Görsel – İşitsel Konuşma Tanıma'da Veri Kaynaştırma Teknikleri

Information Fusion Techniques in Audio-Visual Speech Recognition

H. Karabalkan[1,2], H. Erdoğan[1]

1. Mühendislik ve Doğa Bilimleri Fakültesi
Sabancı Üniversitesi
karabalkan@su.sabanciuniv.edu, haerdogan@sabanciuniv.edu

2. TÜBİTAK Ulusal Elektronik ve Kriptoloji Araştırma Enstitüsü
karabalkan@uekae.tubitak.gov.tr

Özetçe

İnsanın sesi algılamada işitsel bilginin yanında görsel bilgiyi de kullandığı bilinmektedir ancak farklı veri kanallarından gelen bilgiyi nasıl birleştirdiği belirsizliğini korumaktadır. Belirsizikle birlikte Görsel – İşitsel Konuşma Tanıma'daki veri kaynaştırma sürecine duyulan ilgi de artmaktadır. Bu çalışmada, Ardışık Karma Yöntem (AKY) olarak adlandırılan veri kaynaştırma tekniği ile Görsel – İşitsel Konuşma Tanıma'da yeni bir yaklaşım sunulmakta ve bu yaklaşım etkili ve sık kullanılan bir veri kaynaştırma tekniği olan Çok Akımlı Saklı Markov Model (Multiple Stream Hidden Markov Model - MSHMM) ile karşılaştırılmaktadır.

Abstract

It is well known that human perception of speech relies both on audio and visual information. However, the physiology of information fusion process in humans is still indefinite which attracts scientists' attention to information fusion process for Audio-Visual Speech Recognition. In this work, a novel tandem hybrid approach is introduced for an efficient Audio – Visual Speech Recognition system and the performance of the proposed technique is experimentally compared with the widely used Multiple Stream Hidden Markov Model (MSHMM) approach.

1. Giriş

Konuşma, insanın çevresiyle etkileşiminde en sık başvurduğu araçlardan biri olması dolayısıyla insan – bilgisayar arayüzleri açısından kritik önemdedir. Konuşma tanıma sistemleri için, sesin modellenmesinde oldukça etkili teknikler önerilmiştir. Ancak gürültüsüz ortamlarda başarılı sonuçlar veren bu teknikler gürültü seviyesinin artmasıyla ciddi performans kayıplarına maruz kalmaktadır. Oysa konuşma tanımaya ihtiyaç duyulan ortamların gürültüsüz olması garanti edilemez. Problemin çözümü yine insan fizyolojisinde yatmaktadır. İnsan, sesi algılamada işitsel bilginin yanında görsel bilgiyi de kullanır. Hatta görsel bilginin yardımcı değil bütünleyici bilgi olduğu Mac Gurk tarafından kanıtlanmıştır [1].

Görsel – İşitsel Konuşma Tanıma Sistemleri, işitsel bilginin yanında görsel bilgiden de faydalanarak gürültü seviyesinin arttığı ortamlarda da yüksek tanıma oranlarını hedeflemektedir. Bir Görsel – İşitsel Konuşma Tanıma Sistemi, üç alt yapıdan oluşmaktadır. Birincisi, işitsel bilginin analizi, ikincisi görsel bilginin analizi ve üçüncüsü iki bilgi akımının birleştirilmesi ya da kaynaştırılmasıdır.

İşitsel bilginin analizinde, gürültüden daha az etkilenen işitsel özniteliklerin çıkarılması konusunda çeşitli çalışmalar sürdürülse de, Mel Frekansı Kepstral Katsayıları (Mel Frequency Cepstral Coefficients – MFCC) gürültüsüz durumlardaki başarısıyla ETSI (European Telecommunications Standard Institute) standardı kabul edilmiştir [2]. Bu çalışmada da işitsel öznitelik olarak MFCC'ler tercih edilmiştir.

Görsel bilginin analizinde, işitsel analizde olduğu gibi standart kabul edilebilecek teknikler olmasa da çalışmaların yoğunlaştığı yöntemler mevcuttur. Bu çalışmada, basit ve etkili bir görsel öznitelik olarak kabul gören Ayrık Kosinüs Dönüşümü (Discrete Cosine Transform – DCT) katsayıları tercih edilmiştir.

Makalenin odaklandığı nokta ise veri kaynaştırma sürecidir. İşitsel ve görsel öznitelik vektörlerinin, Saklı Markov Modelleri (Hidden Markov Models – HMM) ile modellenmeden önce ön sınıflandırıcı aşamasından geçirilmesine dayanan veri kaynaştırma yöntemleri Ardışık Karma Yöntem (AKY) olarak isimlendirilmektedir. Bilimsel yazında, çeşitli AKY'ler ileri sürülmüştür [6,7]. Bu çalışmada da, veri akımlarının birbirinden bağımsız olarak Gauss Karışımı Modeli (Gaussian Mixture Model – GMM) sınıflandırıcıları ile sınıflandırıldığı ve sonrasında iki sınıflandırıcının Doğrusal Ayırtaç Analizi (Linear Discriminant Analysis – LDA) sınıflandırıcısı ile birleştirildiği bir AKY önerilmektedir. Önerilen algoritma, MSHMM ile karşılaştırılmaktadır.

Makale, giriş bölümüyle birlikte altı bölüme ayrılmıştır. İkinci bölümde işitsel öznitelik çıkarılması, üçüncü bölümde görsel öznitelik çıkarılması, dördüncü bölümde işitsel ve görsel veri akımlarının birleştirilmesi ve HMM modelleme

978-1-4244-4436-6/09/$25.00 ©2009 IEEE

anlatılmıştır. Deneysel sonuçlar beşinci bölümde analiz edilerek, altıncı bölümde vargılar irdelenmiştir.

2. İşitsel Öznitelik Çıkarımı

İşitsel öznitelik olarak ETSI standardı kabul edilen Mel Frekansı Kepstral Katsayıları (Mel Frequency Cepstral Coefficients – MFCC) kullanılmıştır [2]. Ses işaretinin Mel ölçeğinde kepstral analizi ile elde edilen MFCC'ler, insanlarda doğrusal olmayan frekans algısını modellemedeki başarısı dolayısıyla sık kullanılan işitsel özniteliklerdir.

Sesin analizi için genellikle 10ms'de bir alınan 25ms uzunlukta çerçeveler kullanılmaktadır ancak Görsel – İşitsel Konuşma Tanıma'da işitsel ve görsel bilginin senkronizasyonuna ihtiyaç duyulduğundan bu çalışmada 40ms'de bir alınan 100ms uzunluktaki çerçeveler tercih edilmiş ve böylece 25fps'lik görsel bilgiyle senkronizasyon sağlanmıştır.

MFCC'lerin çıkarımında şu temel adımlar atılır: Her bir ses çerçevesine Fourier Dönüşümü uygulanarak frekans spektrumu bulunur. Spektrum, Mel ölçeğine izdüşürülerek logaritması alınır ve ardından DCT uygulanır. Sonuçta ulaşılan DCT katsayılarının genlikleri MFCC'lerdir.

Öznitelik olarak alınan MFCC sayısı, spektrumun gösterimindeki hassasiyeti belirler. MFCC sayısı arttıkça hassasiyet artar. Genellikle ilk 12 MFCC (spektrumdaki düşük frekans katsayıları) ve çerçevedeki enerji alınarak, her çerçeve için 13 boyutlu statik öznitelik vektörü oluşturulur. Dinamik bilgiyi de modellemek için, 13 boyutlu öznitelik vektörünün komşu çerçevelerle birinci ve ikinci türevleri de çıkarılır ve neticede her çerçeve için 39 boyutlu işitsel öznitelik vektörü elde edilir.

3. Görsel Öznitelik Çıkarımı

Görsel öznitelik çıkarımında, işitsel öznitelik çıkarımında olduğu gibi standart haline gelmiş teknikler olmasa da, çalışmaların yoğunlaştığı algoritmalar mevcuttur. Görsel öznitelik çıkarma metotları iki farklı kategoride incelenebilir:

1. Şekil temelli öznitelikler
2. Bölge temelli öznitelikler

Şekil temelli öznitelik olarak, ağzın dik ve yatay açıklık miktarları, ağzın açıklık açısı gibi ölçümler ya da dudak şeklinin parametrik gösteriminin parametreleri kullanılmaktadır. Ancak, şekil temelli özniteliklerin başarısı, dudak çevritinin takip edilmesindeki başarıya bağlıdır ve takip algoritmalarındaki küçük sapmalar dahi tanıma oranlarında büyük hatalara sebep olabilmektedir.

Bölge temelli öznitelikler için ise dudak çevritinin takip edilmesine gerek olmaksızın dudak bölgesinin içinde kalan piksel değerleri kullanılır. Dolayısıyla, öznitelik çıkarımına geçmeden önce en uygun ilgi bölgesi tayin edilmelidir. İlgi bölgesi içinde kalan tüm pikseller üzerinde yapılacak istatistiksel analizde yüksek boyut problemiyle karşılaşılacağından, ilgi bölgesine çeşitli boyut indirgeme yöntemleri uygulanır. Boyut indirgeme, işlemsel yükü azalttığı gibi tanıma sisteminin konuşmacıdan bağımsız olmasına da katkı sağlar.

3.1. İlgi Alanının Çıkarılması

Görsel ses bilgisinin büyük bir kısmının, burnun ucunu ve çeneyi de kapsayan bir dudak bölgesinde olduğu bilinmektedir. Hemen hemen tüm çalışmalarda ilgi alanı değişen boyutlarda da olsa dikdörtgen olarak seçilmiştir. Bu çalışmada dudak bölgesi, yüzün dikey olarak altta kalan %40'lık ve yatay olarak ortadaki %50'lik kısmı kabul edilmiştir. Yüz sezimi için Viola ve Jones'un görsel nesne sezimi metodu kullanılmıştır [3]. Ardışık video kareleri arasındaki sürekliliği sağlamak ve yüz sezimi algoritmasından kaynaklanabilecek sapmaları en aza indirgemek için video kareleri arasındaki ilinti kullanılabilir. Bu nedenle ele alınan video karesindeki dudak bölgesinin bir önceki karedeki dudak bölgesiyle ilintisine bakılmıştır. Her video karesi için faklı boyutta bulunabilecek olan dudak bölgeleri doğrusal ara değerleme ile yeniden boyutlandırılarak 48x64 dudak videoları elde edilmiştir.

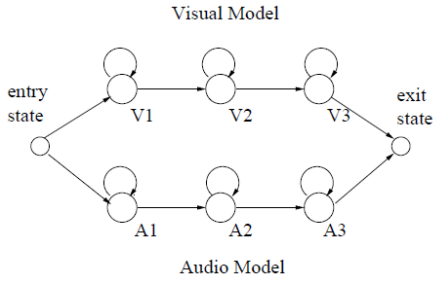
3.2. Ayrık Kosinüs Dönüşümü (DCT)

İlgi alanının saptanmasından sonra boyut indirgeme işlemine geçilir. Ayrık Kosinüs Dönüşümü (DCT) imge sıkıştırmada olduğu gibi görsel konuşma tanımada da etkilidir. İlk olarak Potamianos tarafından konuşma tanımaya uygulanan DCT'nin şekil temelli özniteliklere olan üstünlüğü saptanmıştır [4]. DCT, enerji sıkıştırmadaki başarısının yanında gerçekleme hızı nedeniyle de tercih edilen bir yöntemdir. Bu çalışmada, gri ölçekli dudak imgelerine iki boyutlu DCT uygulanmıştır. Statik öznitelikler olarak ilk 25 alçak frekans DCT katsayıları alınmıştır. İşitsel öznitelik çıkarımında olduğu gibi görsel özniteliklerin de birinci ve ikinci türevleri alınarak neticede 75 boyutlu görsel öznitelik vektörü elde edilmiştir.

4. Veri Kaynaştırma

Daha önce de belirtildiği gibi, insanların konuşma tanıma için işitsel ve görsel bilgiyi nasıl kaynaştırdığı bilinmemektedir. Bu durum, görsel-işitsel konuşma tanıma konusunda çalışma yapanları çeşitli veri kaynaştırma tekniklerini deneyerek karşılaştırmaya yöneltmektedir.

Etkili ve sık kullanılan veri kaynaştırma tekniklerinden biri de Çok Akımlı Saklı Markov Model (Multiple Stream Hidden Markov Model - MSHMM)'lerdir. MSHMM'ler ile birden çok veri akımının farklı ağırlıklar verilerek paralel olarak modellenmesi mümkün olmaktadır [5]. Şekil-1'de 3 durumlu bir MSHMM topolojisi verilmektedir.



Şekil-1 : MSHMM topolojisi

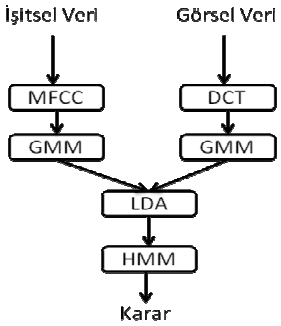
Tek akımlı HMM'lerden farklı olarak MSHMM'lerde, *t* anındaki bir gözlemin *j.* durum tarafından gözlemlenme olasılığı $b_j(o_t)$, tüm veri akımlarındaki gözlemlerin *j.* durum tarafından gözlemlenme olasılıkları $b_{js}(o_{ts})$ 'lerin kombinasyonu olarak ifade edilir. Denklem-1'de de görüldüğü

gibi her akımın bir w_s ağırlığı vardır ve tüm akımların ağırlıkları toplamı 1'e eşittir.

$$b_j(o_t) = \prod_s [b_{js}(o_{ts})]^{w_s} \quad (1)$$

Bu çalışmada önerilen Ardışık Karma Yöntem (AKY), var olan MSHMM yöntemiyle karşılaştırılmaktadır. AKY, öznitelik vektörlerinin HMM'e gözlem vektörü olarak verilmeden önce bir ön sınıflandırma sürecinden geçirilmesi esasına dayanır. Ön sınıflandırma neticesinde elde edilen ardıl olasılık vektörü HMM'e gözlem vektörü olarak sürülür. Bu yöntem çeşitli çalışmalarda incelenmiş ve doğrudan öznitelik vektörlerinin kullanıldığı yöntemlere üstünlükleri kanıtlanmıştır [6,7].

AKY'de atılan adımlar şunlardır: İlk olarak, her bir akım için bağımsız sınıflandırıcılar eğitilir. Daha sonra bu sınıflandırıcılar bir sınıflandırıcı birleştiricisi ile birleştirilir. İşitsel ve görsel öznitelik vektörlerinin bir biri ardına eklenmesiyle oluşturulan görsel-işitsel öznitelik vektörü eğitilmiş olan birleşik sınıflandırıcıdan geçirilir. Sınıflandırıcı çıktısı ardıl olasılık vektörüdür ve bu vektörün boyutu toplam sınıf sayısı kadardır.



Şekil-2 : AKY topolojisi

Ardışık Karma yöntemler, kullanılan tekil sınıflandırıcılar ve sınıflandırıcı birleştiricilerine göre farklılık gösterebilir [6,7]. Bu çalışmada her bir veri akımı Gauss Karışımı Modeli (Gaussian Mixture Model - GMM) sınıflandırıcısı ile eğitilmiş ve ardından her iki veri akımına ait GMM'ler, Doğrusal Ayırtaç Analizi (Linear Discriminant Analysis - LDA) sınıflandırıcısı ile birleştirilmiştir. 75 boyutlu görsel öznitelik vektörünün, 39 boyutlu işitsel öznitelik vektörüne eklenmesiyle elde edilen 114 boyutlu görsel-işitsel öznitelik vektörü GMM-LDA birleşik sınıflandırıcısına verilmiş ve 11 boyutlu ardıl olasılık vektörü elde edilmiştir. Her bir veri çerçevesinden elde edilen 11 boyutlu ardıl olasılık vektörü HMM için bir gözlem vektörüdür. Önerilen AKY'in topolojisi Şekil-2'de görülmektedir.

MSHMM'de 114 boyutlu öznitelik vektörleri kullanılırken AKY'de 11 boyutlu öznitelik vektörleri kullanılmaktadır. Bu da, test sürecinde AKY'in çok daha hızlı gerçeklenebilmesine olanak sağlamaktadır. AKY'deki GMM ve LDC aşamalarının işlem yükü ise HMM aşamasına kıyasla ihmal edilebilir. Bunun sebebi, GMM aşamasında her sınıf için bir GMM eğitilirken HMM aşamasında her HMM durumu için bir GMM eğitilmesidir. LDC aşamasında da yine her sınıf için bir Gauss dağılımı kullanılmaktadır.

5. Sonuçlar

Deneyler, M2VTS veritabanı üzerinde gerçekleştirilmiştir [8]. M2VTS veritabanı, 37 farklı konuşmacının rakamları Fransızca olarak 0'dan 9'a sırayla seslendirdiği video'lardan oluşmaktadır ve farklı zamanlarda kayıtların tekrarlandığı 5 bant bulunmaktadır. Sadece işitsel bilginin kullanıldığı ve sadece görsel bilginin kullanıldığı tanıma deneylerinde 4 bant eğitim verisi olarak, sonuncu bant ise test verisi olarak kullanılmıştır. MSHMM deneyinde 3 bant eğitim için, 4. bant veri akımlarının çeşitli gürültü seviyelerindeki ağırlıklarının saptanması için ayrılmış ve 5. bant test için bırakılmıştır. AKY deneyinde ise GMM sınıflandırıcıları 3 bant, LDA birleştiricisi 4. bant ile eğitilmiş ve testler 5. bant üzerinde gerçekleştirilmiştir.

Gürültülü ortam performanslarını incelemek için ofis ortamında alınan kayıtlara 20dB'den -5dB'ye kadar değişen seviyelerde araba gürültüsü eklenmiştir.

Sınırlı sayıda sınıf bulunduğu için, 10 rakam ve "sessizlik" olmak üzere toplam 11 sınıf, kelime tabanlı tanıma tercih edilmiş ve 10 durumlu, 12 Gauss karışımlı HMM'ler eğitilmiştir. MSHMM'de akımların ağırlıkları her gürültü seviyesi için tanıma oranı performansına bakılarak saptanmıştır. Kullanılan ağırlıklar Tablo-1'de görülmektedir.

Gürültü Miktarı	İşitsel Akım Ağırlığı	Görsel Akım Ağırlığı
Yok	1.0	0.0
20dB	1.0	0.0
15dB	1.0	0.0
10dB	0.8	0.2
5dB	0.4	0.6
0dB	0.2	0.8
-5dB	0.0	1.0

Tablo-1 : MSHMM için akım ağırlıkları

Tablo-2'de verilen farklı gürültü seviyeleri için tanıma oranları incelendiğinde işitsel bilginin gürültü seviyesinin artmasıyla birlikte yetersiz kaldığı ve görsel bilginin işitsel gürültüden etkilenmediği görülmektedir. İşitsel ve görsel bilginin ikisinden de faydalanılan MSHMM ve AKY deneylerinde, tek akımın kullanıldığı deneylere göre üstünlük aşikardır. 5dB gürültü seviyesine kadar MSHMM ve AKY'nin yakın performans gösterdiği ancak 0dB ve üzerindeki gürültü seviyelerinde MSHMM'in işitsel bilginin olumsuz etkisini daha iyi bastırdığı söylenebilir.

Gürültü	İşitsel	Görsel	MSHMM	AKY
Yok	100.00	73.10	100.00	99.35
20dB	100.00	73.10	100.00	98.06
15dB	98.48	73.10	98.29	97.41
10dB	90.91	73.10	92.00	92.31
5dB	53.89	73.10	79.09	83.33
0dB	19.00	73.10	74.74	60.00
-5dB	10.00	73.10	71.92	40.00

Tablo-2 : Farklı gürültü seviyeleri için tanıma oranları (%)

Tanıma performanslarının yanında, iki yöntem çalışma hızları açısından da karşılaştırılmıştır ve öngörüldüğü gibi AKY'in test sürecinin MSHMM'in test sürecine göre daha az zaman aldığı tespit edilmiştir. Tablo-3'de tüm test verisinin işlenmesi için geçen süreler görülmektedir.

	MSHMM	AKY
Süre (sn.)	129.978	31.112

6. Vargılar

Bu çalışmada, Görsel-İşitsel Konuşma Tanıma sistemlerindeki veri kaynaştırma aşaması için yeni bir yaklaşım önerilmiş ve etkili bir veri kaynaştırma tekniği olarak kabul gören MSHMM ile karşılaştırma yapılmıştır. Önerilen yaklaşımın, MSHMM ile yakın performans göstermesi ve özellikle sınıflandırıcı aşamalarının iyileştirmelere açık olması ileriki çalışmalar için umut vericidir. Bunun yanında, test etme sürecinde AKY'in MSHMM'e göre daha az çalışma zamanına ihtiyaç duyduğu gösterilmiştir.

7. Kaynakça

[1] McGurk, H., MacDonald, J., "Hearing Lips and Seeing Voices", *Nature,* vol. 264, 746-748, 1976.

[2] ETSI, ``Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms", *ETSI Standard Document ES 201 108*, Nisan 2000.

[3] Viola, P., and Jones, M., "Rapid object detection using a bossted cascade of simple features.", In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Aralık 2001.

[4] Potamianos, G., Graf, H. P., ve Cosatto, E., "An image transform approach for HMM based automatic lipreading", *Proc. International Conference on Image Processing,* Chicago, IL, vol. I, s. 173-177.

[5] Dupont, S., Luettin, J., "Using the Multi-Stream Approach for Continuous Audio-Visual Speech Recognition: Experiments on the M2VTS Database", *International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[6] Hermansky, H., Ellis, D. P. W., Sharma, S., "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *Proceedings of ICASSP 2000*, vol. 3, 2000.

[7] Gurban, M., Thiran, J. P., "Audio Visual Speech Recognition with a Hybrid SVM-HMM System", *Proc. of Europian Signal Processing Conference,* 2005.

[8] The M2VTS Database, "http://www.tele.ucl.ac.be/PROJECTS/M2VTS/".