzerlik matrisi gösterimi kullanılmıştır [16,21]. Ancak, elde edilen tanıma doğruluk oranları ve analiz esnasında ihtiyaç duyulan ön bilgiler nedeniyle bir takım problemlere sahiptir. Bu problemler iki grupta incelenebilir: (1) Ses sinyalinin ayrık zamandaki analiz çerçevelerinin benzerlik değerleri, birbirinden belirgin farklılıkları olsa bile genelde çok küçük ya da birbirine çok yakın olabilmektedir. Bu nedenle, ses sinyali içerisindeki yapısal özellikler ya da farklılıklar belirgin değildir. (2) Kullanılan eşik değerlerinin sabit ya da parametre olarak girilmesi ve birinci grupta özetlenen problemden dolayı, benzerlik matrisi içerisinde aranılan bölümlerin zamansal yerleri (başlama ve bitiş zamanları) nispeten yüksek sapmalarla ($\sim 3-4\,sn$) tespit edilebilmektedir. Bu durum, özellikle yüksek hassasiyet gerektiren tarama uygulamaları için çok uygun değildir.

Bu çalışmada uygulanan imge işleme teknikleri ile, benzerlik matrisi içerisinde aranılan desenler belirgin hale getirilmektedir. Bu desenlerin yerlerinin, benzerlik matrisi içerisinde doğru biçimde tespit edilebilmesi için uygun eşik değerlerinin kullanılması gerekmektedir. Ancak, bu eşik değerleri farklı ses verileri için farklı değerler alabilmektedir. Bu nedenle, uyarlanabilir eşik değerlerine ihtiyaç duyulmaktadır. Aksi takdirde, analiz edilen her ses verisi için eşik değerinin parametre olarak girilmesine ihtiyaç duyulacaktır. Bu çalışmada geliştirilen yöntem ile, her ses verisi için uyarlanabilir eşik değerleri elde edilebilmektedir [29]. Yukarıda özetlenen iki problem için geliştirilen çözüm, izleyen bölümde detaylı biçimde açıklanmıştır.

3. SES İÇERİK ANALİZİ

Ses içerik analizinde ilk adımı bölütleme işlemi oluşturmaktadır. Bu çalışmada kullanılan bölütleme yaklaşımı, belirli uzunluktaki ardışık çerçeveler için çıkarılan öznitelik vektörleri arasındaki ani değişimleri temel almaktadır. Bu nedenle, seçilen öznitelik ailesi analiz ve getirim işlemlerinde önemli rol oynamaktadır. Seçilen öznitelik nekadar güçlü ise; diğer bir deyişle farklı gürültü ve kayıt ortamlarında üretilmiş bir ses verisinin karakteristik özelliklerini ne kadar iyi temsil ediyorsa, sonuçlar da o oranda doğru olacaktır.

MFCC öznitelik ailesi, Otomatik Ses Tanıma (OST) birliği tarafından çok iyi bilinen ve konuşma tanıma uygulamalarında yaygın olarak kullanılan bir öznitelik olmasına rağmen, MPEG-7 ASF özniteliği, ses öznitelik ailesine oldukça yeni olup, MFCC özniteliği kadar uygulanmamıştır. Bu çalışmanın bir amacı da, MPEG-7 standardı tarafından güçlü bir ses tanıma özniteliği olarak duyurulan MPEG-7 ASF özniteliğinin otomatik ses içerik analizine uygulanarak MFCC özniteliği ile bir karşılaştırma ortamı sağlamaktır [30]. Bu nedenle, çalışmamızda MPEG-7 ASF ve MFCC öznitelik ailelerinin analiz altyapısında kullanılmasına karar verilmiştir.

İzleyen bölümlerde, bu özniteliklerin önerilen otomatik ses içerik analizindeki kullanımları açıklanmaktadır.

3.1 Öznitelik Çıkarımı

Alt-seviyeli öznitelikler, ses içerik analizi yöntemlerinin doğruluk oranlarını önemli ölçüde etkilemektedir. İzleyen iki alt bölümde, bu çalışmada kullanılan MPEG-7 ASF ve MFCC öznitelikleri tanıtılmış ve çıkarım yöntemleri açıklanmıştır.

3.1.1 MPEG-7 standardı ve ASF özniteliği

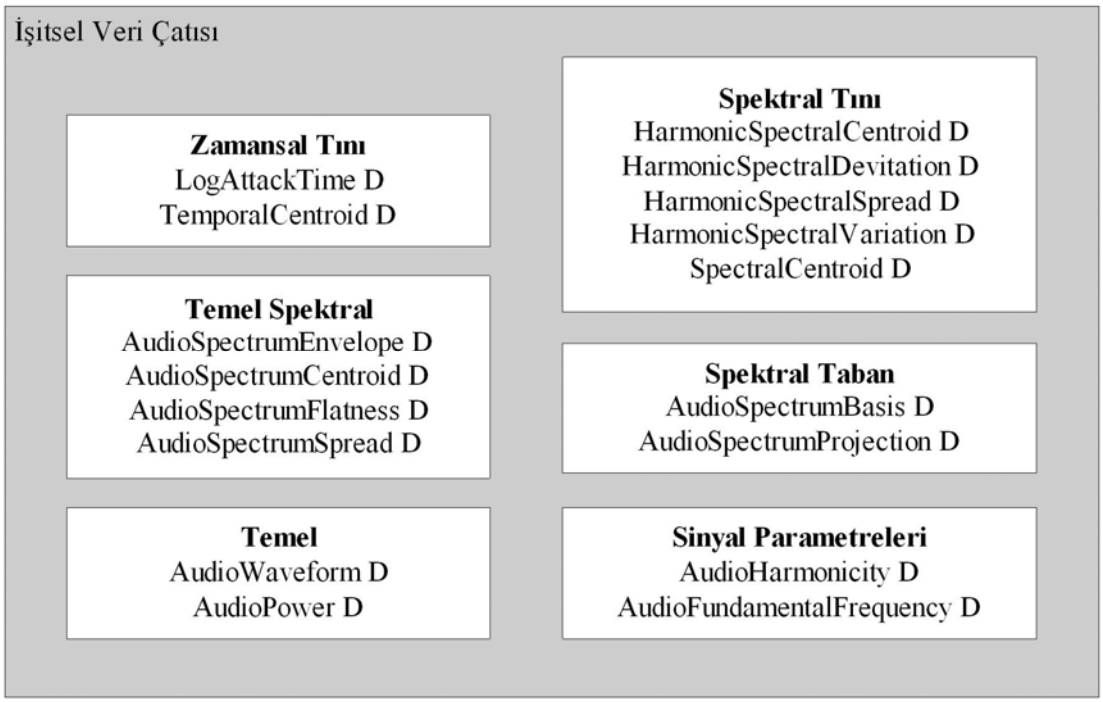
MPEG-7 ASF özniteliği, MPEG-7 standardının dördüncü bölümü olan ses bölümüne ait alt-seviyeli bir özniteliktir. Çokluortam içerik tanımlama arayüzü olarak bilinen

MPEG-7 standardı, çokluortam içeriğine ait özniteliklerin açıklanması için bir üstveri sistemi sağlayarak, çok geniş yelpazedeki çokluortam uygulamalarını ve gereksinimlerini adreslemektedir. Bu standartta, görsel ve işitsel verilerin dizinlenmesini, getirilmesini ve dünya genelinde birlikte çalışabilirliğini sağlayabilmek amacıyla bir dizi tanımlayıcı ve tanımlayıcı araçlar tanıtılmaktadır [30]. Standart içerisinde ayrıca, tanımlayıcı ve tanımlayıcı araçları genişletebilmek amacıyla *Description Definition Language* (DDL) olarak adlandırılan XML tabanlı bir dil önerilmiştir. Tanımlayıcılar, bir ses verisinden elde edilen alt- ve üst-seviyeli öznitelikleri açıklamakta kullanılırken, tanımlayıcı araçlar bir dizi tanımlayıcıyı ve aralarındaki ilişki(leri) açıklamak amacıyla kullanılmaktadır. Renk, hareket ve ses enerjisi gibi alt-seviyeli özniteliklere karşılık gelen bir çok tanımlayıcı otomatik olarak elde edilebilmektedir. Bu çalışmanın dışında kalan üst-seviyeli özniteliklerin çıkarılması için insan yardımı gerekirken, genellikle yarı otomatik olarak elde edilebilmektedir.

MPEG-7 standardında tanımlı olan teknolojiler belirli bir uygulamayı hedeflemeyip, aksine, mümkün olduğunca geniş bir yelpazadeki uygulama alanlarını desteklemeyi amaçlamaktadır. MPEG-7 standardı toplam sekiz bölümden oluşmaktadır ve bu bölümlerden dördüncüsü ses verisi için önerilen tanımlayıcı ve tanımlayıcı araçları anlatmaktadır. İşitsel veri çatısı, yüksek seviyeli işitsel veri uygulamalarının oluşturulmasında kullanılabilecek alt-seviyeli tanımlayıcıları içermektedir. Standartda tanımlı bu tanımlayıcıların kullanılmasıyla, uygulamaların birlikte işlerliği hedeflenmektedir, diğer bir deyişle, her uygulamanın işitsel verileri depolama şekli aynı olacaktır.

MPEG-7 standardının işitsel veri çatısında, üst seviyeli uygulamalarda kullanılmak üzere toplam on yedi adet alt-seviyeli tanımlayıcı bulunmaktadır (Şekil 3.1). Zaman ya da frekans düzleminden çıkarılabilen bu öznitelikler toplam altı grupta incelenebilmektedir:

1. Temel tanımlayıcılar
2. Temel spektral tanımlayıcılar



Şekil 3.1. MPEG-7 işitsel veri çatısı – alt seviyeli tanımlayıcılar

3. Spektral taban tanımlayıcıları
4. Sinyal parametre tanımlayıcıları
5. Zamansal tını tanımlayıcıları
6. Spektral tını tanımlayıcıları

İşitsel veri çatısında, bir ses verisinin alt-seviyeli özniteliklerle ilişkilendirilmesinin iki yolu vardır. Bu yollardan birisi, özniteliklerin sabit ses aralıklarından çıkarılması, diğeri ise ses verisinin farklı bölütlerinden çıkarılmasıdır. Bu öznitelikler aşağıda kısaca açıklanmıştır. Bu çalışmada kullanılan MPEG-7 ASF özniteliği ayrı bir paragrafta detaylı biçimde açıklanacaktır. *Temel tanımlayıcılar*: Bu sınıfta iki tanımlayıcı bulunmaktadır: *AudioWaveform* ve *AudioPower*. *AudioWaveform* tanımlayıcısı, örneklem periyodu içerisindeki bir ses sinyalinin maksimum ve minimum örnek değerini (genlik) açıklamaktadır. Bu tanımlayıcı, bir ses sinyalininin dalga biçimini çizdirirken kullanışlı olmaktadır. *AudioPower* tanımlayıcısı ise ses verisinin gücünü açıklamaktadır. Bu güç, seçilen çerçevedeki bir örneklemin anlık gücünü temsil etmektedir.

Temel spektral tanımlayıcılar: Bu sınıftaki tanımlayıcılar frekans düzleminden çıkarılmaktadır. Spektral tanımlayıcılar, insan kulağına en yakın gösterimleri temsil etmektedir ve bu nedenle çok önemlidir. Bu sınıfta toplam dört adet tanımlayıcı bulunmaktadır: *AudioSpectrumEnvelope*, *AudioSpectrumFlatness*, *AudioSpectrum-Spread* ve *AudioSpectrumCentroid*. *AudioSpectrumEnvelope* tanımlayıcısı spektogram görüntülemek ya da arama ve karşılaştırmalarda genel amaçlı tanımlayıcı olarak kullanılabilmektedir. Diğer üç tanımlayıcı ise *AudioSpectrumEnvelope* özniteliğini tamamlayan özniteliklerdir. *AudioSpectrumCentroid* tanımlayıcısı *AudioSpectrum-Envelope* özniteliğinin kütle merkezini açıklamaktadır ve frekans değerlerinin yüksek ya da düşük frekanslardan oluşup oluşmadığını göstermektedir. *AudioSpectrumSpread* tanımlayıcısı, güç spektrumunun merkez çevresinde ya da bütün spektrum boyunca yayıldığı bilgisini göstermektedir. Bu tanımlayıcı, ton benzeri ya da gürültü benzeri sesleri ayırtetmekte kullanılabilmektedir. *AudioSpectrumFlatness* tanımlayıcısı, bir ses sinyalini kısa süreli bölütler halinde ifade eder ve her bölütün frekans düzlemindeki karşılığının düzlük özelliklerini tanımlar. Bu öznitelik, *AudioSpectrumEnvelope* ile çıkarılan her bant için hesaplanır. Bu tanımlayıcı, ses çiftlerinin etkin bir şekilde karşılaştırılması için kullanılabilmektedir. Örnek uygulama olarak ses parmakizi gösterilebilir.

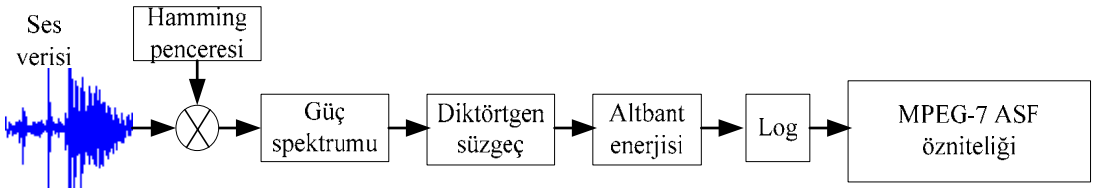
Spektral taban tanımlayıcılar: Bu sınıftaki *AudioSpectrumBasis* ve *AudioSpectrum-Projection* tanımlayıcıları, bir ses sinyalinin yüksek boyutlu spektrumunu, daha küçük boyutlu bir spektruma dönüştürmeyi sağlamaktadır. Bu özellik, etkin sınıflandırma algoritmalarının kullanılabilmesini mümkün kılmaktadır. *AudioSpectrumBasis* tanımlayıcısı, yüksek boyutlu spektrumun açıklamalarını, küçük boyutlu hale dönüştürmede kullanılacak taban fonksiyonlarını içermektedir. *AudioSpectrumProjection* tanımlayıcısı ise, *AudioSpectrumBasis* taban fonksiyonlarının yansıtılması neticesinde elde edilen spektrumun küçük boyutlu halini temsil etmektedir.

Sinyal parametre tanımlayıcılar: Standartda tanımlı olan iki adet sinyal parametre tanımlayıcısı vardır. Bu tanımlayıcılar *AudioHarmonicity* ve *AudioFundamental-Frequency* olarak adlandırılmaktadır. *AudioFundamentalFrequency* tanımlayıcısı, bir

ses sinyalinin temel frekansını açıklamaktadır. *AudioFundamentalFrequency*, müzik ve konuşma perdelerinin öngörülmesinde etkili sonuçlar vermektedir. *AudioHarmonicity* tanımlayıcısı, bir ses sinyalinin harmoniklik derecesini açıklamaktadır. Bu açıklama, harmonik spektrumu olan ve olmayan sesleri ayırabilmeyi mümkün kılar (örneğin müzik sesi ve gürültü).

Zamansal tını tanımlayıcılar: Zamansal tını tanımlayıcıları genellikle bir ses bölütü içerisindeki sinyal zarfının parametrelerini açıklamak için kullanılmaktadır. Bu tanımlayıcılar *LogAttackTime* ve *TemporalCentroid* olarak adlandırılmaktadır. *LogAttackTime* tanımlayıcısı bir sesin yükseldiğini tahmin etmeye çalışmaktadır. Bu tanımlayıcı, sinyalin yükselmeye başlaması ile kararlı hale gelmesi arasındaki süreyi saniye cinsinden logaritmik (onluk tabanında) olarak açıklamaktadır. *TemporalCentroid* tanımlayıcısı, ses bölütünün uzunluğuna bağlı olarak enerjinin zaman içerisinde nerede yoğunlaştığını açıklamaktadır.

Spektral tını tanımlayıcılar: Bu sınıfta, standartda tanımlı beş tanımlayıcı bulunmaktadır. Bu tanımlayıcılar *HarmonicSpectralCentroid*, *HarmonicSpectral-Deviation*, *HarmonicSpectralSpread*, *HarmonicSpectralVariation* ve *Spectral-Centroid* olarak adlandırılmaktadır. Tanımlayıcıların hepsi seslerin doğrusal frekans uzayındaki zamansal karakteristiklerini açıklamaktadır. Bu nedenle, zamansal ve spektral tını tanımlayıcıları birlikte kullanıldığında, müzik enstrumanlarının tınısını açıklamakta etkili olmaktadır. *SpectralCentroid* tanımlayıcısı *AudioSpectrum-Centroid* tanımlayıcısı ile aynı olup tek farkı, *SpectralCentroid*' in doğrusal güç spektrumundan çıkarılıyor olmasıdır. *HarmonicSpectralCentroid* tanımlayıcısı, bir spektrumun harmonik tepelerinin genlik-ağırlıklı ortalamaları olarak açıklanmaktadır. Diğer merkez tanımlayıcılarına benzemekle birlikte tek farkı, müzik tonunun harmonik kısımlarına uygulanıyor olmasıdır. *HarmonicSpectralDeviation* tanımlayıcısı, logaritmik-genlik bileşenlerinin genel spektral zarftan olan spektral sapmayı açıklamaktadır. *HarmonicSpectralSpread* tanımlayıcısı, anlık *HarmonicSpectralCentroid* tarafından normalize edilmiş spektrumun harmonik tepelerinin genlik-ağırlıklı standart sapması olarak açıklanmaktadır.



Şekil 3.2. MPEG-7 ASF özniteliğinin çıkarım yöntemi

HarmonicSpectralVariation tanımlayıcısı, iki komşu çerçevenin harmonik tepelerinin genliği arasındaki normalize edilmiş ilintisi olarak açıklanmaktadır.

MPEG-7 ASF özniteliği, bir sinyalin spektral içeriği hakkında oldukça derlitoplu (kısa) bilgiler vermesinin yanısıra, hemen hemen insan kulağının logaritmik tepkisini yansıttığı için bu çalışmada tercih edilmiştir [31].

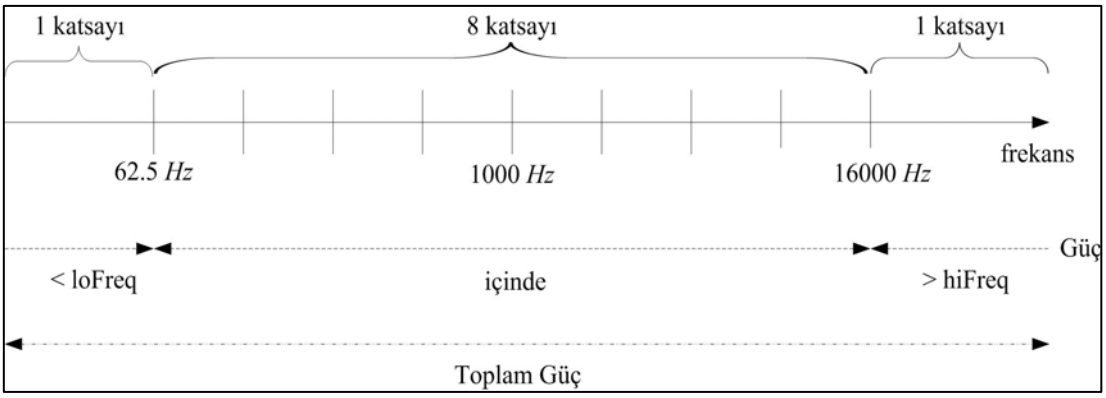
ASF özniteliği kabaca bir ses sinyalinin kısa zamanlı güç spektrumunun düzlük özelliğini açıklamaktadır. ASF özniteliğinin bir ses sinyalinden çıkarılmasını gösteren blok şema Şekil 3.2'de gösterilmiştir.

Bu çalışmada ASF özniteliği tek kanallı ses sinyallerinden çıkarılmaktadır. Bu nedenle, öznitelik çıkarımına başlamadan önce, iki kanallı bir ses sinyali tek kanallı bir dalga biçimine dönüştürülmektedir. Bu işlem, iki kanallı bir ses sinyalinin karşılık gelen bütün örneklem değelerinin aritmetik ortalaması ile hesaplanmıştır. Tek kanala dönüştürülen sayısal ses sinyali üzerine 30 *ms*'lik bir *Hamming* pencere fonksiyonu uygulanarak hızlı *Fourier* dönüşümü ile frekans düzlemine geçilir. Kullanılan Hamming penceresi,

$$\omega_i = \left(0,54 - 0,46 \cos\left(\frac{2\pi i}{N} \right) \right) \qquad 1 \le i \le N \tag{3.1}$$

fonksiyonu ile tanımlıdır. Burada *N*, *Hamming* penceresindeki örnek sayısını temsil etmektedir. Standart tarafından önerilen yöntem, ASF özniteliğinin güç spektrumundaki belirli frekans ve çözünürlük aralığından çıkarılması şeklindedir.

Standardın önermiş olduğu frekans aralığının alt ve üst sınırları sırasıyla $62,5$ *Hz* ve 16 *kHz* (işitme duyusunun üst sınırına karşılık gelmektedir), çözünürlük aralığı ise 0,0625 – 8 *oktav* arasındadır.



Şekil 3.3. MPEG-7 ASF özniteliğine ait alt-bantların elde ediliş biçimi

Bu parametrelere bağlı olarak, bir analiz çerçevesindeki alt-bant (katsayı) sayısı aşağıdaki biçimde tanımlanmaktadır:

$$Band \quad Sayısı = \frac{\log_2 \left(hiFreq / loFreq \right)}{octave \,\mathrm{Re}\, solution} \tag{3.2}$$

Burada *hiFreq* ve *loFreq* sırasıyla analiz edilen çerçevenin üst ve alt frekans sınırlarını, *octaveResolution* ise bant çözünürlüğünün logaritmik değerini temsil etmektedir. *loFreq* ve *hiFreq* arasındaki bütün bantlar eşit uzunluktadır. Eş. 3.2'den elde edilen bant sayılarına ek olarak fazladan iki katsayı daha kullanılmaktadır. Bu katsayılardan birincisi $0 - loFreq$ arasındaki enerjiyi, diğeri ise f_S örneklem hızı olmak üzere $hiFreq - f_s/2$ arasındaki enerjiyi temsil etmektedir (Şekil 3.3). Sonuç olarak, bir analiz çerçevesi için ASF öznitelik vektörü (V_i), her alt-bant için hesaplanan spektral gücün geometrik ortalamasının aritmetik ortalamasına oranı ile hesaplanmaktadır:

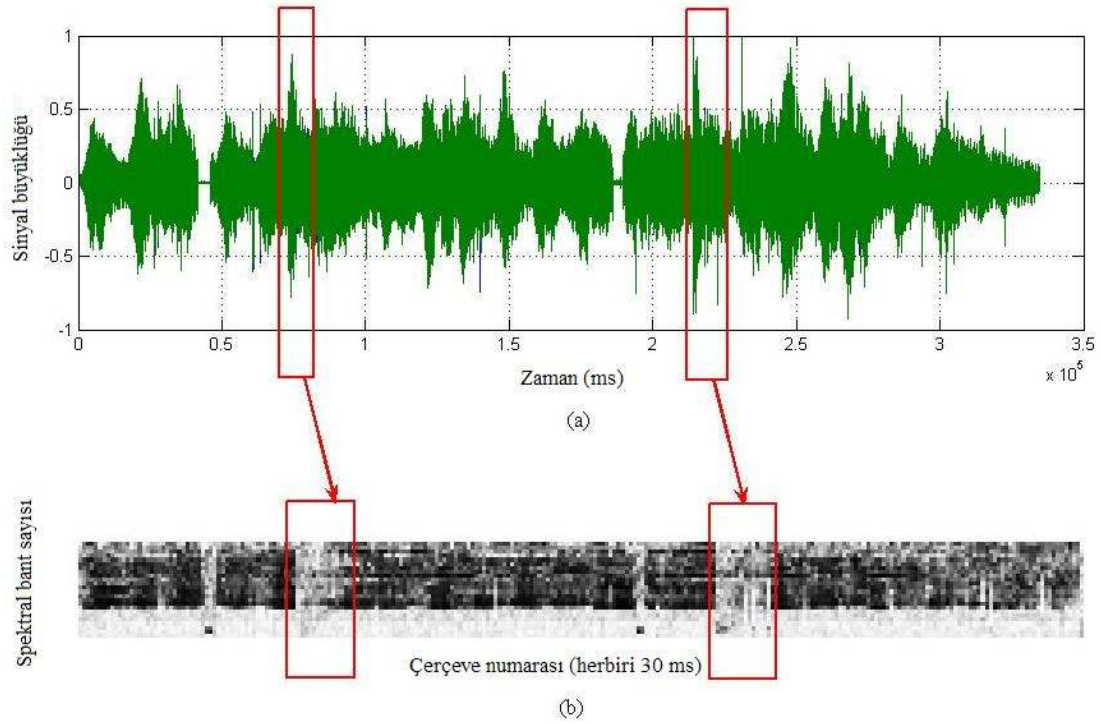
$$V_i = \frac{\sqrt[N]{\prod_i^N P_i}}{\left(\sum_i^N P_i\right)\Big/ N} \tag{3.3}$$

Burada P, her alt bandın güç spektrumunu, N ise her alt bandın uzunluğunu temsil etmektedir. Buna göre, bir ses sinyalinin bütün analiz çerçeveleri için elde edilen öznitelik bilgisi nxm boyutunda bir matristir. Burada n değeri, ses sinyalindeki analiz çerçeve sayısını, m ise analiz çerçevesinin logaritmik alt bant sayısını temsil etmektedir.

Bu çalışmada kullanılan analiz çerçeve boyutu 30 *ms* olarak seçilmiştir ve örtüşmeyen çerçeve kaymaları ile bütün ses sinyaline uygulanmaktadır. Bu nedenle, analiz öğe boyu 30 *ms*'dir. t bir ses sinyalinin süresini temsil etmek üzere, bu ses sinyali için hesaplanan ASF öznitelik matrisinin satır sayısı $n = (t/30\,ms)$ olmaktadır. Ayrıca, bu matrisin her satırı 30 ms'lik analiz çerçevesi için ASF öznitelik vektörünü temsil etmektedir. Sonuç olarak, bir ses sinyali için elde edilen ASF öznitelik matrisi aşağıdaki biçimde tanımlanmaktadır:

$$ASF = \begin{bmatrix} V_{1,1} & V_{1,2} & \cdots & V_{1,j} \\ V_{2,1} & \cdot & & \\ \cdots & & \cdot & \\ V_{i,1} & & & V_{i,j} \end{bmatrix} \qquad 1 \le i \le n,\ \ 1 \le j \le m \quad ve \quad n, m \in Z^+ \tag{3.4}$$

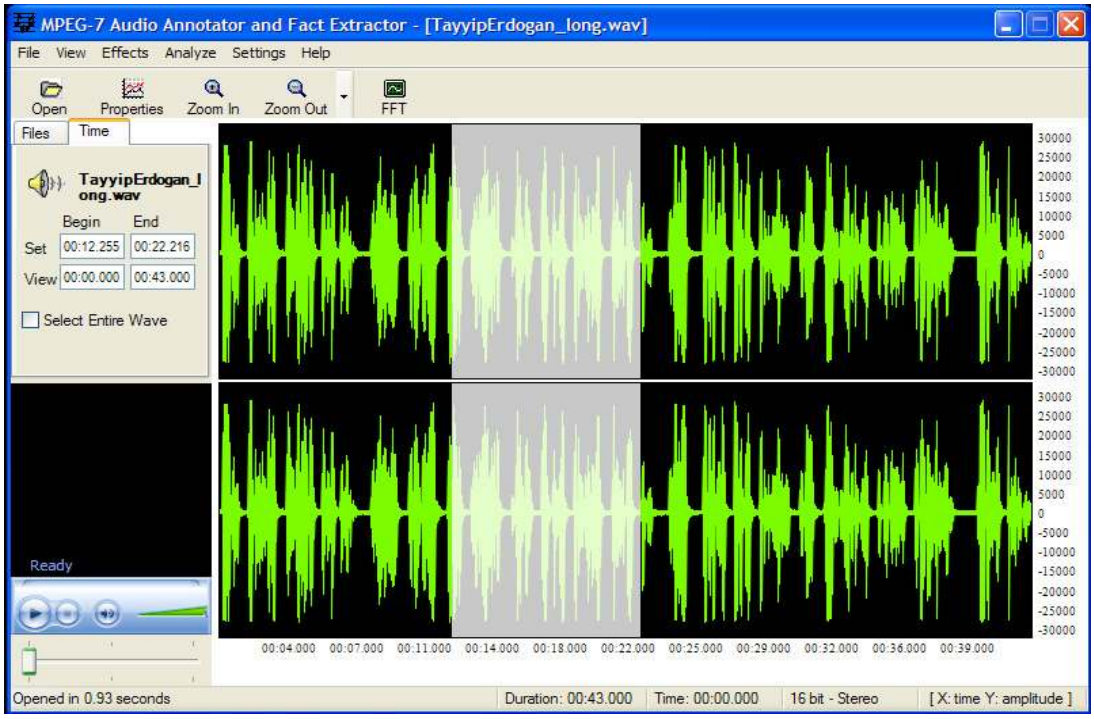
Bir ses sinyali için elde edilen ASF öznitelik matrisinin iki boyutlu görüntüsü Şekil 3.4'te verilmiştir. Şekil 3.4.a'da *Vivaldi*'nin "Four Season – Spring" eserinin dalga biçimi gösterilmektedir. Bu dalga biçiminin içerisine gerçek hayata uygun olması için rastgele pozisyonlara iki silah sesi efekti arkaplan sesi olarak eklenmiştir. Şekil 3.4.b, bu sinyal için elde edilmiş ASF öznitelik matrisinin iki boyutlu görüntüsünü göstermektedir. Eklenen arkaplan sesleri, ASF öznitelik matrisinin iki boyutlu görüntüsünde açık biçimde görülmektedir (Şekil 3.4.b). Bu durum, ASF öznitelik vektörünün ayırtedici özelliğini (gücünü) açık biçimde göstermektedir.



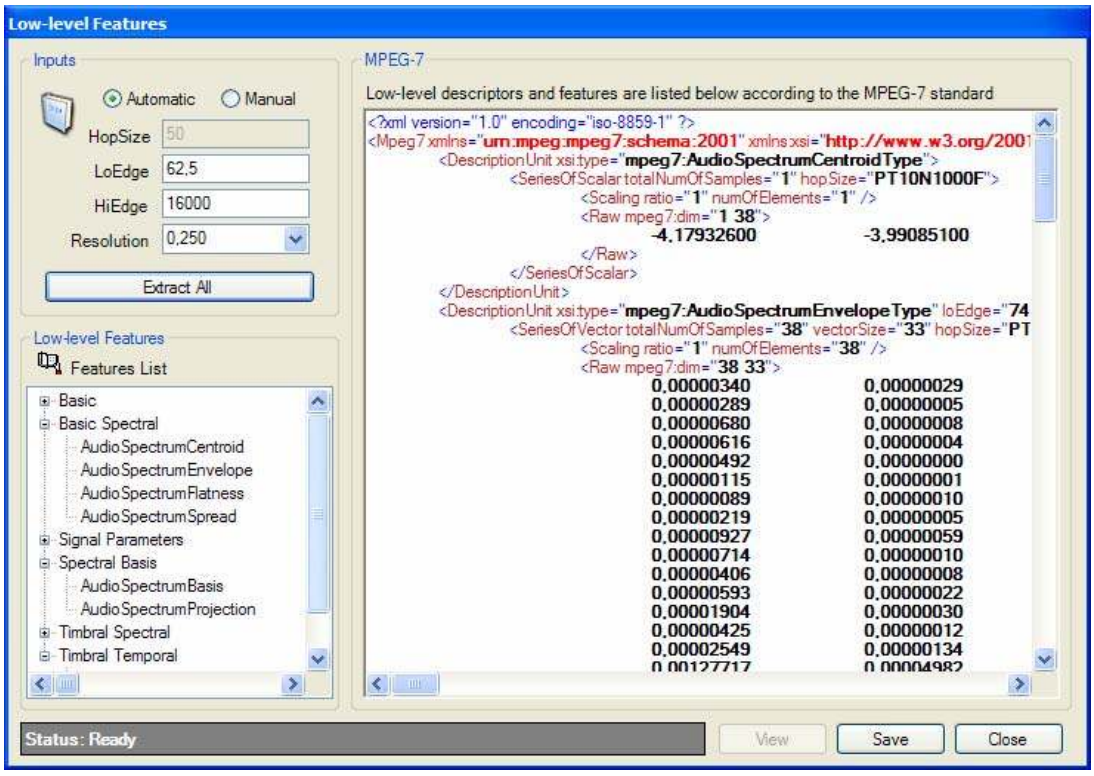
Şekil 3.4. *Vivaldi*'nin "Four Season – Spring" eseri (a) dalga biçimi (b) ASF özniteliğinin iki boyutlu görüntüsü

Tez çalışması boyunca, MPEG-7 ASF özniteliği de dahil olmak üzere, MPEG-7 işitsel veri çatısında tanımlı olan toplam on yedi adet alt-seviyeli özniteliğin çıkarılabilmesini sağlayan ses düzenleyici araç yazılımı geliştirilmiştir.

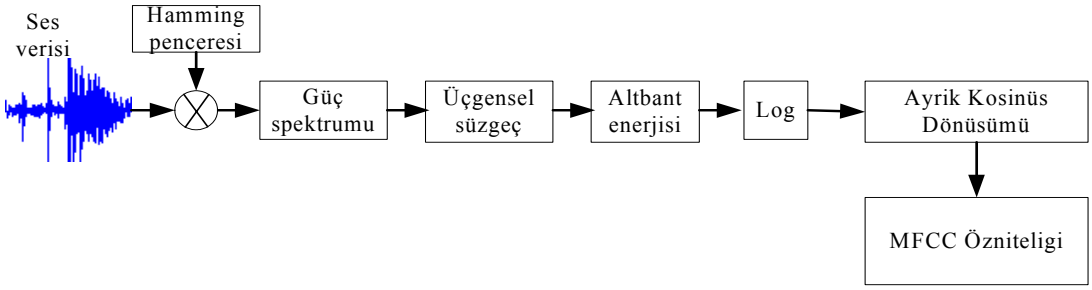
Resim 3.1 ve Resim 3.2'de ekran görüntüleri verilen ve nesne yönelimli tekniklerle geliştirilen araç yazılım, çok parçacıklı programlama mimarisine sahip olup, *wav* biçiminde saklanan ses verilerinin okunmasını, ekranda çizdirilmesini, üst verilerin (dosya boyutu, kanal sayısı, kodlama bilgileri, örneklem değeri, süre, yazar vb.) çıkarılmasını, zaman düzlemindeki örnek değerlerin elde edilmesini, hızlı Fourier dönüşümü (FFT) ile zaman düzlemindeki değerlerin frekans düzlemine dönüştürülmesini ve ses dosyalarının oynatılması gibi işlevleri desteklemektedir (Resim 3.1). Bu işlevlere ek olarak, araç yazılıma yüklenen bir ses dosyasının seçilen bir bölümüne ya da tamamına ait istenilen alt-seviyeli öznitelikler otomatik olarak çıkarılarak, XML tabanlı olan MPEG-7 DDL biçiminde açıklamaları üretilebilmektedir (Resim 3.2).



Resim 3.1. Ses düzenleme ve öznitelik çıkarıcı araç yazılımın ses düzenleme arayüzü



Resim 3.2. Ses düzenleme ve öznitelik çıkarıcı araç yazılımın MPEG-7 işitsel çatısı için alt-seviyeli öznitelik çıkarım arayüzü



Şekil 3.5. MFCC özniteliğinin çıkarım yöntemi

Önerilen otomatik ses içerik analiz yönteminde kullanılan MPEG-7 ASF özniteliği, geliştirilen ses düzenleme ve öznitelik çıkarıcı araç yazılımı vasıtasıyla elde edilmiştir.

3.1.2 MFCC özniteliği

MFCC özniteliği, ses spektrumunun kısa gösterimleri için kullanılan bir diğer öznitelik ailesidir. Bu öznitelik, frekans bantlarına MPEG-7 ASF özniteliğinden farklı olarak Mel ölçeği adı verilen bir yöntemle karar verir. Mel frekans ölçeğinde, 1 kHz'in altında doğrusal ölçek, 1 kHz'in üstünde logaritmik ölçek kullanılmaktadır. Bu özniteliğin en yaygın kullanım alanlarından birisi konuşma tanıma uygulamalarıdır. Bunun yanısıra, müzik yapısının tespit edildiği bazı uygulamalarda da MFCC özniteliği kullanılmaktadır [32, 33].

MFCC özniteliğinin çıkarım yöntemine ait blok şema Şekil 3.5'te gösterilmiştir. Genel çıkarım adımları aşağıda biçimde listelenebilir [34]:

1. Sinyal, analiz çerçevelerine ayrılır.
2. Her çerçeve için Fourier dönüşümü hesaplanır.
3. Frekans spektrumu log ölçeğine dönüştürülür.
4. Mel ölçeklendirmesi gerçekleştirilir.
5. Öznitelik katsayıları (MFCC katsayıları), ayrık kosinüs dönüşümü yöntemiyle elde edilir.

3.2 Yapısal Benzerlik Analizi

Müzik ve konuşma sinyalleri genellikle zaman içerisinde kendilerini tekrar eden bazı desenlere sahiptir. Müzik sinyali içerisindeki nakarat bölümleri ve konuşma sinyali içerisinde tekrar eden kelimelere ait sesler bu desenlere örnek olarak verilebilir. Konuşma verisi içerisinde birbirini tekrar eden birçok kelime bulunabilir. Bu kelimeler arasından doğru seçimler yapılabilirse, ilgili konuşmanın içeriği (konusu) hakkında anlamlı bilgiler elde edilebilir.

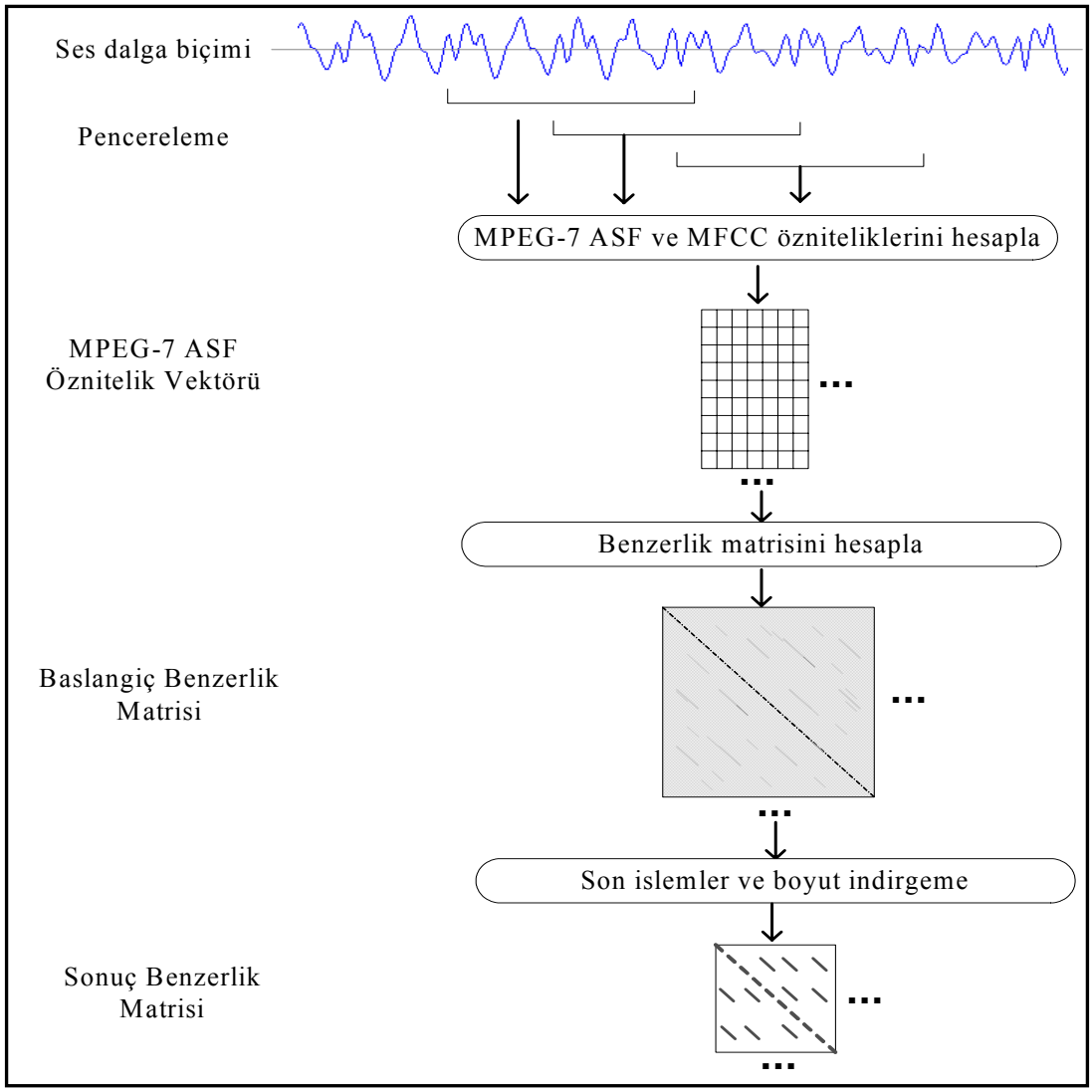
Bu çalışmada, sırasıyla konuşma ve müzik verilerinden otomatik olarak elde edilmeye çalışılan *anahtar kavram* ve *nakarat* kavramları aşağıdaki biçimde tanımlanmıştır.

Nakarat. Bir şarkı sözü içerisinde kendisini genellikle birden fazla kez tekrar eden ses bölümleri.

Anahtar kavram. Bir konuşma verisi içerisinde kendisini tekrar eden en uzun kelimelerin oluşturduğu ses bölümleri.

Müzik ve konuşma sinyallerinden elde edilen tekrar eden desen yapıları, ses bilgi geri getirim sistemleri için çok önemlidir. Örneğin bu analizin sonucu, ses getirim ve tarama uygulamalarında, ses içeriklerini dizinlemek amacıyla kullanılabilmektedir. Bu çalışmada önerilen yapısal benzerlik analiz yaklaşımına ait blok şema Şekil 3.6'da gösterilmiştir. Kullanılan yaklaşımın bölütleme kısmında, Wellhausen tarafından gerçekleştirilen zamansal bölütleme yöntemi temel alınmıştır [36]. Wellhausen, bölütleme yaklaşımı olarak bir müzik sinyalinin bütün analiz çerçeveleri için elde edilen MPEG-7 Audio Spectrum Envelope spektral öznitelik vektörleri arasındaki ani değişimlerin tespit edilmesi yaklaşımını kullanmıştır.

Bu çalışmada önerilen yapısal benzerlik analizi yöntemi, bir ses verisinin bütün analiz çerçeveleri için hesaplanan MPEG-7 ASF ya da MFCC öznitelik vektörleri arasındaki benzerlik değerleri ile "benzerlik matrisi" yapısını



Şekil 3.6. Yapısal benzerlik analizi yöntemi

oluşturmaktadır. Başlangıçta elde edilen benzerlik matrisi içerisindeki tekrar eden desenlerin köşegensel çizgiler biçiminde görünmesi beklenmektedir. Ancak, benzerlik matrisi içerisinde görünmesi beklenen köşegensel çizgiler matrisin bu halinde yeterince belirgin olmadığından tanıma doğruluk oranı da düşük olmaktadır.

Tanıma doğruluk oranını etkileyen en önemli özelliklerden birisi kullanılan öznitelik tipidir. Bu nedenle, güçlü ses özniteliklerinin kullanılması gerekmektedir. Önceki bölümlerde açıklanan nedenlerden dolayı, önerilen yöntemin öznitelik altyapısı olarak MPEG-7 ASF ve MFCC öznitelikleri kullanılmıştır. Tekrar eden desen

analizinde diğer önemli husus da, desenlerin orijinal ses verisi içerisindeki doğru yerlerinin tespit edilmesidir. Bu nedenle, benzerlik matrisi içerisinde tekrar eden desenleri temsil eden köşegensel çizgilerin sınırlarının net olması gerekmektedir. Başlangıçta elde edilen benzerlik matrisindeki bu sınırlar genellikle çok net olmadığından, çeşitli imge işleme teknikleri kullanılarak tekrar eden desenlere ait sınırların belirginleştirilmesine çalışılmıştır. Bir ses verisi içerisinde birbirini tekrar eden birden fazla grup desen olabilmektedir. Son olarak, bu gruplar arasından en anlamlı desen grubunu seçebilmek amacıyla k-*means* gruplama algoritması kullanılarak en anlamlı desen grubu seçilmektedir [37]. İzleyen bölümlerde bu adımlar ayrıntılı biçimde açıklanmıştır.

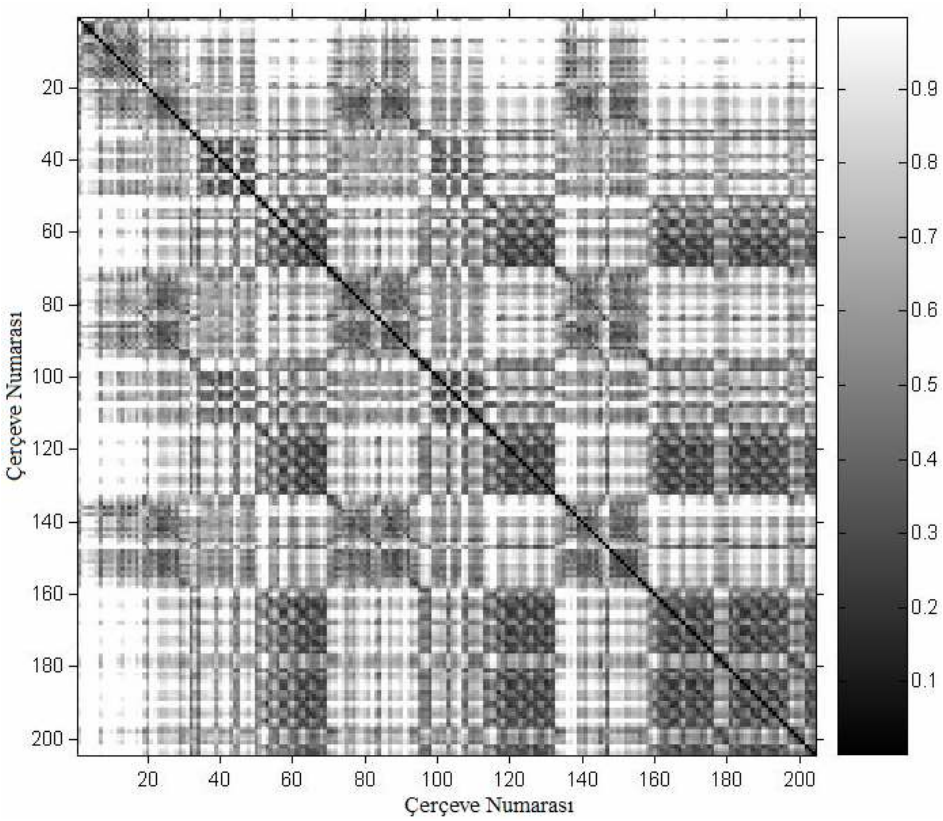
3.2.1 Benzerlik hesabı

Müzik ve konuşma verileri içerisindeki tekrar eden desenlerin tespit edilebilmesi için ilk adımı, Eş. 3.4'de verilen ASF öznitelik matrisi içerisindeki bütün öznitelik vektörleri arasındaki benzerliklerin hesaplanması oluşturmaktadır (ASF öznitelik matrisi içerisindeki her satırın, ses sinyali içerisindeki bir analiz çerçevesi için çıkarılan öznitelik vektörü olduğunu hatırlayınız). Buna göre, birbirine çok benzeyen (yakın olan) öznitelik vektörleri için çok küçük benzerlik değerleri elde edilecektir.

V_i ve V_j sırasıyla ASF benzerlik matrisi içerisindeki *i* ve *j*'nci çerçevelerin öznitelik vektörleri ise, bu vektörler arasındaki öklit uzaklığı aşağıdaki biçimde tanımlanır:

$$\sigma(V_i,V_j)=\sqrt{\sum_{k=1}^{m}\left(V_{i_k}-V_{j_k}\right)^2} \qquad \left(i,\ \ j,\ \ m\ \in\ Z^+\right) \tag{3.5}$$

Burada $\sigma(V_i,V_j)$ *i* ve *j* çerçevelerinin birbirlerine olan benzerliklerini ve *m* vektör uzunluklarını temsil etmektedir. Bu hesaplama, ses sinyalinin bütün çerçeveleri için hesaplanırsa, *n*x*n* boyutunda ana köşegene göre simetrik bir kare matris elde edilmektedir. Burada *n*, ses sinyali içerisindeki 30 *ms* uzunluğundaki analiz çerçevelerinin sayısını temsil etmektedir. Böyle bir benzerlik matrisinin iki boyutlu



Şekil 3.7. MPEG-7 ASF özniteliği ile hesaplanan başlangıç benzerlik matrisi (Beatles – *Lucy in the sky* şarkısı için benzerlik matrisi)

görüntüsü Şekil 3.7'de gösterilmiştir. Bu resimdeki (i, j) konumundaki bir piksel, ses sinyali içerisindeki i ve j'nci analiz çerçevesinin birbirine olan benzerlik değerini temsil etmektedir. Bu resim üzerinde benzerlik ne kadar fazla ise o oranda siyaha yakın bir piksel, benzerlik ne kadar az ise o oranda beyaza yakın bir piksel elde edilmektedir.

Şekil 3.7'den de görüldüğü üzere, benzerlik matrisinin bu biçimi, tekrar eden desenleri açık olarak ortaya çıkarmamaktadır. Çünkü, benzerlik hesabında kullanılan öznitelik vektör değerleri çok farklı iki analiz çerçevesini temsil etseler bile, birbirine çok yakın ve küçük değerler alabilmektedir. Bu nedenle, elde edilen benzerlik matrisine bir dizi imge işleme tekniklerinin uygulanması yerinde olacaktır. Bu işleme geçmeden önce, elde edilen benzerlik matrisinin boyutlarının çok büyük olması nedeniyle, önemli bilgileri kaybetmeden boyut indirgeme işlemi gerçekleştirilecektir.

Böylece, son işlemlerin gerektirdiği hesaplama ve bellek karmaşıklığı aza indirgenmiş olacaktır.

3.2.2 Boyut indirgeme

Eş. 3.4'te oluşturulan başlangıç benzerlik matrisi, bir ses sinyalinin 30 *ms*'lik analiz çerçeveleri için elde edilmiştir. Bir müzik ya da konuşma verisi içerisinde aranan desenler genellikle 30 *ms*'lik sıklıklarla tekrar etmemektedir ya da etseler bile önemli bir bilgiyi göstermemektedir. Bunun yanısıra, yukarıda da bahsedildiği üzere benzerlik matrislerin boyutları genellikle çok büyük olduğundan hem bellek hem de hesaplama karmaşıklığına neden olmaktadır. Örneğin, örnek değerleri 16-*bit* ile saklanan 1 dakikalık bir müzik verisinin benzerlik matrisi ile gösterimi için gereken bellek miktarı yaklaşık olarak 8 *MB* civarında olacaktır. Bu motivasyondan yola çıkarak, benzerlik matrisinin boyutunu önemli bilgileri kaybetmeden belirli oranda azaltabiliriz. Bunun için önerdiğimiz yöntem, *d*x*d* boyutundaki bir maskeyi benzerlik matrisinin üzerinde örtüşmeyen biçimde gezdirirken, bu maskenin benzerlik matrisi içerisinde karşılık gelen *d*x*d*'lik alan içerisindeki en küçük değeri alma prensibine dayanmaktadır. Bu noktada, en küçük değerlerin seçilmesi önemlidir, çünkü benzerlik matrisi içerisindeki en küçük değerler, analiz çerçeveleri arasındaki en büyük benzerlikleri temsil etmektedir. Bu işlem sonucunda elde edilen benzerlik matrisi $\tilde{\sigma}$ ile temsil edilirken yeni matris boyutu $\lfloor n/d \rfloor \mathrm{x} \lfloor n/d \rfloor$ olacaktır. Burada *d* değeri aynı zamanda sıkıştırma katsayısı olarak da tanımlanmaktadır.

3.2.3 İmge işleme tekniklerinin ses içerik analizine uygulanması

Tekrar eden desenlerin, başlangıçta oluşturulan benzerlik matrisi içerisinde -45^{o} yönünde köşegensel çizgiler şeklinde görülmesi beklenmektedir. Ancak, Şekil 3.7'de elde edilen benzerlik matrisinden bu desenleri elde etmek mümkün değildir. Çünkü, bu benzerlik matrisinde köşegensel çizgiler belirgin olmayıp, aynı zamanda yatay ve dikey çizgiler tarafından bozulmaya uğramıştır. Bu problemi çözmek için iki aşamalı bir yöntem önerilmektedir.

Birinci aşamada, öznitelik vektörleri arasındaki benzerlikleri belirgin hale getirebilmek için iki ardışık işlem uygulanmaktadır. Bu işlemlerden birincisinde, düzgün bir dağılım sağlayabilmek için, $\tilde{\sigma}$ benzerlik matrisinin benzerlik değerleri 0 ile 1 arasına normalize edilmektedir. Normalizasyon işlemi aşağıdaki biçimde tanımlıdır:

$$\bar{\sigma}_{x,y} = \frac{\tilde{\sigma}_{x,y} - \min(\tilde{\sigma})}{\max(\tilde{\sigma}) - \min(\tilde{\sigma})} \tag{3.6}$$

Burada x ve y sırasıyla benzerlik matrisinin satır ve sütun indislerini temsil ederken, $\bar{\sigma}_{x,y}$, (x, y) pozisyonundaki normalize edilmiş matris değerini göstermektedir. Normalize edilen benzerlik matrisine uygulanan ikinci işlem ise, benzerlik matrisinde tekrar eden köşegensel çizgileri güçlendirerek ortaya çıkarmayı amaçlamaktadır. Bu nedenle, köşegensel bir toplama işlemi, elemanları normalize edilen benzerlik matrisine uygulanmaktadır:

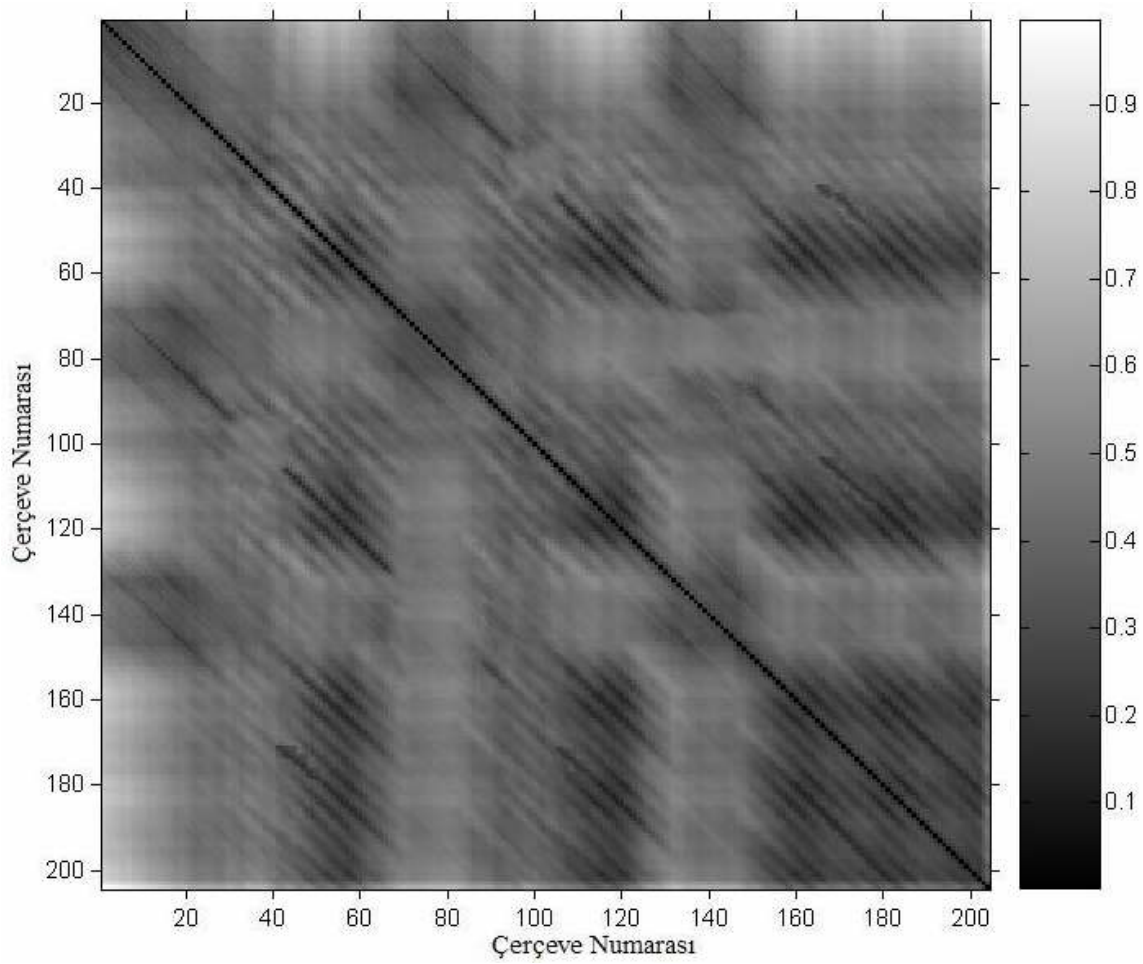
$$\hat{\sigma}_{x,y} = \sum_{i=1}^{k} \bar{\sigma}_{x+i,y+i} \tag{3.7}$$

Burada $\hat{\sigma}_{x,y}$, (x, y) pozisyonundaki pikselin güçlendirilmiş değerini, k ise güçlendirme parametresini göstermektedir. k parametresinin seçimi, aranılan desenin boyutuyla doğrudan ilişkili olduğu için önemlidir. Deneysel çalışmalarda, müzik ve konuşma verisi içerisindeki tekrar eden desenlerin tespit edilmesinde $k = \{5,6,7,8,9,10,11,12\}$ değerlerinin uygun olduğu gözlemlenmiştir. $k = 12$ için elde edilen benzerlik matrisinin iki boyutlu görüntüsü Şekil 3.8'de gösterilmektedir.

İkinci aşamada, -45^{o} yönündeki çizgilerin elde edilebilmesi için kenar tespit etme yöntemi kullanılmaktadır. *Kenar* kavramı, resim içerisindeki bir nesne ile zemin arasındaki sınır olarak tanımlanır ve gri tonlamalı bir resim için, gri seviyeli piksel değerlerindeki ani değişiklikler ile ifade edilir. Bu değişikliklerin tespit edilebilmesi

için kullanılan yöntemlere kenar tespit etme yöntemi adı verilmektedir. Kenar tespit etme yöntemleri, bir resim içerisindeki nesne sınırlarını ortaya çıkardığından, imge işleme alanında önemli bir yere sahiptir. Bu yöntemler, ön bilgi kullanan ve ön bilgi kullanmayan yöntemler olarak iki ayrı grupta incelenebilmektedir [38]. Kullanılan ön bilgiler uygulamaya göre değişiklik göstermekle birlikte, bu bilgilere tespit edilecek kenarların tonlama bilgisi, yapısı ve yönü örnek olarak verilebilir. Bu yöntemlerin dezavantajı sadece belirli özelliklerdeki resimler üzerinde kullanılabilmeleridir. Diğer yandan, ön bilgi gerektirmeyen yöntemler, belirli özellikteki resimlerle sınırlı olmadığından daha esnek bir kullanıma sahiptir ve yerel işleme prensibine dayalıdır, yani, bir pikselin kenar pikseli olup olmadığı, bu pikselin komşu piksellerine bağlıdır. Bu çerçevede Mohsen, kenar tespit etme yöntemlerini dört grupta incelemiştir [39]. Sürekli zamandaki bir resmin birinci dereceden türevi alındığında, kenarlar yönünde yerel maksimumlar elde edilecektir. Bu nedenle, ilk gruptaki yöntemler, birinci dereceden türev alma prensibine dayanmaktadır ve bu işlem gradyan işleçleri kullanılarak gerçekleştirilmektedir [40-42]. Gradyan işleçleri aynı zamanda maske olarak da adlandırılmaktadır. Bu gruptaki yöntemlerin avantajları arasında, hesaplama karmaşıklıklarının düşük olması, kenar ve bu kenarların yönlerinin tespit edilebilmesi gösterilebilir. İkinci gruptaki yöntemler, ikinci dereceden türev alma prensibine (Laplas işleci) dayanmaktadır [43]. Bu yaklaşımın en önemli dezavantajı, gürültüye karşı olan duyarlılığıdır. Üçüncü gruptaki yöntemlerde, Laplas işleci ile Gauss süzgeci yöntemi birleştirilmiştir [44]. Son gruptaki yöntemler ise, kenar bulma işleminden önce imge düzleştirme yaklaşımını kullanarak imge gürültüsünü azaltmayı hedeflemektedir. Bu yaklaşımı kullanan işleçlere sırasıyla Canny ve Shen-Castan yöntemleri örnek olarak gösterilebilir [43, 45]. Son iki gruptaki yöntemler, yüksek hesaplama karmaşıklığına sahiptir.

Bu çalışmada, ihtiyaç duyulan gereksinimleri karşılaması ve düşük hesaplama karmaşıklığına sahip olması nedeniyle, birinci gruptaki *maske* yöntemi tercih edilmiştir. Temel olarak bir maske, orijinal benzerlik matrisine kıyasla oldukça küçük olan bir kare matristir. Bu matrisin tipik boyutları uygulamaya bağlı olarak genellikle 3x3, 5x5, 7x7, vb. olabilmektedir. Bu çalışmada kullanılan maskenin boyutu 11 x 11 olarak belirlenmiştir. Nispeten daha büyük maske boyutunun



Şekil 3.8. Normalize edilmiş ve köşegensel çizgileri güçlendirilmiş benzerlik matrisi (Beatles – *Lucy in the sky* şarkısı için benzerlik matrisi)

$$maske = \begin{bmatrix} 10 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & 10 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 10 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 10 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 10 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & 10 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 10 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & 10 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 10 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 10 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 10 \end{bmatrix}_{11x11} \quad (3.8)$$

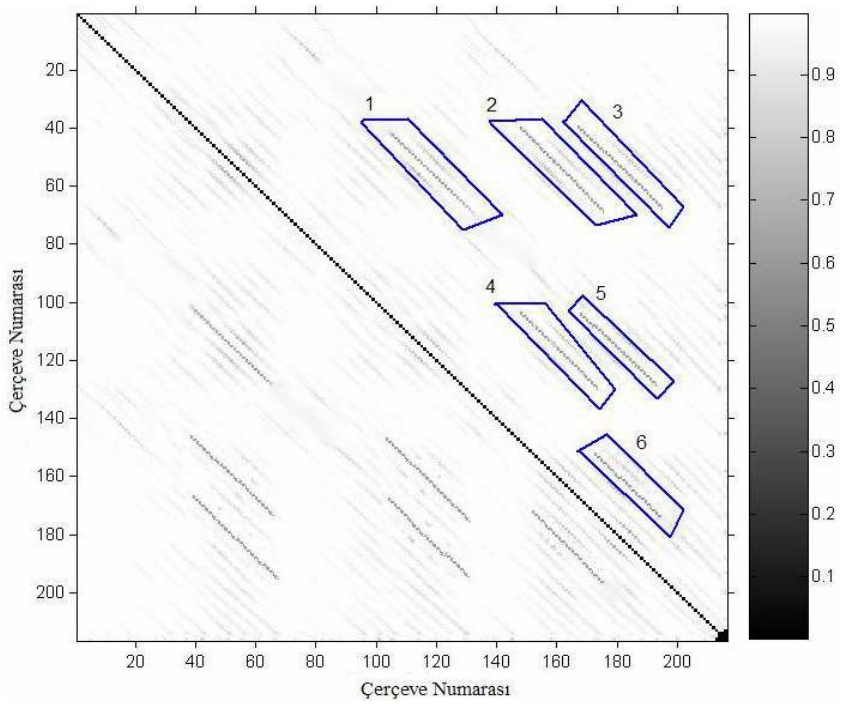
seçilmesiyle, aranılan kenar piksellerine daha çok vurgu yapılmaktadır. Bu maskenin ana köşegen elemanları 10 tamsayı değerlerinden oluşurken, diğer elemanlar -1 değerinden oluşmaktadır (Eş. 3.8). Maske tanımlamada dikkat edilmesi gereken nokta, maskeyi oluşturan değerlerin toplamının sıfır olmasıdır. Böylece, sabit bir zeminde gezdirilen bir maske hiçbir tepki vermeyecektir.

Tanımlanan bu maskenin merkez elemanı, $\hat{\sigma}$ benzerlik matrisinin $(0,0)$ koordinatından başlamak suretiyle bir piksellik kaymalar ile soldan-sağa ve yukarıdan-aşağıya doğru gezdirilmektedir. Her gezdirme işleminde, maske ile benzerlik matrisinin karşılık gelen alanı içerisindeki bütün elemanların çarpımlarının toplamı hesaplanır. Hesaplanan bu sonuç, maskenin merkez elemanının benzerlik matrisi içerisinde karşılık geldiği piksel koordinatına yazılır. Maskenin benzerlik matrisi içerisindeki bir piksele olan etkisi aşağıdaki şekilde tanımlanmaktadır:

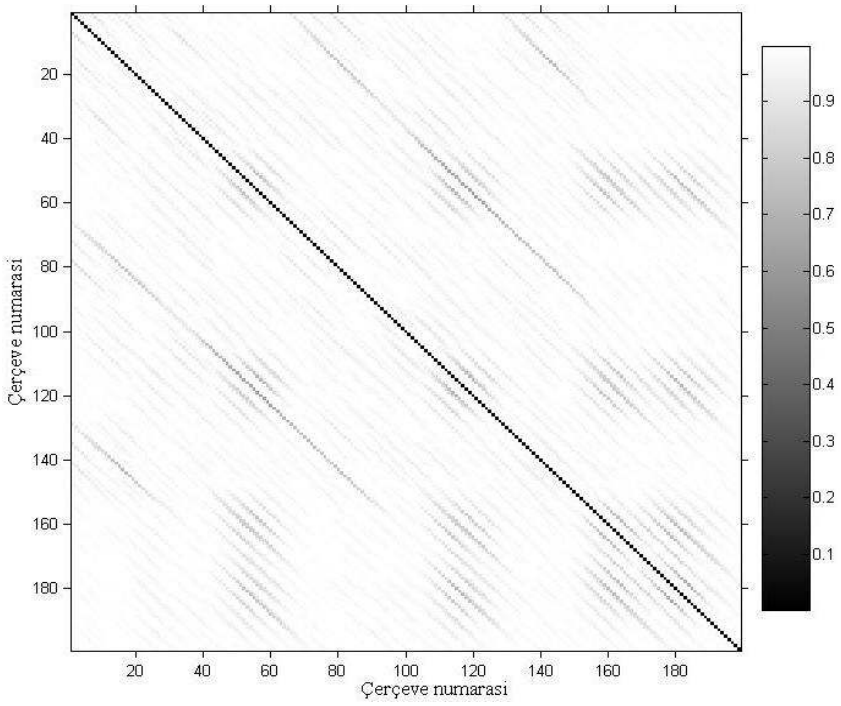
$$R = \sum_{i=1}^{b^2} \omega_i z_i \tag{3.9}$$

Burada, z_i benzerlik matrisi içerisinde ω_i maske elemanının karşılık geldiği benzerlik değerini, R herhangi bir noktadaki maske tepkisini, b ise maskenin satır ve sütün sayısının çarpımını temsil etmektedir. Bu işlem $\hat{\sigma}$ benzerlik matrisinin bütün elemanları için uygulandığında, -45^o yönündeki bütün çizgileri ortaya çıkarmaktadır. Bu işlem neticesinde oluşan sonuç benzerlik matrisi $\breve{\sigma}$ ile temsil edilmektedir ve iki boyutlu görüntüsü Şekil 3.9'da gösterilmiştir.

Aynı müzik dosyasının, MFCC özniteliği kullanıldığında elde edilen benzerlik matrisinin iki boyutlu görüntüsü Şekil 3.10'da gösterilmiştir. Şekil 3.10'dan da görüldüğü üzere, *The Beatles – Lucy in the sky* şarkısı için elde edilen benzerlik matrisi üzerindeki tekrar eden desenler, MPEG-7 ASF özniteliği için elde edilen benzerlik matrisi kadar açık değildir. Diğer bir deyişle, Şekil 3.9'da 1 numaralı etiket ile işaretlenmiş olan desen, Şekil 3.10'da daha uzun olup, nakarat başlama zamanı doğru biçimde tespit edilememektedir.



Şekil 3.9. İmge işleme teknikleri neticesinde MPEG-7 ASF özniteliği için elde edilmiş sonuç benzerlik matrisi (Beatles – *Lucy in the sky* şarkısı)



Şekil 3.10. İmge işleme teknikleri neticesinde MFCC özniteliği için elde edilmiş sonuç benzerlik matrisi (Beatles – *Lucy in the sky* şarkısı)

Bu durum, oluşturulan ses veritabanındaki diğer müzik sesleri için de benzerdir. Bu farklığın, seçilen öznitelik ailelerinin kullandığı frekans ölçeklerinden kaynaklandığı düşünülmektedir ve Bölüm 3.4'teki deneysel sonuçlar ve değerlendirmeler kısmında ayrıntılı biçimde irdelenmiştir. Sonuç olarak bu durum, kullanılan öznitelik ailesinin önemini sergilemektedir ve verilen örnekte (Şekil 3.9 ve Şekil 3.10) MPEG-7 ASF özniteliği, önerilen analiz yöntemi için daha doğru bir sonuç üretmektedir.

3.3 Anlamlı Desenlerin Çıkarılması

Şekil 3.9'dan görüldüğü üzere, bir ses sinyali içerisindeki birbirine benzer analiz çerçevelerinin oluşturduğu yapı, benzerlik matrisi içerisinde köşegensel çizgiler biçiminde görülmektedir. Bir sonraki adım, bu köşegensel çizgilerin yorumlanarak ses sinyali içerisindeki yerlerinin tespit edilmesidir.

3.3.1 Desenlerin yorumlanması

Şekil 3.9'da, Beatles'ın "*Lucy in the sky*" şarkısı için elde edilen altı adet tekrar eden desen görülmektedir. Bu desenler resim üzerinde çokgensel şekillerle işaretlenerek numaralandırılmıştır ve ilgili şarkı içerisindeki nakarat bölümlerini temsil etmektedir. Şekil 3.9'daki benzerlik matrisinin ana köşegen elemanlarının tamamı sıfırdır, çünkü 30 *ms*'lik her analiz çerçevesinin kendisine olan benzerlik değeri (uzaklığı) sıfırdır.

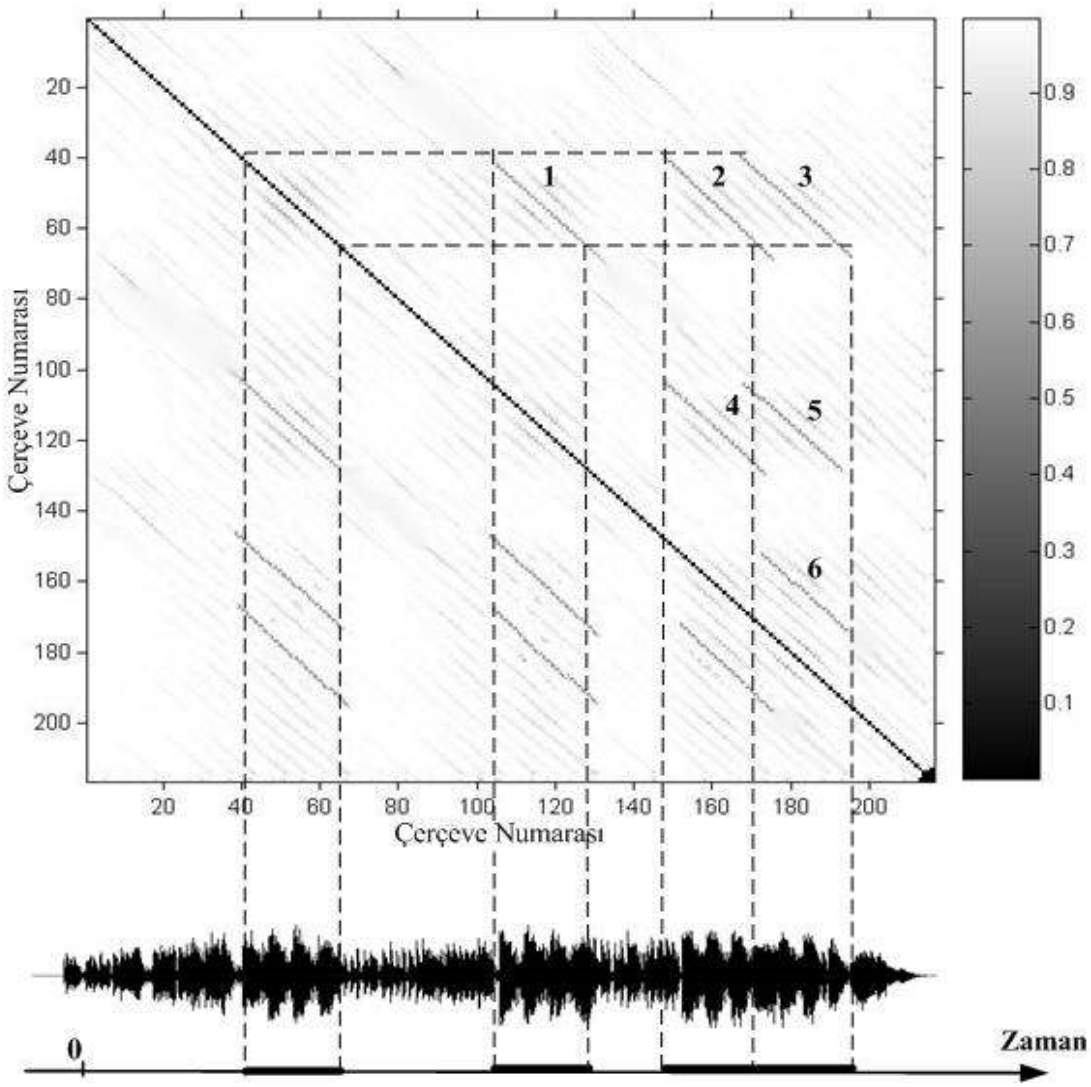
Şekil 3.9 üzerinde 1 numaralı etiketle işaretlenmiş köşegensel çizgi, bu sinyal içerisindeki nakarat bölümünün birinci tekrarını gösterirken 2 ve 3 numaralı köşegensel çizgiler sırasıyla ikinci ve üçüncü tekrarları göstermektedir. Bu sinyal içerisindeki nakarat bölümünün ilk göründüğü kısım ana köşegen üzerinde olduğu için gizli kalmıştır. Benzerlik matrisi, ilgili ses sinyalinin bütün görüntüsünü yansıttığından 4 ve 5 numaralı köşegensel çizgiler sırasıyla 2 ve 3 numaralı nakarat bölümlerinin tekrarını göstermektedir. Resim üzerindeki 6 numaralı desen ise 5 numaralı desenin tekrarını temsil etmektedir. Sonuç olarak, analiz edilen ses sinyali içerisindeki 4, 5 ve 6 numaralı desenler desen çıkarım işleminde gerekli değildir.

Desen çıkarım işlemi, resim içerisinde tekrar eden bölümlerin ses sinyali içerisindeki zamansal konumlarını tespit etmek amacıyla gerçekleştirilir. Zamansal konum bilgisi, başlama ve bitiş özellikleri ile tanımlanmaktadır. Şekil 3.9'da görülen şarkıya ait nakaratın ilk pozisyonu, 1 numaralı desenin izdüşümünün önce ana köşegene, daha sonra zaman eksenine yansıtılması neticesinde bulunmaktadır (Şekil 3.11). Böylece, ana köşegen tarafından gizlenmiş olan ilk nakarat bölümünün başlama ve bitiş zamanları elde edilebilmektedir.

Bu noktada, neden doğrudan zaman eksenine değil de, öncelikle ana köşegen üzerine izdüşümün alındığı sorusu akla gelmektedir. Bu yöntemin altında yatan düşünce, analiz edilen müzik sinyali içerisindeki nakarat bölümünün ilk geçtiği yerin ana köşegen tarafından gizlenmiş olmasıdır. Nakaratın ilk geçtiği bölümünün (ana köşegen üzerinde kalan kısım) başlangıç ve bitiş zamanlarını elde edebilmek için, öncelikle bu nakaratı temsil eden desenin, benzerlik matrisi içerisindeki başlama ve bitiş satır numaralarını tespit etmek gerekmektedir. Bu satır numaralarını tespit edebilmek için, ilk nakarat bölümünün tekrarlarını temsil eden 1, 2 ve 3 numaralı desenlerden herhangi birisinin başlama ve bitiş indis değerlerini ana köşegene yansıtmak yeterlidir. Bu mantıkla devam edilirse, 1, 2 ve 3 numaralı nakarat bölümlerinin ses sinyali içerisindeki başlama ve bitiş çerçeve numaraları, ilgili desenlerin doğrudan zaman eksenine izdüşümü ile bulunmaktadır.

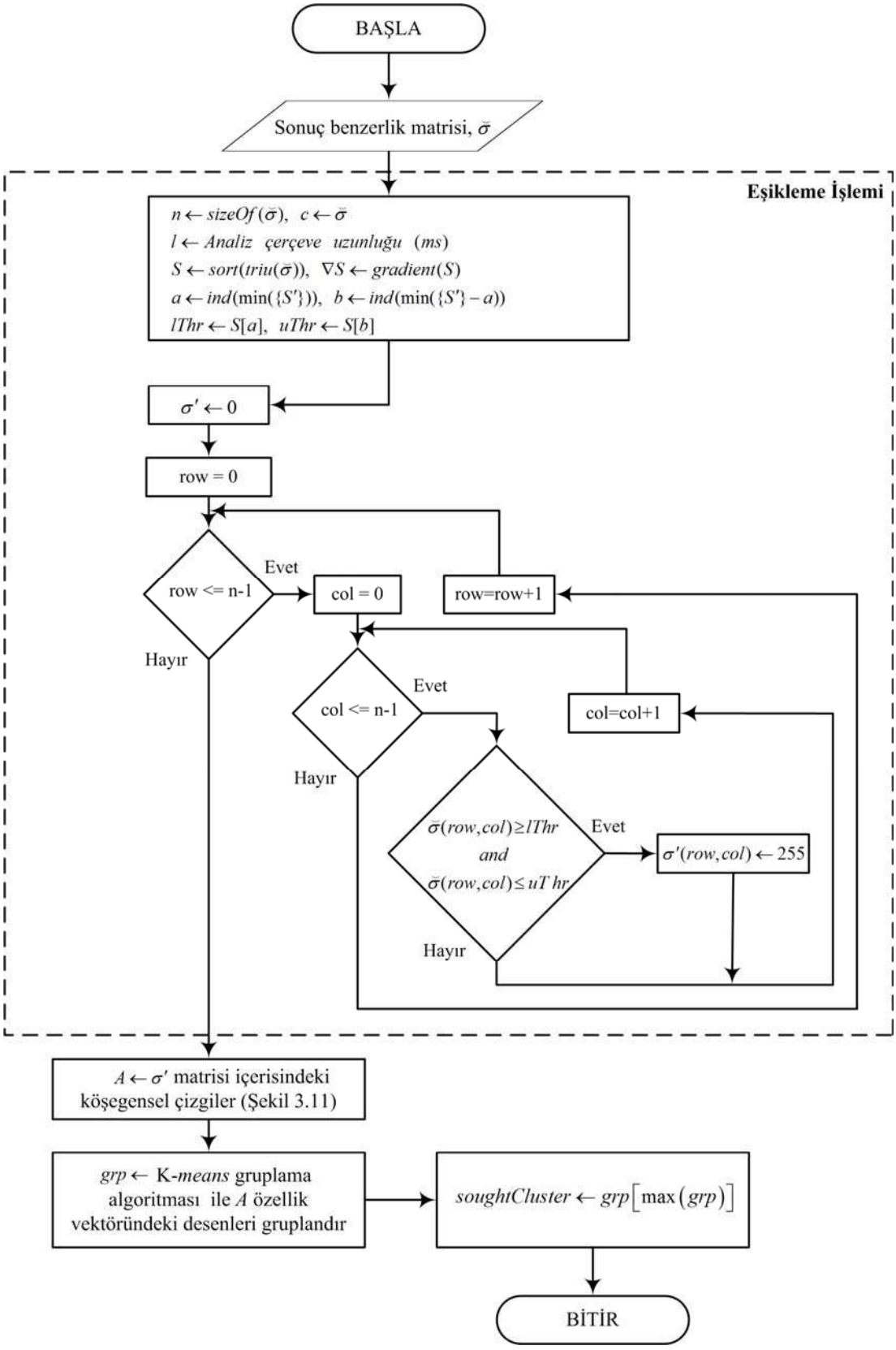
3.3.2 Anlamlı desenlerin çıkarımı

Bazı durumlarda, gerçekleştirilen yapısal içerik analizi işlemi sonucunda, ses sinyali içerisinde farklı uzunluklarda tekrar eden desenler tespit edilebilir. Böyle bir durumda, hangi desen grubunun en anlamlı desen grubu olduğunun tespiti önemlidir. Bu çalışmada, benzerlik matrisi içerisindeki en uzun desenlerin oluşturduğu gruba ait olan desenler en anlamlı desen olarak kabül edilmiştir. Böylece, ses sinyalini en iyi ifade eden bilginin elde edileceği düşünülmektedir. Bu yaklaşım sadece anlamlı desenleri tespit etmekle kalmayıp, aynı zamanda konuşma verileri içerisinde çok sık bir kullanıma sahip olan duraklama kelimelerine ait seslerin de gözardı edilmesini



Şekil 3.11. Desen pozisyonlarının izdüşüm yöntemi ile benzerlik matrisinden çıkarımı (MPEG-7 ASF özniteliği, Beatles – *Lucy in the sky* şarkısı)

sağlamaktadır. Sonuç benzerlik matrisi içerisindeki en anlamlı desenlerin çıkarılması için geliştirilen algoritmanın akış şeması Şekil 3.12'de verilmiştir. Algoritma, verilen bir benzerlik matrisinden en anlamlı desenleri üç aşamalı bir işlemden sonra çıkarmaktadır. Birinci aşamada, verilen bir benzerlik matrisi için alt ve üst eşik değerleri hesaplamaktadır. İkinci aşamada, hesaplanan eşik değerleri kullanılarak benzerlik matrisindeki tekrar eden desenler yorumlanmaktadır. Son aşamada ise, k-*means* gruplama algoritması kullanılarak yorumlanan desenler içerisinden en anlamlı desen grubu seçilmektedir. Kullanılan boyut ve yer tabanlı gruplama, ses verisinin arama ve tarama verimliliğini önemli ölçüde artırmaktadır.



Şekil 3.12. Benzerlik matrisi içerisindeki anlamlı desenlerin çıkarımı için akış şeması

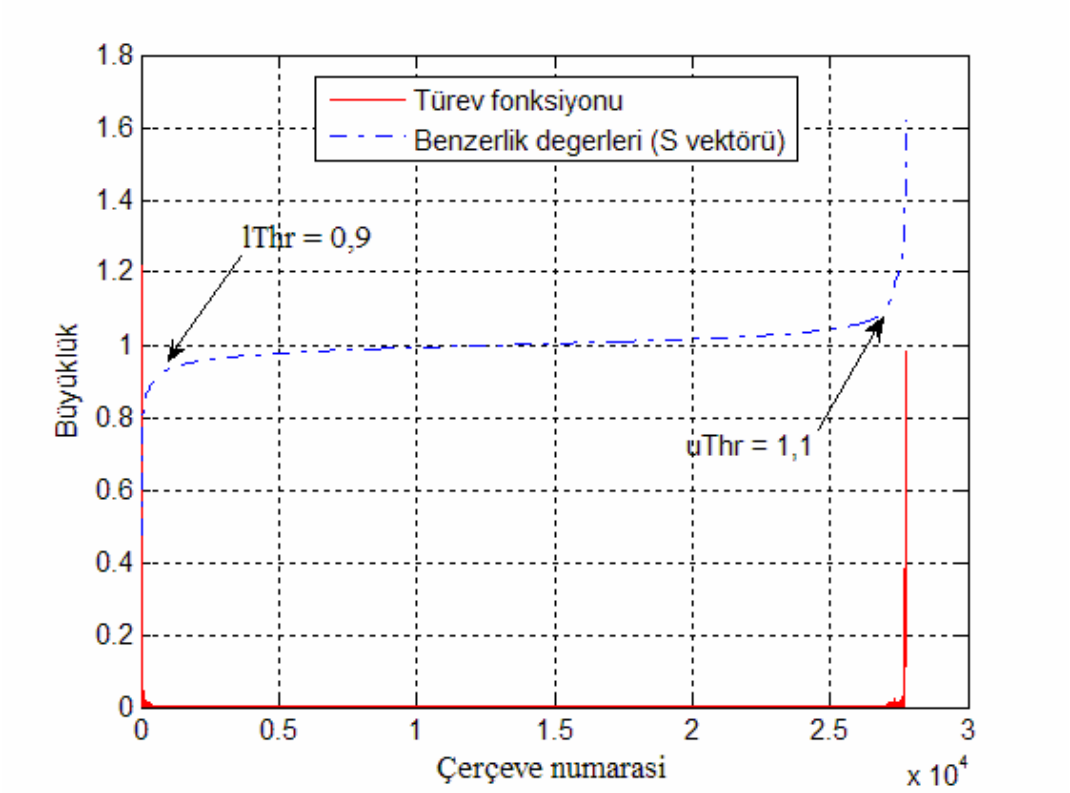
Benzerlik matrisi içerisindeki tekrar eden desenleri ifade eden köşegensel çizgilerin sınırları, ilgili desenin ses verisi içerisindeki yerini göstermektedir. Bu sınırların, benzerlik matrisi içerisinde doğru biçimde tespit edilebilmesi için uygun eşik değerlerinin kullanılması gerekmektedir. Ancak, bu eşik değerleri farklı ses verileri için farklı değerler alabilmektedir. Bu nedenle uyarlanabilir eşik değerlerine ihtiyaç duyulmaktadır. Aksi takdirde, analiz edilen her ses verisi için eşik değerinin parametre olarak girilmesi gerekmektedir.

Algoritma, *lThr* ve *uThr* adı verilen ve sırasıyla alt ve üst eşik değeri anlamına gelen iki eşik değeri hesaplamaktadır. Bu işlem için öncelikle, verilen bir benzerlik matrisinin üst üçgensel bölümü S isimli vektör içerisine alfabetik olarak artan sırada sıralanmaktadır. Sıralama işleminin benzerlik matrisinin üst üçgensel bölümü üzerinde yapılmasının nedeni, matrisin ana köşegene göre simetrik olmasıdır. Algoritma, maksimum değişim oranının yönünü S vektörünün birinci dereceden türevini hesaplayarak bulur ve ∇S ile gösterilir. Daha sonra, ∇S içerisindeki birinci ve ikinci en küçük değerlerin indislerinin S içerisinde karşılık geldiği değerler sırasıyla *lThr* ve *uThr* değişkenlerine atanır. Bu değerler, tekrar eden desenlerin sınırlarını belirlemede kullanılacak eşik değerlerini tanımlamaktadır. Eşik değerlerinin hesaplanması işlemi, Perry Farell'in "Tahitian Moon" isimli şarkısı için Şekil 3.13'te gösterilmiştir.

Eşikleme işlemi, $\breve{\sigma}$ sonuç benzerlik matrisi olmak üzere aşağıdaki biçimde gerçekleştirilir:

$$\sigma'(x,y) = \begin{cases} 0 & lThr \le \breve{\sigma}(x,y) \le uThr \\ 255 & diğer \quad durumlar \quad için \end{cases} \tag{3.10}$$

Burada, (x,y) bir pikselin gri tonlamalı renk değerini, σ' ise eşikleme işlemi neticesinde elde edilen benzerlik matrisini temsil etmektedir. Şekil 3.13'teki S vektörüne ait grafikteki düz alan korunması gereken desenlere ait pikselleri ifade ederken, geri kalan kısım istenmeyen bölüm (zemin) olarak algılanmaktadır. Sonuç



Şekil 3.13. Perry Farell'in "Tahitian Moon" şarkısı için hesaplanan eşik değerleri

olarak, piksel değeri *lThr* ve *uThr* arasında olan elemanlar korunurken, diğer elemanlar benzerlik matrisinden silinmektedir. Eş. 3.10'da kullanılan 0 ve 255 değerleri sırasıyla 256 renk paletindeki siyah ve beyaz piksel renk değerlerini temsil etmektedir. Eşikleme işleminden sonraki adım, σ' benzerlik matrisindeki aday desenlere ait özelliklerin (*oid*, *length*, *starting* ve *ending*) hesaplanması işlemidir. Bu işlem Şekil 3.11'de ayrıntılı olarak açıklandığı için bu kısımda değinilmeyecektir. Bu işlemin sonucunda bir yapı dizisi elde edilir ve A ile gösterilir. A dizisi içerisindeki her elemanın yapısı aşağıdaki biçimde tanımlıdır:

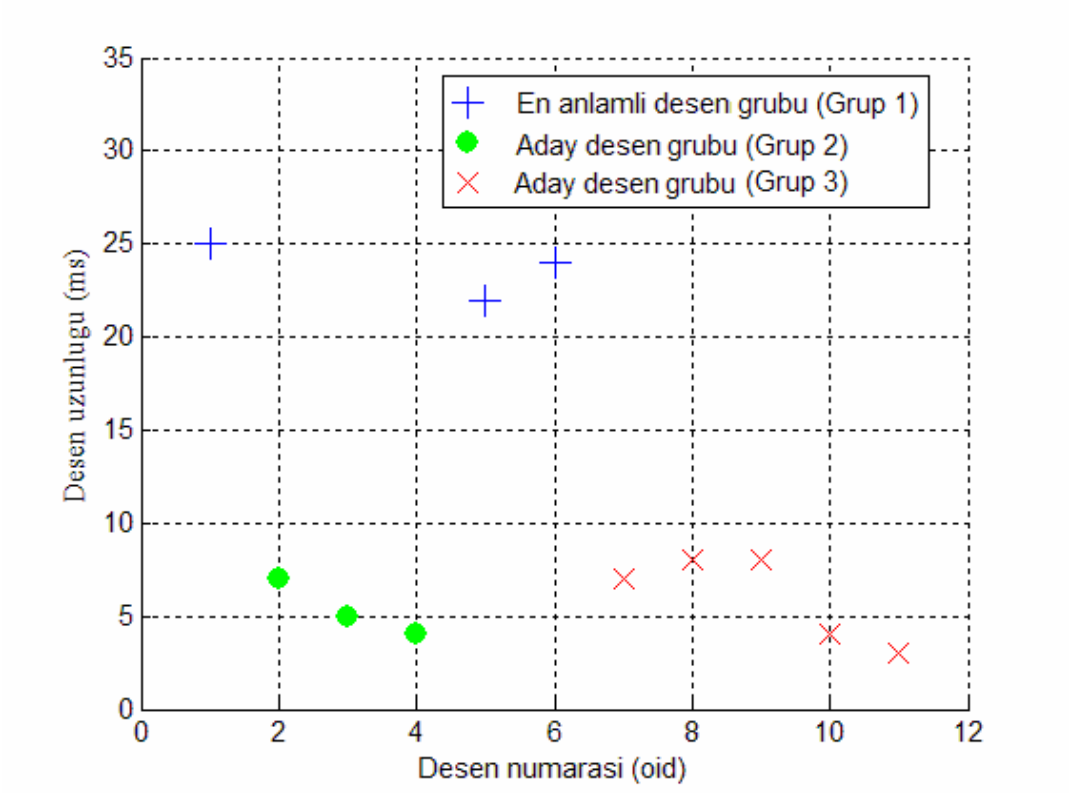
```
struct element {
        unsigned int oid;
        unsigned int length;
        unsigned int starting;
        unsigned int ending;
}
```

Burada, *oid* aday desenin tekil numarasını, *length* aday desenin *ms* cinsinden süresini, *starting* ve *ending* parametreleri ise sırasıyla aday desenin ses sinyali içerisindeki başlama ve bitiş zamanlarını temsil etmektedir.

Bazı durumlarda analizi yapılan ses sinyali içerisinde birbirinden farklı uzunluklarda tekrar eden aday desenlerle karşılaşılabilmektedir. Bu çalışmadaki diğer bir amacımız da, bu desen grupları içerisinden en anlamlı olanların tespit edilmesi olduğundan, zamansal süre cinsinden birbirine yakın uzunlukta olan desenler ile olmayan desenlerin gruplandırılması için bir yönteme ihtiyaç duyulmaktadır. Bu problemi çözmek için bu çalışmada k-*means* olarak adlandırılan gruplandırma algoritması kullanılmıştır. Gruplama yöntemleri imge işleme, otomatik öğrenme ve veri madenciliği alanlarında sıkça kullanılmaktadır. Buradaki temel amaç, birbiriyle benzer özellikler taşıyan bir dizi nesnenin aynı grup içerisinde toparlanması olarak açıklanabilir. K-*means* algoritması bu amaç için kullanılan güdümsüz bir gruplandırma algoritmasıdır.

K-*means* algoritması bu çalışmada şu şekilde kullanılmıştır: Grup sayısı (k) belirlendikten sonra, A vektörü için başlangıç merkez koordinatları sıralı olarak hesaplanır. Örneğin $k = 3$ için, A vektörünün ilk üç elemanı başlangıç merkez koordinatları olarak belirlenir. Daha sonra, A vektörünün her bir elemanının bu merkez koordinatlarına olan uzaklıkları öklit yöntemi ile hesaplanır. Birinci iterasyondan sonra, her grubun üyelerini bilmek koşuluyla, yeni üyelikler için her gruba ait yeni merkez değerleri hesaplanır. Bu işlem, hiçbir eleman yeni bir gruba dahil olmayıncaya kadar aynı şekilde devam ettirilir. Elde edilen deneysel sonuçlar neticesinde, müzik ve konuşma verileri içerisinde tekrar eden nakarat ve anahtar kavramlara ait desenlerin tespit edilebilmesi için, $k = 3$ değerinin uygun olduğu gözlemlenmiştir. K-*means* algoritması ile gerçekleştirilen bir gruplandırma işlemi, Perry Farell'in "Tahitian Moon" isimli şarkısı için Şekil 3.14'te gösterilmiştir.

Şekil 3.14'te görüldüğü üzere, "Tahitian Moon" şarkısı için üç farklı desen grubu tespit edilmiştir. Bu desen grupları içerisinde, en uzun desen üyesine sahip olan grup, anlamlı desen grubu olarak seçilmiştir.



Şekil 3.14. Perry Farell "Tahitian Moon" şarkısı için k-*means* yöntemi ile gruplama

3.4 Deneysel Sonuçlar ve Değerlendirme

Ses içerik analizi için geliştirilen yöntem popüler pop ve rak şarkılarından oluşan bir veritabanı üzerinde test edilmiştir. OST birliği tarafından yaygın olarak kullanılan MFCC özniteliği de iki amaç doğrultusunda deneysel sonuçlara dahil edilmiştir. Bunlardan birincisi, MPEG-7 ASF ve MFCC özniteliklerinin ses ve konuşma verileri içerisindeki tekrar eden desenlerin tespitinde doğruluk oranlarının karşılaştırılması, ikincisi ise, bu çalışmada önerilen yaklaşımın sonuçlarını benzer çalışmalarla karşılaştırmaktır. Çünkü benzer çalışmalarda MFCC özniteliği sıkça kullanılmıştır [16, 23].

Deneysel sonuçlar için, toplam 4 saat ve 28 dakikalık 64 şarkı ve 32 dakikalık konuşma kaydından oluşan bir ses veritabanı kurulmuştur. Şarkıların büyük çoğunluğu *The Beatles* isimli sanatçıdan seçilmiştir. Geliştirilen yöntemin ürettiği sonuçların tanıma doğruluk oranları aşağıdaki biçimde hesaplanmıştır:

$$Tanıma \quad doğruluk \quad oranı = \frac{A \bigcap M}{M} \tag{3.11}$$

Burada A, geliştirilen sistem ile otomatik olarak elde edilen bölümleri, M ses verisi içerisinden el ile çıkarılan bölümleri, $A \bigcap M$ ise, otomatik olarak tespit edilen bölümlerin el ile tespit edilen bölümler içerisindeki uyuşmasını temsil etmektedir. Ses sinyali içerisindeki anlamlı desenlerin zamansal yerleri, Eş. 3.7'de belirlenen k değerine bağlı olmak suretiyle $\pm 150\,ms$ ya da $\pm 360\,ms$ sapmayla tespit edilebilmektedir. Bu değerler sırasıyla, Eş. 3.7'deki $k = 5$ ve $k = 12$ değerleri ve $30\,ms$'lik analiz çerçeve boyutu için elde edilmiştir.

3.4.1 Müzik verisi

Veritabanında kullanılan bütün şarkılar iki kanal, 16-*bit* ve 44 *kHz'lik* örneklem oranına sahiptir. Bu çalışmada önerilen yöntemin doğruluk oranını ölçebilmek için, Eş 3.11'deki eşitlik kullanılmıştır. El ile çıkarılan nakarat bölümlerinin sayısının doğruluğunu onaylayabilmek için, Allan W. Pollack'ın Web üzerindeki "Notes On" isimli çalışması ile karşılaştırılmıştır [46]. Müzik verisi için elde edilen sonuçlar Çizelge 3.1'de verilmiştir. Çizelgenin açık olması için veritabanındaki bütün şarkı isimleri listelenmemiştir. Bu çalışmada önerilen yöntemin MPEG-7 ASF özniteliği ile birlikte kullanıldığında toplamda 172 nakarat bölümünün 148 adetini ve yerini doğru biçimde tespit ettiği gözlemlenmiştir. Geliştirilen yöntem bu değerler ile %86'lık $(148/172)$ bir tanıma doğruluk oranına sahiptir. Öznitelik olarak MFCC kullanıldığında, önerilen yöntemin toplamda 172 nakarat bölümünün 141 adetini ve yerini doğru biçimde tespit ettiği gözlemlenmiştir. Geliştirilen yöntem bu değerler ile %82'lik $(141/172)$ bir tanıma doğruluk oranına sahiptir. Sonuçlardan da görüleceği üzere, MPEG-7 ASF özniteliği, MFCC özniteliğine göre biraz daha iyi bir tanıma doğruluk oranına sahiptir.

Bu çalışmada gerçekleştirilen yöntemin doğruluk oranını benzer yöntemlerle karşılaştırmak için, Foote'nun çalışmasında kullanılan şarkılar da test veritabanına eklenmiştir [16].

Çizelge 3.1. Müzik verileri için nakarat bölümlerinin tanıma doğruluk oranları (k-*means* gruplama yöntemi için $k = 3$)

Şarkı Adı	Sanatçı	Süre (sa:dk:sn)	El ile	Geliştirilen yöntem (ASF)	Geliştirilen yöntem (MFCC)	[16]
Wild Honey	U2	3:46.612	3	3	2	3
Magical Mystery	The Beatles	2:24.320	3	2	2	3
Tahitian Moon	Perry Farell	3:47.395	3	3	3	3
Lucy in the Sky	The Beatles	3:27.438	4	4	4	3
The Zephyr Song	Chili Peppers	3:53.560	3	3	3	2
Hash Pipe	Weezer	3:06.592	3	3	3	2
Optimistic	Radiohead	5:16.029	2	2	2	2
		Alt toplam	21	20	19	18
		Alt tanıma doğruluk oranı	–	%95	%91	%86
Fast Car	Tracy Chapman	4:56.636	3	2	2	–
Blue Hotel	Chris Isaak	3:13.358	3	3	3	–
Always	Jon Bon Jovi	4:19.918	3	3	3	–
Yesterday	The Beatles	2:05.126	4	3	2	–
Revolution	The Beatles	4:15.790	3	1	1	–
...	...	...	...	...	...	...
		Toplam	172	148	141	–
		Toplam tanıma doğruluk oranı	–	%86	%82	–

Bu şarkılar, Çizelge 3.1'in son sütununda gösterilen ilk yedi şarkıdan oluşmaktadır. Bu çalışmada önerilen yöntem, MPEG-7 ASF özniteliği kullanıldığında Çizelge 3.1'deki toplam 21 nakarat bölümünden 20 adetini ve yerini doğru biçimde tespit ederken, MFCC özniteliği kullanıldığında toplam 19 adetini ve yerini doğru biçimde tespit etmektedir. Bu değer Foote'nun çalışmasında 18 olarak rapor edilmiştir. Geliştirilen yöntem MPEG-7 ASF özniteliği ile kullanıldığında tanıma doğruluk oranı %95, MFCC özniteliği ile birlikte kullanıldığında bu sonuç %91 iken, Foote'nun çalışmasındaki tanıma doğruluk oranı ise %86'dır. Bu sonuçlarıyla geliştirilen yöntem en iyi tanıma doğruluk oranını MPEG-7 ASF özniteliği ile elde etmektedir.

3.4.2 Konuşma verisi

Geliştirilen sistemin konuşma verisi için test edilebilmesi için yaklaşık olarak 32 dakikalık yapay konuşma kaydı oluşturulmuştur. Yapay konuşma kaydının oluşturulmasının nedeni, uygulamaya uygun daha önceden işaretlenmiş konuşma verilerinin mevcut olmayışıdır. Geliştirilen yönteme uygun yapay konuşma verisinin

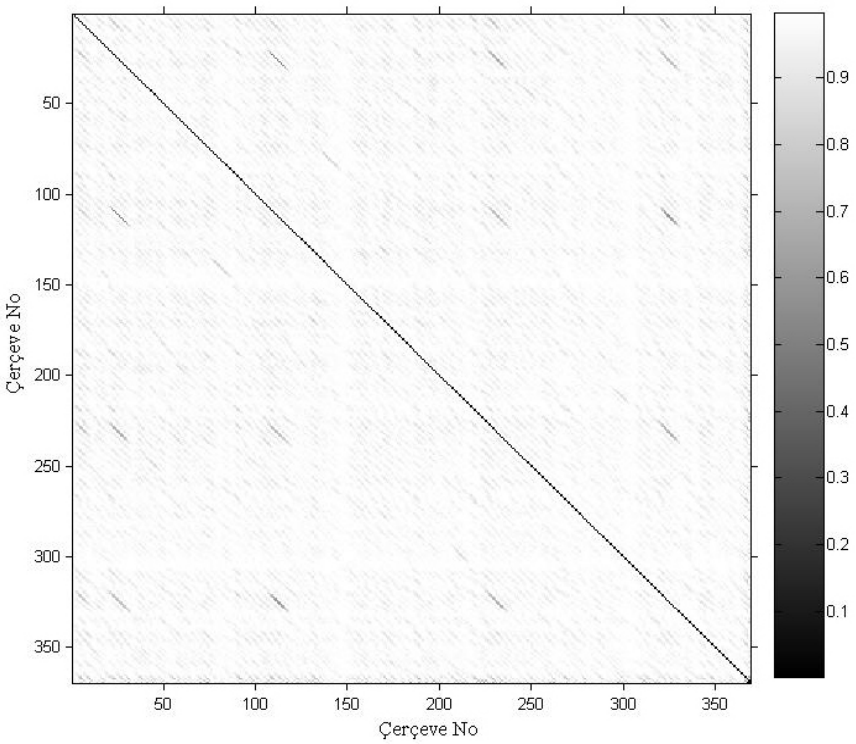
oluşturulabilmesi için, içerisinde kendisini belirli sıklıklarda tekrar eden üç adet konuşma metni hazırlanmıştır. Bu metinlerden birincisinin konusu *Pakistan Depremi* olup, NTVMSNBC İnternet haber portalından temin edilmiştir [47]:

"Pakistan yardım bekliyor. Pakistan depreminin 7,6 büyüklüğünde olduğu ve depremin bilançosunun giderek ağırlaştığı bildiriliyor. Merkez üssü Keşmir bölgesi olan Pakistan depreminde, ölü sayısının kırk bine yaklaştığı bildiriliyor. Hükümet kaynakları otuz ile kırk bin arasında ölü olduğunu, yaralı sayısınınsa altmış bini aştığını belirtti. ABD Pakistan depremi için acil yüz bin dolar göndereceğini açıkladı ve ABD helikopterlerinin bölgede kullanılabileceğini belirtti. Türk Kızılay'ı da Pakistan depremi nedeniyle kurtarma ekibi ve yardım malzemesi gönderdi".

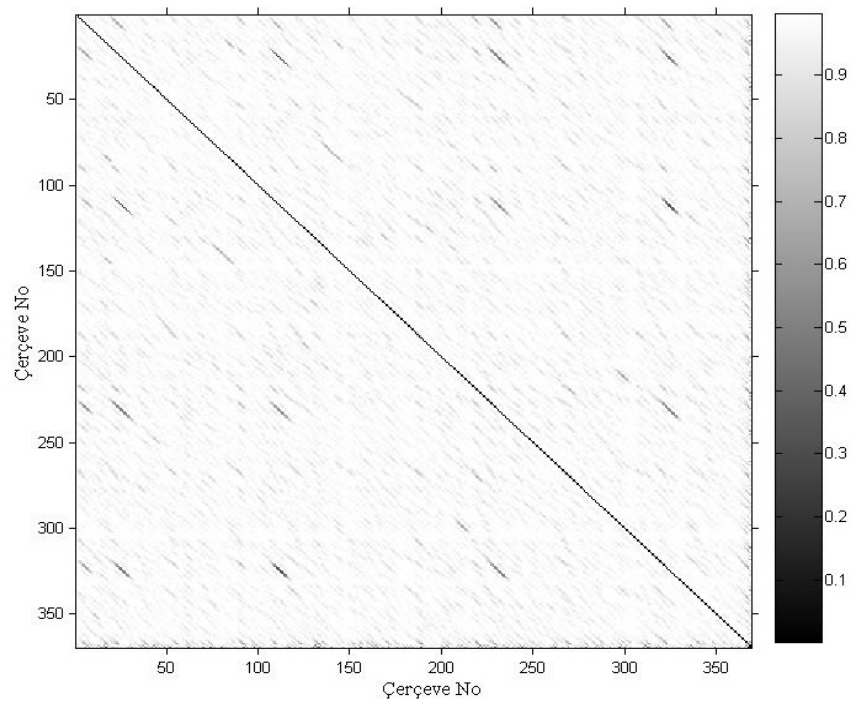
Diğer iki metin ise bilgisayar bilimleri ders notlarından şeçilmiştir. Bu metinler daha sonra bir mikrofona konuşmak suretiyle CD kalitesinde (44 *kHz*, 16-*bit*) kayıt edilmiştir. Kayıtlardaki akustik değişimlerden etkilenmemek için, ilgili metinlerin herbiri beş kadın ve beş erkek konuşmacı tarafından okunmuştur. Geliştirilen yöntemin konuşma verileri için test edilmesi için, müzik verilerinde kullanılan yaklaşımın kullanılmıştır (Eş 3.11).

Yukarıda metin içeriği verilen konuşmanın, MPEG-7 ASF ve MFCC öznitelikleri ile elde edilen benzerlik matrisleri sırasıyla Şekil 3.15 ve Şekil 3.16'da verilmiştir. Bu konuşmanın anahtar kavramı, bütün konuşmacılar tarafından, konuşma içerisinde dört kez tekrar eden *Pakistan Depremi* olarak belirlenmiştir. Şekil 3.15 ve Şekil 3.16'da elde edilen benzerlik matrislerinin üst kısımlarında nispeten daha kısa biçimde görünen köşegensel çizgiler, bu konuşma metninin anahtar kavramı olan *Pakistan Depremi* kelimelerine ait seslerin, *Pakistan* kısımlarını ifade etmektedir.

Konuşma verisi için elde edilen sonuçlar Çizelge 3.2'de listelenmiştir. Geliştirilen yöntem MPEG-7 ASF özniteliği ile birlikte kullanıldığında, konuşma verileri içerisindeki toplam 120 anahtar kavramdan 93 adetini ve yerini doğru biçimde tespit etmiştir. Bu değerler ile elde edilen tanıma doğruluk oranı $93/120 = \%78$ olarak gözlemlenmiştir. Elde edilen bu sonuç müzik verileri ile karşılaştırıldığında daha düşük bir tanıma doğruluk oranı elde edilmiştir. Bu sonuca, kayıt karakteristiklerinin ve konuşmacıların seslerindeki ton değişiklerinin sebep olduğu düşünülmektedir.



Şekil 3.15. İmge işleme teknikleri neticesinde MPEG-7 ASF özniteliği için elde edilen sonuç benzerlik matrisi (Konuşma metni – *Pakistan Depremi*)



Şekil 3.16. İmge işleme teknikleri neticesinde MFCC özniteliği için elde edilen sonuç benzerlik matrisi (Konuşma metni – *Pakistan Depremi*)

Çizelge 3.2. Konuşma verileri için anahtar kavram tanıma doğruluk oranları (k-*means* gruplama yöntemi için $k = 3$)

Konuşma Metni Kaynağı	Anahtar Kavram	Geliştirilen yöntem (ASF)		Geliştirilen yöntem (MFCC)	
		Kadın	Erkek	Kadın	Erkek
Haber Metni	*Pakistan Depremi*	18/20	14/20	19/20	17/20
Ders Notu	*Nesne yönelimli Analiz ve tasarım*	13/15	11/15	14/15	12/15
Ders Notu	*Java dili*	20/25	17/25	23/25	19/25
Toplam tanıma doğruluk oranı		$93/120 = \%78$		$104/120 = \%87$	

Geliştirilen yöntem MFCC özniteliği ile birlikte kullanıldığında, konuşma verileri içerisindeki toplam 120 anahtar kavramdan 104 adetini ve yerini doğru biçimde tespit etmiştir. Bu değerler ile elde edilen tanıma doğruluk oranı $104/120 = \%87$ olarak gözlemlenmiştir. MPEG-7 ASF özniteliği ile karşılaştırıldığında, MFCC özniteliği daha yüksek bir tanıma doğruluk oranı elde etmiştir. Bu farklılık, MFCC özniteliğinin bazı kadın ve erkek konuşmacılara ait sesleri daha iyi tanıması sonucunda oluşmuştur. Bu durumun, MFCC özniteliği ile MPEG-7 ASF özniteliklerinin kullandığı frekans ölçeklerinden kaynaklandığı düşünülmektedir. MFCC özniteliğinin kullandığı frekans ölçeği (1 *kHz*'in altında doğrusal ölçek, üstünde logaritmik ölçek), MPEG-7 ASF özniteliğinin kullandığı frekans ölçeğine (Eş. 3.2'de açıklandığı şekliyle bütün frekanslar için logaritmik ölçek) göre daha güçlüdür. Çünkü, insanlara ait konuşma sinyallerinin büyük çoğunluğu 1 *kHz'in* altındadır.

4. İÇERİK-TABANLI SORGU VE TARAMA

Klasik bilgi getirim sistemleri, ilişkisel veritabanlarında birbirleri ile ilişkili tablolar içerisinde saklanan verilere erişimi form benzeri arayüzler ile sağlamaktadır. İstemciler bir sorgu ifadesini, kendilerine sağlanan bu formlar üzerinden aranılan bilgilere ait kesin özellikler tanımlamak suretiyle belirtmektedirler. Ancak, bu yöntem istemcilerin genellikle sorgulanacak verinin yapısı hakkında önbilgiye sahip olmamaları nedeniyle, çokluortam bilgi getirim sistemlerinde istenmeyen sonuçların getirilmesine neden olmaktadır. Bu problemin çözümü için uygun yöntemlerden birisi, istemcilerin iyi tasarlanmış sorgu arayüzleri üzerinden *yaklaşık* (*benzerlik*) *sorgular* girebilmeleridir. Bu özellik, tam uyuşmanın arandığı klasik bilgi getirim sistemleri ile çokluortam bilgi getirim sistemleri arasındaki en önemli farklılıklardan birisidir. Yaklaşık sorgu özelliğini destekleyebilmek için, karmaşık bir yapıya sahip olan çokluortam verilerine ait, verimli saklama ve getirim modellerinin tasarlanarak geliştirilmesi gerekmektedir [11, 27, 48-50].

Önceki bölümde, ses içerik analizi çerçevesinde, konuşma ve müzik verileri içerisindeki tekrar eden desenlerin çıkarılması için geliştirilen yöntem tanıtılmıştır. Çıkarılan bu bilgiler, ses dizinleme ve getirim uygulamalarında çok önemli bir yere sahiptir. Böylece istenilen bilgiye hızlı ve doğru biçimde ulaşabilmek mümkün olacaktır. Bu bölümde, ses içerik analizi çerçevesinde elde edilen bilgileri kullanan içerik tabanlı bir ses getirim sistemi modeli önerilmektedir. Önerilen model ile ses verisi anlamsal ve yapısal olmak üzere iki seviyede sorgulanabilmektedir. Anlamsal sorgular, oluşturulan ses efekt veritabanı sayesinde ve *nakarat*/*anahtar kavram* sorguları ile gerçekleştirilebilmektedir. İstemciler ses efekt sorgularını, model tarafından önceden tanımlanan anahtar kelimeler ile ifade edebilmektedir. Bu anahtar kelimelere silah patlaması, patlama, hayvan sesi gibi ses efektleri örnek olarak verilebilmektedir. Veritabanı yeni ses efektleri ile eğitilmek suretiyle, bu efektler için de sorgu ifadeleri girebilmek mümkündür. İstemcilerin ifade edebileceği anlamsal sorgulara "*silah patlama sesi içeren bütün sesleri bul*" şeklindeki bir sorgu örnek olarak verilebilir.

4.1 Desen Eşleme Yöntemi

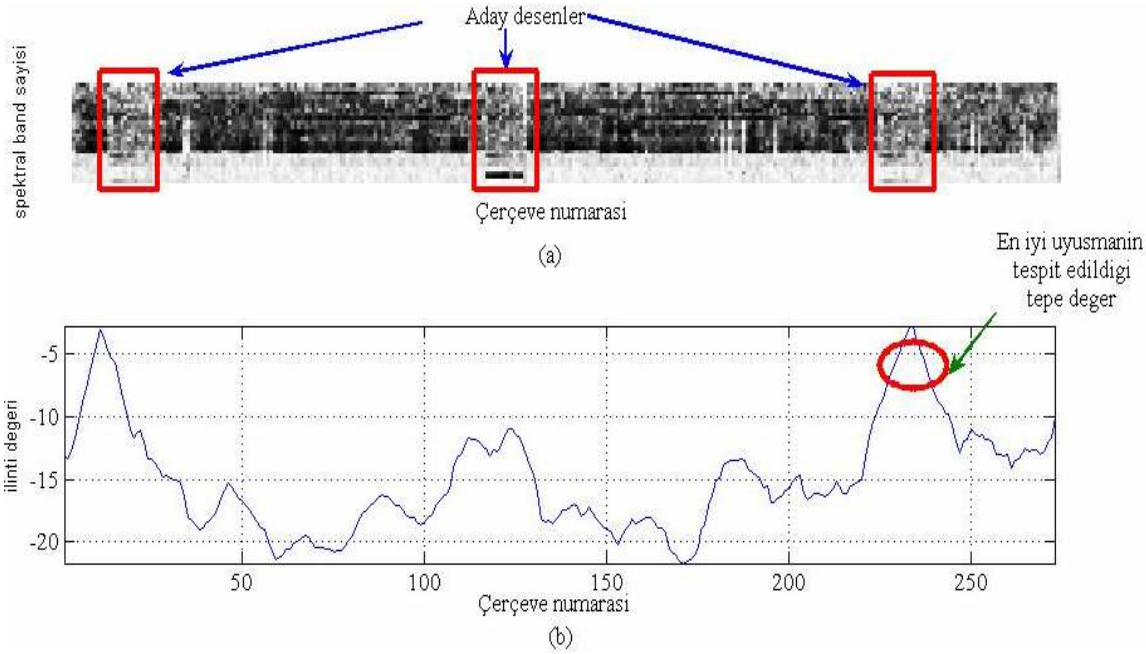
Ses verilerinin içerik tabanlı getiriminde, sorgu elemanı bir ses dosyası iken, sorgu sonucu ise benzerliklerine göre sıralanmış bir dizi ses dosyasıdır. Bu nedenle, sorgu elemanının ses veritabanı içerisinde aranması için uygun bir yönteme ihtiyaç duyulmaktadır. Lu'nun belirttiği üzere, iki ses dosyasının içerik tabanlı karşılaştırılması için en uygun yöntem, karşılaştırma işlemin ses öznitelikleri üzerinden gerçekleştirilmesi şeklindedir [43]. Bu yaklaşımda, genellikle vektörler biçiminde temsil edilen öznitelikler kümesinin birbirlerine olan benzerlikleri hesaplanmak suretiyle karşılaştırma işlemi gerçekleştirilmektedir. Bu yöntem aynı zamanda desen eşleme problemi olarak da adlandırılabilmektedir.

Bu çalışmada, öznitelik vektörlerinin birbirlerine olan benzerliklerini hesaplayabilmek amacıyla, bu vektörlerin karşılıklı ilintileri hesaplanmaktadır. Karşılıklı ilinti, iki vektörün birbirlerine olan benzerliklerinin (ya da benzemezliklerinin) bir ölçüsünü temsil etmektedir.

Verilen $f(x,y)$ ve $w(x,y)$ öznitelik matrislerinden $w(x,y)$, MxN boyutundaki $f(x,y)$ matrisi içerisinde aranan JxK boyutundaki desen matrisi olsun. Bu durumda $f(x,y)$ ve $w(x,y)$ arasındaki karşılıklı ilinti aşağıdaki şekilde hesaplanır:

$$c(x,y) = \sum_{m=0}^{J-1} \sum_{n=0}^{K-1} f(m,n) w(x+m, y+n) \tag{4.1}$$

Burada $x = 0,1,2,...,M-1$ ve $y = 0,1,2,...,N-1$ olmak üzere w ve y'nin örtüştüğü bölgelerde toplama işlemi gerçekleştirilmektedir. c'nin maksimum değeri w'nin y içerisinde en iyi uyuştuğu pozisyon bilgisini vermektedir. Bir ses sinyali içerisindeki iki farklı analiz çerçevesine ait öznitelikleri ifade eden V_i ve V_j' nin tanımlandığı Eş. 3.5'ten faklı olarak Eş. 4.1'deki matrisler $J \leq M$ ve $K \leq N$ koşulunu sağlayan iki farklı ses sinyaline ait öznitelikleri temsil etmektedir.

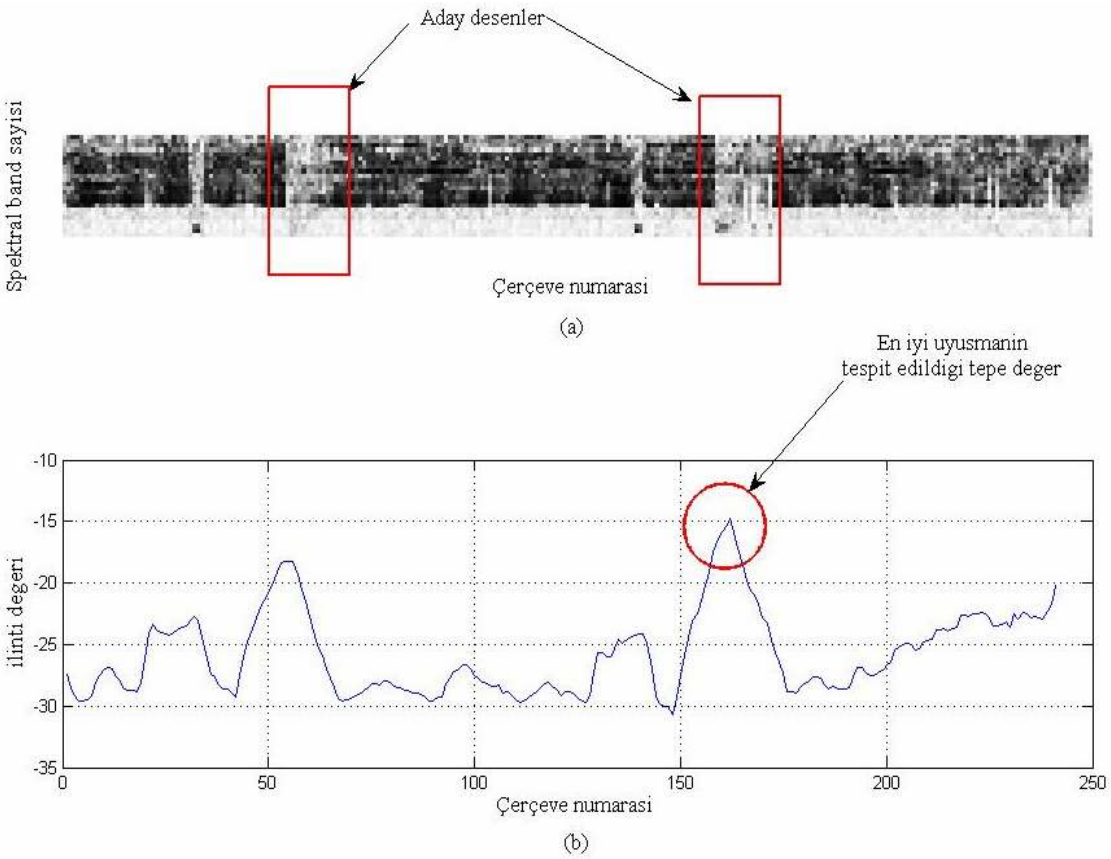


Şekil 4.1. Senaryo 1 için karşılıklı ilinti sonucu (a) Ses sinyalinin MPEG-7 ASF için elde edilmiş öznitelik matrisi (b) Karşılıklı ilinti grafiği

Eş. 4.1'de verilen desen eşleme yönteminin sonuçlarını gösterebilmek amacıyla iki farklı senaryo hazırlanmıştır.

Senaryo 1: Silah patlama sesini içeren üç farklı ses efekti, Vivaldi'nin "Four Season-Spring" eserinin içerisine rastgele pozisyonlara arkaplan sesi olarak eklenmiştir. Daha sonra, eklenen üç silah sesi efektinden farklı bir silah sesi efekti sorgu elamanı olarak tanımlanmıştır.

Bu sorgu işlemi için elde edilen karşılıklı ilinti sonucu Şekil 4.1'de verilmiştir. Şekil 4.1.a'da, *Senaryo 1*'deki ses sinyali için hesaplanmış MPEG-7 ASF özniteliğinin iki boyutlu görüntüsü verilmiştir. Dikkat edilirse, öznitelik görüntüsünde dikdörtgenlerle işaretlenmiş bölümler, bu ses sinyaline arkaplan sesi olarak eklenen ses efektlerini temsil etmektedir ve belirgin bir şekilde görünmektedir. Şekil 4.1.b'de ise, sorgu elamanının *Senaryo 1* için oluşturulan ses sinyali içerisinde, öznitelikler üzerinden karşılıklı ilintilerinin hesaplanmasıyla elde edilen karşılıklı ilinti grafiği görülmektedir. Bu grafik üzerinde işaretlenen tepe değeri, sorgu elemanının ses



Şekil 4.2. Senaryo 2 için karşılıklı ilinti sonucu (a) Ses sinyalinin MPEG-7 ASF için elde edilmiş öznitelik matrisi (b) Karşılıklı ilinti grafiği

sinyali içerisinde en iyi uyuşmanın olduğu yeri temsil etmektedir.

Senaryo 2: Silah patlama sesini içeren iki farklı ses efekti, Vivaldi'nin "Four Season-Spring" eserinin içerisine rastgele pozisyonlara arka plan sesi olarak eklenmiştir. Daha sonra, ses sinyaline eklenen ikinci silah sesi efekti sorgu elamanı olarak tanımlanmıştır.

Bu sorgu işlemi için elde edilen karşılıklı ilinti sonucu Şekil 4.2'de verilmiştir. Şekil 4.2.a'da, *Senaryo 2*'deki ses sinyali için hesaplanmış MPEG-7 ASF özniteliğinin iki boyutlu görüntüsü verilmiştir. Dikkat edilirse, öznitelik görüntüsünde dikdörtgenlerle işaretlenmiş bölümler, bu ses sinyaline arka plan sesi olarak eklenen ses efektlerini belirgin şekilde temsil etmektedir. Şekil 4.2.b'de ise, sorgu elamanının *Senaryo 2* için oluşturulan ses sinyali içerisinde, öznitelikler üzerinden karşılıklı ilintilerinin

hesaplanmasıyla elde edilen karşılıklı ilinti grafiği görülmektedir. Bu grafik üzerinde işaretlenen tepe değeri, sorgu elemanının ses sinyali içerisindeki en iyi uyuşmanın olduğu yeri temsil etmektedir.

Yukarıda hazırlanan iki senaryodan elde edilen sonuçlar, kullanılan MPEG-7 ASF özniteliği ve karşılık ilinti yaklaşımının ses verilerinin içerik tabanlı getiriminde oldukça ayırtedici olduğunu göstermektedir. Özellikle birinci senaryoda, sorgu elemanı, ses verisi içerisindeki silah sesi efektlerinden farklı olmasına rağmen, arama işlemi neticesinde doğru sonuç elde edilmiştir. Bu özelliği ile, çalışmada önerilen "*silah patlama sesi içeren bütün sesleri bul"* şeklindeki anlamsal sorgular sistem tarafından doğru biçimde değerlendirilebilmektedir.

Karşılıklı ilinti hesabı ile en iyi uyuşmanın sağlandığı çerçeve numarası, karşılıklı ilinti vektöründen (c) tespit edilebilmektedir. Bir Analiz çerçeve uzunluğu ($30\ ms$) bilindiğine göre, aranılan desenin ilgili ses sinyali içerisindeki yeri, başlama ve bitiş zamanları cinsiden aşağıdaki biçimde tespit edilmektedir:

$$t_{start} = ind(\max(c)) * t_{frame} \qquad t_{end} = t_{start} + J * t_{frame} \tag{4.2}$$

Burada t_{start} ve t_{end} sırasıyla aranılan desenin ses sinyali içerisindeki başlama ve bitiş zamanlarını, $ind(\max(c))$ en iyi uyuşmanın görüldüğü indis numarasını, t_{frame} bir analiz çerçevesinin *ms* cinsinden uzunluğunu ve J ise sorgu elemanına ait öznitelik vektörünün analiz çerçeve sayısını temsil etmektedir. Bu bilgi, önerilen sistemin sadece içerik tabanlı ses getirim kısmında değil, aynı zamanda ses sinyalinin içerik tabanlı taranması kısmında da kullanılmaktadır.

4.2 Nokta Sorguları

Çokluortam verilerinin sorgulanma biçimlerinden birisi olan nokta sorgu tipinde, kullanıcı sorgusu bir öznitelik vektörü ile ifade edilir ve bu vektör, ses veritabanındaki daha önceden hesaplanmış öznitelik vektörleri ile karşılaştırılarak en

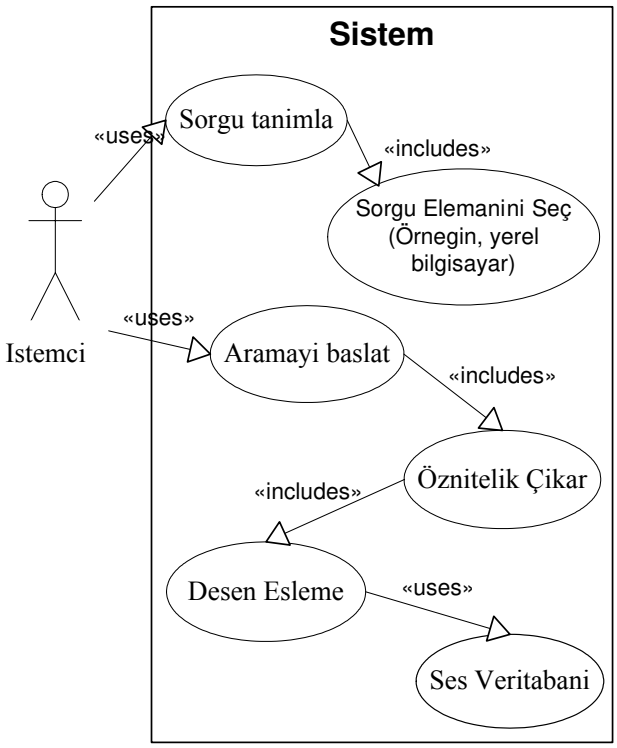
iyi uyuşmanın olduğu ses sinyali sonuç olarak getirilmektedir. Karşılaştırma işlemi, Eş. 4.1'de verilen karşılıklı ilinti hesaplama yöntemi ile gerçekleştirilir. Daha açık bir şekilde ifade etmek gerekirse, sırasıyla aşağıdaki işlem adımları uygulanır:

1. İstemci öncelikle benzerini bulmak istediği ses dosyasını ve karşılaştırmada kullanılacak öznitelik tipini (MPEG-7 ASF ya da MFCC) belirler.
2. İstemci sorgusunu sisteme gönderir.
3. Sistem, kendisine sunulan sorgu için,
A. Sorgu elemanına ait seçilen öznitelik vektörünü (MPEG-7 ASF ya da MFCC) çıkarır.
B. Bu vektör ile veritabanındaki tüm öznitelik vektörleri arasındaki karşılıklı ilintiler hesaplanmak suretiyle karşılaştırma yapılır.
C. Karşılaştırma sonucunda, aday sesler ilgililik (benzerlik) oranına göre sıralanır.
D. Sıralamada en yüksek benzerlik değerine sahip ses sinyali sonuç olarak döndürülür.

Sistem ile kullanıcı etkileşimlerini gösteren UML *use-case* şeması Şekil 4.3'de verilmiştir. Bu sorgu tipinde, sisteme örnek olarak bir ses verisi verilir ve bu ses verisine en çok benzeyen 0..1 adet ses verisi sonuç olarak geri döndürülür. Sisteme örnek olarak verilen ses, yerel bilgisayarda önceden kayıtlı bulunan bir ses verisi olabileceği gibi, sorgu anında bir mikrofona mırıldanmak suretiyle oluşturulan bir ses kaydı da olabilmektedir.

4.3 Aralık Sorguları

Çokluortam verilerinin sorgulama biçimlerinden bir diğeri de *aralık sorgusu* olarak adlandırılmaktadır. Bu sorgu tipi, çokluortam verilerinin sorgulanmasında çok önemli bir yere sahiptir. Çünkü, bu tür sorgularda genellikle benzerlik uyuşması aranmaktadır. Bu sorgu tipinde, kullanıcı sorgusu bir öznitelik vektörü ve bir benzerlik aralığı ile ifade edilir.



Şekil 4.3. Nokta sorgu tipi için kullanıcı-sistem etkileşimini gösteren UML use-case şeması

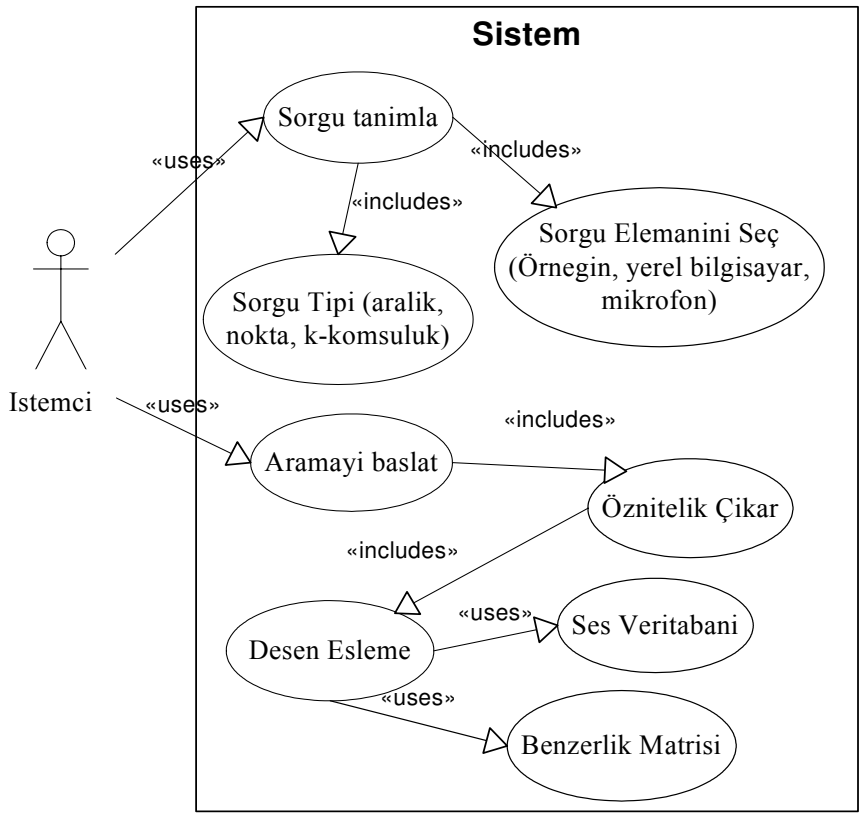
Sorgu elemanını temsil eden öznitelik vektörü, ses veritabanındaki daha önceden hesaplanmış öznitelik vektörleri ile karşılaştırılmak suretiyle, sorguda belirlenen benzerlik aralığı içerisinde bir uyuşmaya sahip olan ses sinyalleri sonuç olarak getirilmektedir. Benzerlik aralığı, yüzdelik olarak (Örneğin %80 - %90) ya da 0 ile 1 arasında bir değer olarak verilebilmektedir. Nokta sorgulardan farklı olarak, bu sorgu işleminin neticesinde birden fazla sonuç geri döndürülebilmektedir. Aralık sorguları ile iki farklı durum söz konusu olabilmektedir: (1) Sisteme girilen sorgunun bir ses sinyali içerisinde aranması (tarama), (2) sisteme girilen sorgunun ses veritabanındaki bütün elemanlar arasında aranması.

Birinci durumda, verilen sorgu ve benzerlik aralığına uygun bir ses sinyali içerisindeki aranılan bütün bölümler sonuç olarak döndürülmektedir. Böyle bir sorgu için gerçekleştirilen işlemler sırasıyla aşağıdaki gibidir:

1. İstemci öncelikle aramak istediği sorgu elemanını, karşılaştırmada kullanılacak öznitelik tipini (MPEG-7 ASF ya da MFCC) ve benzerlik aralığını belirler.
2. İstemci sorgusunu sisteme gönderir.
3. Sistem, kendisine sunulan sorgu için,
A. Sorgu elemanına ait seçilen öznitelik vektörünü (MPEG-7 ASF ya da MFCC) çıkarır.
B. Bu vektör ile taramanın yapılacağı ses sinyalinin öznitelik vektörleri arasındaki karşılıklı ilinti hesaplanır.
C. Hesaplanan ilintiden en iyi uyuşmanın elde edildiği indis numarası Eş. 4.2'deki yöntemle ilinti vektöründen tespit edilir. (Bu indis numarası aynı zamanda orijinal ses sinyali içerisindeki çerçeve numarasıdır).
D. Sorguda belirtilen benzerlik aralığındaki sonuçları hesaplamak için,
I. En iyi uyuşmanın olduğu çerçeve numarası için benzerlik matrisindeki aynı satıra konumlanılır (Benzerlik matrisindeki her satırın bir analiz çerçevesinin diğer çerçevelere olan benzerlik değerlerini depoladığını hatırlayınız).
II. Bu satırdaki değerler indis numaraları korunmak suretiyle alfabetik olarak artan biçimde sıralanır (En küçük değerin en büyük benzerliği ifade ettiğini hatırlayınız).
III. Sıralanan elemanlar arasından, benzerlik aralığı sınırları içerisinde olan elemanlar indis numaraları ile döndürülür.
IV. Döndürülen bu indis değerleri, sorgu elemanının verilen benzerlik aralığı için ses sinyali içerisindeki bölümlerin çerçeve numaralarıdır. Bu çerçeveler istemciye sonuç olarak geri döndürülür.

Geleneksel yöntemde kullanıcı sorgusu, ilgili ses sinyali içerisinde belirli uzunluktaki kaymalarla ayrı ayrı eşleştirilmek suretiyle benzerlikler aranmaktadır. Arama işleminin sonucunda, bütün benzerlik değerleri sıralanmak suretiyle, kullanıcı sorgusunda belirtilen benzerlik aralığına uyan bölümler sonuç olarak döndürülmektedir.

Bir ses sinyali için aralık sorguların, benzerlik matrisi gösteriminden yararlanılarak değerlendirilmesi, arama esnasındaki desen eşleme karmaşıklığını azaltmaktadır.



Şekil 4.4. Aralık ve En Yakın k-komşuluk sorgu tipleri için kullanıcı-sistem etkileşimini gösteren UML use-case şeması

Daha açık ifade etmek gerekirse, bu yaklaşımla, karşılaştırma işlemi neticesinde bulunan en iyi sonuç için, ilgili ses sinyali içerisinde istemcinin verdiği benzerlik aralığını sağlayan diğer bütün karşılaştırmaların yapılma zorunluluğu ortadan kaldırılmıştır. Aralık sorguları için, sistem ile kullanıcı etkileşimlerini gösteren UML *use-case* şeması Şekil 4.4'de verilmiştir.

İkinci durumda, verilen sorgu ses veritabanındaki bütün elemanlar için gerçekleştirilmektedir. Bu sorgu biçiminde sırasıyla aşağıdaki işlemler gerçekleştirilmektedir:

1. İstemci öncelikle aramak istediği sorgu elemanını, karşılaştırmada kullanılacak öznitelik tipini (MPEG-7 ASF ya da MFCC) ve benzerlik aralığını belirler.
2. İstemci sorgusunu sisteme gönderir.

3. Sistem, kendisine sunulan sorgu için,
A. Sorgu elemanına ait seçilen öznitelik vektörünü (MPEG-7 ASF ya da MFCC) çıkarır.
B. Bu vektör ile ses veritabanındaki tüm öznitelik vektörleri arasındaki karşılıklı ilintiler hesaplanmak suretiyle karşılaştırma yapılır.
C. Karşılaştırma sonucunda, aday sesler ilgililik oranına göre (benzerlik oranı) sıralanır.
D. Sıralamada, verilen benzerlik aralığı içerisinde olan ses sinyal(ler)i sonuç olarak döndürülür.

Sonuç olarak sistemin bu özelliği, bir ses sinyalinin içeriğinin taranması ve ses ipuçları ile video verilerinin indekslenmesi gibi uygulamalarda büyük önem arzetmektedir.

4.4 En Yakın K-komşuluk Sorguları

En yakın k-komşuluk sorgularında, aralık sorgularına çok benzer bir yaklaşım kullanılmaktadır. İki sorgu biçimi arasındaki tek fark, sisteme benzerlik aralığı bilgisi yerine bir k tamsayısı verilmektedir. k tamsayısının anlamı, yapılan benzerlik sorgusu neticesinde, elde edilen aday sonuçlardan en yüksek benzerlik oranına sahip ilk k tane sonucun geri döndürülmesi şeklindedir. En yakın k-komşuluk sorguları için sistem ile kullanıcı etkileşimlerini gösteren UML *use-case* şeması Şekil 4.4'de verilmiştir.

4.5 Nakarat Sorguları

Üçüncü bölümde, ses içerik analizi çerçevesinde müzik ve konuşma verileri içerisindeki nakarat ve anahtar kavramların otomatik olarak çıkarılması için geliştirilen yöntem açıklanmıştı. Elde edilen bu bilgiler ile ses verilerinin yapısal ve ve anlamsal seviyede sorgulanması ve taranması mümkündür. Bir konuşma verisinin anahtar kavramlar üzerinden taranması ve sorgulanması, müzik verisinin nakarat bölümleri için taranması ve sorgulanması kullanıcı gereksinimlerine örnek olarak

verilebilmektedir. Bu gereksinimlere ihtiyaç duyulan uygulamalara, büyük miktarda ses verisi içeren sayısal arşivler (radyo ve televizyon arşivleri gibi) ve müzik telif hakkı uygulamaları örnek olarak verilebilmektedir. Özellikle müzik telif hakkı uygulamalarında nakarat sorguları önemli bir yere sahiptir. Çünkü, müzik verilerinden elde edilen nakarat bölümleri, her müzik için ayırtedici olmasının yanısıra tekil bir bilgidir. Böylece, bir müzik verisinin, başka bir müzik verisine benzeyip benzemediği, ya da hangi oranda benzediği ilgili verilerin sadece nakarat bölümlerinin sorgulanması neticesinde elde edilebilir. Aşağıdaki sorgular, geliştirilen sistem tarafından desteklenen nakarat ve anahtar kavram sorgularına örnek olarak verilebilir:

Sorgu 1: "Verilen sorgu nakaratına/anahtar kavramına en çok benzeyen ses verilerini getir" – Nokta sorgu biçimi, anlamsal sorgu.

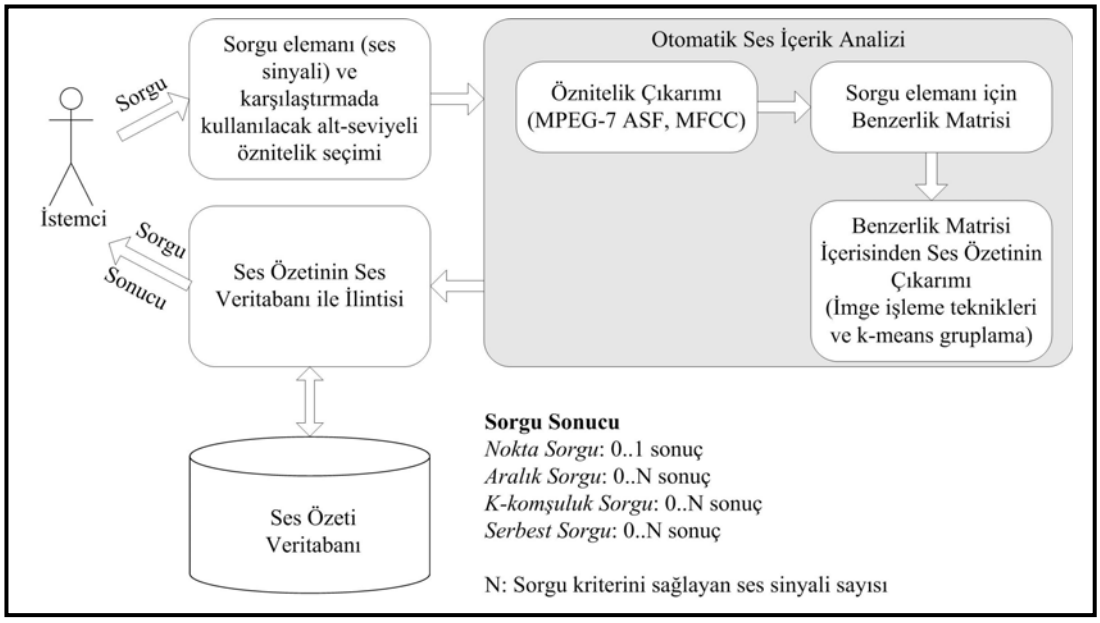
Sorgu 2: "Verilen sorgu nakaratına/anahtar kavramına %80 ile %90 arasında benzeyen ses verilerini getir" – Aralık sorgu biçimi, anlamsal sorgu.

Sorgu 3: "Verilen sorgu nakaratına/anahtar kavramına en çok benzeyen ilk 5 ses verisini getir" – En yakın k-komşuluk sorgu biçimi, anlamsal sorgu.

Sorgu 4: "Veritabanı elemanlarının verilen nakarat/anahtar kavrama olan benzerliklerini bul" – Serbest sorgu biçimi, anlamsal sorgu.

Yukarıda verilen anlamsal sorguların yanıtlanması için kullanılan yöntem Şekil 4.5'te gösterilmiştir. Buna göre, sorgunun yanıtlanması için sırasıyla aşağıdaki işlem adımları gerçekleştirilmektedir:

1. İstemci öncelikle aramak istediği sorgu elemanını, karşılaştırmada kullanılacak öznitelik tipini (MPEG-7 ASF ya da MFCC) ve sorgu tipini (nokta, aralık, k-komşuluk) belirler.
2. İstemci sorgusunu sisteme gönderir.



Şekil 4.5. Nakarat ve anahtar kavramlar için anlamsal sorgu yöntemi

3. Sorgu elemanına ait seçilen öznitelik vektörü (MPEG-7 ASF ya da MFCC) öznitelik çıkarım aracı tarafından çıkarılır.

4. Bu vektör için benzerlik matrisi yaratılır.

5. Benzerlik matrisi içerisindeki nakarat/anahtar kavram bilgisi Bölüm 3'te açıklanan ses içerik analizi yöntemiyle çıkarılır. Bu noktadan itibaren çıkarılan bilgi *ses özeti* olarak adlandırılacaktır.

6. Çıkarılan ses özeti bilgisi, ses özeti veritabanındaki özetler ile Eş. 4.1'deki yöntemle karşılaştırılarak benzerlikler hesaplanır.

7. Hesaplanan benzerlik değerleri ilgililik oranlarına göre sıralanır.

8. Sorgu 1 için (*nokta sorgu biçimi*),

A. Sıralamada en yüksek benzerlik değerine sahip ses sinyali sonuç olarak geri döndürülür.

9. Sorgu 2 için (*aralık sorgusu biçimi*),

A. Sıralamada, verilen benzerlik aralığı içerisinde olan ses sinyal(ler)i sonuç olarak geri döndürülür.

10. Sorgu 3 için (*en yakın k-komşuluk sorgu biçimi*),

A. Sıralamada, en çok benzeyen ilk k tane ses sinyali sonuç olarak geri döndürülür.

11. Sorgu 4 için (*serbest sorgu biçimi*),

A. Yedinci adımda elde edilen liste sonuç olarak geri döndürülür.

Önerilen yöntemin nakarat ve anahtar kavram sorgularına getirdiği için en önemli avantaj, karşılaştırmaların bir ses sinyalinin bütünü üzerinden değil, sadece önceden çıkarılmış özetleri üzerinden gerçekleştirilmesidir. Böylece, verilen sorgunun arama karmaşıklığı azaltılmaktadır.

5. SONUÇ VE ÖNERİLER

Bu çalışmada, ses içerik analizi çerçevesinde, müzik ve konuşma verileri içerisinde tekrar eden desenlerin otomatik olarak tespit edilmesini sağlayan yeni bir yöntem geliştirilerek, elde edilen sonuçların içerik tabanlı ses getirim ve tarama sistemlerine uygulanabilmesi için bir model önerilmiştir.

Müzik ve konuşma verileri içerisindeki tekrar eden desenleri tespit edebilmek amacıyla, *benzerlik matrisi* adı verilen bir gösterim esas alınmıştır [35]. Bu gösterimin özgün hali, müzik ve konuşma verileri içerisindeki sırasıyla nakarat ve anahtar kavram gibi tekrar eden yapıları otomatik olarak tespit edebilmek için uygun değildir. Bu nedenle, benzer çalışmalardan farklı olarak, çeşitli imge işleme teknikleri ses içerik analizine uygulanmıştır. Uygulanan imge işleme yöntemleriyle, ses verisi içerisindeki tekrar eden desenler daha belirgin hale getirilmiştir. Elde edilen deneysel sonuçlarda, tekrar eden desenlerin tanıma doğruluk oranlarının arttığı gözlemlenmiştir.

Tekrar eden desenlerin tespit edilmesi esnasındaki diğer önemli nokta da, bu desenlerin ses verisi içerisindeki zamansal yerlerinin düşük bir hatayla elde edilmesidir. Şöyle ki, benzerlik matrisi içerisindeki tekrar eden desenleri ifade eden köşegensel çizgilerin sınırları ilgili desenin ses verisi içerisindeki yerini göstermektedir. Bu sınırların, benzerlik matrisi içerisinden doğru biçimde tespit edilebilmesi için uygun eşik değerlerinin kullanılması gerekmektedir. Ancak, bu eşik değerleri farklı ses verileri için farklı değerler alabilmektedir. Bu nedenle uyarlanabilir eşik değerlerine ihtiyaç duyulmaktadır. Aksi takdirde, analiz edilen her ses verisi için eşik değerinin parametre olarak girilmesine ihtiyaç duyulacaktır. Bu çalışmada geliştirilen algoritma ile, her ses verisi için uyarlanabilir eşik değerleri elde edilebilmektedir [29]. Sonuç olarak, ses sinyali içerisinde tekrar eden desenlerin zamansal yerleri en fazla $\pm 360\,ms$ sapmayla tespit edilebilmektedir.

Geliştirilen ses içerik analizi yönteminin öznitelik altyapısında MFCC ve MPEG-7 ASF öznitelik aileleri kullanılmıştır. MFCC öznitelik ailesi, OST birliği tarafından

yaygın olarak kullanılan spektral sınıfa ait bir öznitelik iken, MPEG-7 ASF özniteliği nispeten çok daha yeni bir spektral özniteliği olup, ses içerik analizi ve içerik-tabanlı ses getirim sistemlerinde MFCC özniteliği kadar kullanılmamıştır. Her iki özniteliğin geliştirilen sistemin öznitelik altyapısında kullanılmasıyla, uygun bir karşılaştırma ortamı da sağlanmıştır.

Deneysel çalışmalar neticesinde, müzik ve konuşma verileri içerisindeki nakarat ve anahtar kavramlar MPEG-7 ASF özniteliği kullanıldığında sırasıyla %86 ve %78 tanıma doğruluk oranları ile elde edilmiştir. MFCC öznitelik ailesi kullanıldığında bu değerlerin müzik verisi için %82, konuşma verisi için %87 olduğu gözlemlenmiştir. Bu sonuçlardan, MPEG-7 ASF özniteliğinin müzik verisi içerisindeki nakaratları önerilen yöntemle MFCC özniteliğine kıyasla daha iyi çıkardığı görülmektedir. Bu durum konuşma verisi için tam tersidir, yani MFCC özniteliği, konuşma verileri içerisindeki anahtar kavramları daha doğru bir şekilde çıkarmaktadır. Toparlamak gerekirse, MPEG-7 ASF ve MFCC öznitelik ailelerinin müzik ve konuşma verileri içerisindeki tekrar eden desenlerin çıkarılması kapsamında elde edilen sonuç ve tecrübeler aşağıdaki şekilde listelenebilir:

1. Müzik ses sinyalleri için, MPEG-7 ASF özniteliğinin MFCC özniteliğine kıyasla daha iyi bir tanıma doğruluk oranına sahip olduğu tecrübeler neticesinde gözlemlenmiştir.
2. Konuşma ses sinyalleri için, MFCC özniteliğinin MPEG-7 ASF özniteliğine kıyasla daha iyi bir tanıma doğruluk oranına sahip olduğu tecrübeler neticesinde gözlemlenmiştir.
3. MFCC özniteliğinin MPEG-7 ASF özniteliğine kıyasla, benzerlik matrisi içerisindeki zemin gürültülerine karşı daha güçlü olduğu gözlemlenmiştir. MFCC özniteliğinin bu özelliği, desen çıkarım algoritmasının doğruluk oranını yükseltebilmektedir. Ancak, müzik verileri için genel tanıma doğruluk oranı düşük olduğundan MPEG-7 ASF özniteliği tercih edilmiştir.
4. MFCC özniteliği, spektral alt bantları belirlemek için Mel ölçeğini kullanırken, MPEG-7 ASF özniteliği logaritmik ölçek kullanmaktadır. Bu özelliğiyle, MFCC özniteliği konuşma verilerinde daha yüksek bir tanıma doğruluk oranına sahiptir.

Çünkü konuşma sesinin büyük çoğunluğu 1 *kHz'in* altında olup, MFCC özniteliği bu frekans aralığı için doğrusal ölçek kullanmaktadır.
5. Önerilen analiz aşamasında, MPEG-7 ASF özniteliğinin MFCC özniteliğine kıyasla daha az zaman ve bellek gereksinimi olduğu gözlemlenmiştir.

Bu avantaj ve dezavatajlar çerçevesinde, müzik verilerinin içerik tabanlı analiz ve getiriminde MPEG-7 ASF özniteliğinin; konuşma verilerinin içerik tabanlı analiz ve getiriminde ise MFCC özniteliğinin kullanılmasının daha uygun olacağı düşünülmektedir.

Bu tezde ayrıca, elde edilen yapısal analiz sonuçlarını kullanan ve ses verisinin çok farklı şekillerde sorgulanmasını ve taranmasını sağlayan bir model önerilmiştir. Önerilen model nakarat sorguları, ses efektlerinin sorgulanması ve örnek vererek sorgulama (ÖVS) gibi sorgu biçimlerini desteklemek suretiyle, ses verisinin yapısal ve anlamsal seviyede esnek biçimde sorgulanabilmesini ve taranmasını sağlamaktadır. Önerilen model, çokluortam bilgilerin sorgulanması esnasında çok önemli bir yere sahip olan nokta, aralık ve en yakın k-komşuluk sorgu ve tarama biçimlerini de desteklemektedir. Bu sorgu biçimlerinden aralık ve k-komşuluk sorgularının bir ses sinyali için yanıtlanması esnasında, daha önceden oluşturulan benzerlik matrisinden yararlanılmaktadır. Bu yaklaşım sayesinde, arama esnasındaki desen eşleme karmaşıklığı ortadan kaldırılmaktadır.

Bu tez çalışmasında elde edilen sonuçların birçok uygulama alanı bulunmaktadır. Bu alanlara içerik-tabanlı ses getirim sistemleri, ses çalıcı uygulamaları (bir bölümden diğer bölümlere hızlı biçimde geçiş), ses parmak izi uygulamaları [1], ses işaretleme araçları ve ses ipuçlarının kullanımı ile video dizinleme ve sınıflandırma uygulamaları örnek olarak verilebilir.

Gelecek çalışma planları arasında, önerilen sorgu modeline uygun bir sorgu dilinin geliştirilerek Web tabanlı bir sorgu arayüzüyle entegrasyonu ve ses içerik analizinin video verilerine uygulanması yer almaktadır.

KAYNAKLAR

1. Sert, M., Baykal, B., Yazıcı, A., "A robust and time-efficient fingerprinting model for musical audio", ***IEEE International Symposium on Consumer Electronics***, St. Petersburg, Russia, 237-242 (2006).

2. Lloyd, S.P., "Least squares quantization in PCM", ***IEEE Transaction on Information Theory***, IT(2):129-137 (1982).

3. Cheng, Y., "Content-based musical retrieval on acoustical data", Doktora Tezi, ***Stanford Üniversitesi***, 1-7 (2003).

4. Smith, M.A., Kanade, T., "Multimodel video characterization and summarization", ***The Kluver International Series in Video Computing*** (2004).

5. Adams, W.H., Lyengar, G., Lin, C-Y., Naphade, M.R., Neti, C., Nock, H.J., Smith, J.R., "Semantic indexing and multimedia content using visual, audio, and text cues", ***EURASIP Journal on Applied Signal Processing,*** 2003(2):170-185 (2003).

6. Nam, J., Çetin, A.E., Tevfik, A.H., "Speaker identification and video analysis for hierarchical video shot classification", ***IEEE International Conference on Image Processing***, Santa Barbara, CA, 550-555 (1997).

7. Wang, Y., Liu, Z., Huang, J-C., "Multimedia content analysis using both audio and visual clues", ***IEEE Signal Processing Magazine***, 12-36 (2000).

8. Aigrain, P., Zhang, H., Petkovic, D., "Content-based representation and retrieval of visual media: A state-of-the-art review", ***Multimedia Tools and Applications***, 3(3):179-202 (2006).

9. Bach, J.R., "The virage image search engine: An open framework for image management", ***In proceedings of SPIE'96***, San Jose, CA, 76-87 (1996).

10. Niblack, W., Zhu, X., "Updates to the QBIC system", ***In proceedings of SPIE'98***, San Jose, CA, 150-161 (1998).

11. Amato, G., Mainetto, G., Savino, P., "An approach to a content-based retrieval of multimedia data", ***Multimedia Tools and Applications***, 7(1/2): 9-36 (1998).

12. Lu, L., Jiang, H., Zhang, H., "A robust audio classification and segmentation method", ***In Proceedings of the ACM International Conference on Multimedia***, Ottowa, Canada, 203-211 (2001).

13. Tzanetakis, G., Cook, P., "Multifeature audio segmentation for browsing and annotation", ***In IEEE WASPAA Conference***, New Paltz, NY, 103-106 (1999).

14. Pfeiffer, S., "Pause concepts for audio segmentation at different semantic levels", ***In Proceedings of the ACM International Conference on Multimedia***, Ottowa, Canada, 187-193 (2001).

15. Chai, W., Vercoe, B., "Structural analysis of musical signals for indexing and thumbnailing", ***In Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries***, Houston, Texas, 223-226 (2003).

16. Cooper, M., Foote, J., "Summarizing popular music via structural similarity analysis", ***In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics***, New Paltz, NY, 127-130 (2003).

17. Chai, W., Vercoe, B., "Music thumbnailing via structural analysis", ***In Proceedings of ACM Multimedia Conference*** , New Paltz, NY, 223-226 (2003).

18. Bartsch, M., Wakefield, G., "Audio thumbnailing of popular music using chroma-based representations", ***IEEE Transaction on Multimedia***, 7(1):96-104 (2005).

19. Aucouturier, J.J., Sandler, M., "Finding repeating patterns in acoustic musical signals: Applications for audio thumbnailing", ***AES International Conference on Virtual, Synthetic and Entertainment Audio***, Espoo, Finland, 204-207 (2002).

20. Xu, C., Zhu, Y., Tian, Q., "Automatic music summarization based on temporal, spectral and cepstral features", ***IEEE International Conference on Multimedia & Expo***, Lusanne, Switzerland, 117-120 (2002).

21. Goto, M., "A chorus-section detection method for musical audio signals", ***In IEEE International Conference on Acoustics, Speech and Signal Processing***, Hong Kong, China, 437-440 (2003).

22. Bartsch, M., Wakefield, G., "To catch a chorus: Using chroma-based representations for audio thumbnailing", ***In IEEE Workshop on Applications of Signal Processing on Audio and Acoustics***, New Paltz, NY, 15-18 (2001).

23. Lu, L., Wang, M., Zhang, H.J., "Repeating pattern discovery and structure analysis from acoustic music data", ***In Proceedings of the ACM International Conference on Multimedia***, New York, NY, 275-282 (2004).

24. Dannenberg, S.P., "Pattern discovery techniques for music audio", ***Journal of New Music Research***, 32(2):153-163 (2003).

25. Peeters, G., Burthe, A.L., Rodet, X., "Toward automatic music audio summary generation from signal analysis", ***International Conference on Music Information Retrieval and Related Activities***, Paris, France, 86-92 (2002).

26. Lu, L., Wenyin, L., Zhang, H.J., "Audio Textures: Theory and applications", ***IEEE Transactions on Speech and Audio Processing***, 12(2):156-167 (2004).

27. Wold, E., Blum, T., Keislar, D., "Content-based classification, search and retrieval of audio", ***IEEE Multimedia***, 3(3): 27-36 (1996).

28. Zhang, T., Jay Kuo, C.-C., "Content-based classification and retrieval of audio", ***In Proceedings of SPIE'98***, San Diego, 211-222 (1998).

29. Sert, M., Baykal, B., Yazıcı, A., "Generating expressive summaries for speech and musical audio using self-similarity clues", ***In IEEE International Conference on Multimedia & Expo***, Toronto, Ontario, Canada, 169-172 (2006).

30. MPEG-7, "MPEG-7 requirements document v8", ***ISO/IEC JTC1/SC29/WG11/N2727***, ***Seoul Meeting***, 48-97 (1999).

31. Burnett, I., Jackson, M., Lindsay, A.T., Quackenbush, S., "Fundamentals of audio descriptions, an introduction to MPEG-7 multimedia content description interface", ***John Wiley & Sons Ltd***, England, 283-298 (2002).

32. Steelant, D., DeBaets, B., DeMeyer, H., Leman, M., Martens, S.-P., Clarisse, L., Lesaffre, M., "Discovering structure and repetition in musical audio", ***In Eurofuse***, Varanna, Italy, 11-17 (2002).

33. Logan, B., Chu, S., "Music summarization using key phrases", ***In IEEE International Conference on Acoustics, Speech and Signal Processing***, İstanbul, Türkiye, 749-752 (2000).

34. Rabiner, L., Juang, B.H., "Fundamentals of speech recognition", ***Prentice Hall***, 98-99 (1993).

35. Foote, J., "Visualizing music and audio using self similarity", ***ACM Multimedia***, 1:77-80 (1999).

36. Wellhausen, J., Crysandt, H., "Temporal audio segmentation using MPEG-7 descriptors", ***In Proceedings of Storage and Retrieval for Media Databases***, Santa Clara, CA, 380-387 (2003).

37. Lloyd, S.P., "Least squares quantization in PCM", IEEE Transaction on Information Theory, IT(2):129-137 (1982).

38. Ziou, D., Tabbone, S., "Edge detection techniques – an overview", ***Pattern Recognition and Image Analysis***, 8(4):537-559 (1998).

39. Sharifi, M., Fathy, M., Tayefeh, M., "A classified and comparative study of edge detection algorithms", ***In Proceedings of the International Conference on Information Technology: Coding and Computing***, CA, 117-120 (2002).

40. Gonzales, R.C., Woods, R.E., "Digital Image Processing 2nd ed.", ***Prentice Hall***, New Jersey, 136-138 (2002).

41. Gonzales, R.C., Woods, R.E., "Digital Image Processing 2nd ed.", ***Prentice Hall***, New Jersey, 578-579 (2002).

42. Kirsch, R.A., "Computer determination of the constituent structure of biological images", ***International Journal of Computers and Biomedical Research***, 4(3):315-328 (1971).

43. Canny, J.F., "A computational approach to edge detection", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 8(6):679-698 (1986).

44. Marr, D., Hidreth, E.C., "Theory of edge detection", ***In Proceedings of the Royal Society***, London B, 187-217 (1980).

45. Shen, J., Castan, S., "An optimal linear operator for step edge detection", ***Computer Vision, Graphics, and Image Processing: Graphical Models and Understanding***, 54(2):112-133 (1992).

46. İnternet: Pollack, A.W., "Notes on series", http://www.icce.rug.nl/soundspaces/-DATABASES/AWP/awpnotes_on.html/ (2006).

47. İnternet: NTVMSNBC, http://www.ntvmsnbc.com.tr/ (2006).

48. Gudivada, V., Raghavan, V., "Content-based image retrieval systems: Guest editor's introduction", ***IEEE Computer Magazine***, 18-22 (1995).

49. Ghias, A., Logan, J., "Query-by-humming musical information retrieval in an audio database", ***ACM Multimedia Conference***, San Jose, CA, 231-236 (1995).

50. Lu, L., Zhang, H., "Content-based audio classification and segmentation by using support vector machines", ***Multimedia Systems***, (8):482-492 (2003).

51. Lu, G., "Indexing and retrieval of audio: A survey", ***Multimedia Tools and Applications***, 15(3):269-290 (2001).

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı	: SERT, Mustafa
Uyruğu	: T.C.
Doğum tarihi ve yeri	: 13.12.1974 Mersin
Medeni hali	: Bekar
Telefon	: 0 (312) 234 10 10
Faks	: 0 (312) 234 10 51
e-mail	: msert@baskent.edu.tr.

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek lisans	Gazi Üniversitesi /Elektronik-Bilg. Eğt.	2001
Lisans	Gazi Üniversitesi/ Elektronik-Bilg. Eğt.	1997
Lise	Mersin Anadolu Tek. Lisesi, Bilg. Böl.	1992

İş Deneyimi

Yıl	Yer	Görev
1997-2006	Başkent Üniversitesi	Öğretim Görevlisi

Yabancı Dil

İngilizce

Yayınlar

1. Sert, M., Baykal, B., Yazıcı, A., "Structural and semantic modeling of audio for content-based querying and browsing", In Proceedings of International Conference on Flexible Query Answering Systems, Lecture Notes in Artificial Intelligence (LNAI), Milano, Italy, 319-330 (2006).

2. Sert, M., Baykal, B., Yazıcı, A., "Generating expressive summaries for speech and musical audio using self-similarity clues", In IEEE International Conference on Multimedia & Expo, Toronto, Ontario, Canada, 169-172 (2006).

3. Sert, M., Baykal, B., Yazıcı, A., "A robust and time-efficient fingerprinting model for musical audio", In IEEE International Symposium on Consumer Electronics, St. Petersburg, Russia, 237-242 (2006).

4. Sert, M., Baykal, B., "Web-based Query Engine for Content-based and Semantic Retrieval of Audio", In Proceedings of The Eighth IEEE Int. Symp. on Consumer Electronics (ISCE 2004), Reading, UK, 485-490 (2004).

5. Sert, M., Baykal, B., "An Approach to the Semantic Modeling of Audio Databases", In Proceedings of Int. Conf. on E-Business and Telecommunication Networks (ICETE'04), Setúbal, PORTUGAL, 385-392 (2004).

6. Sert, M., Baykal, B., "A Web Model for Querying, Storing, and Processing Multimedia Content", In Proceedings of Information and Knowledge Sharing (IASTED - IKS'03), AZ, USA, 277-282 (2003).

Hobiler

Bilgisayar teknolojileri, Futbol, Masa tenisi, Yüzme, Kitap okuma, Sinemaya gitme