Yüz İfadesi Canlandırma için Konuşma Sinyalinden Otomatik Duygu Tanıma

Automatic Emotion Recognition for Facial Expression Animation from Speech

Elif Bozkurt[1], Engin Erzin[1], Çiğdem Eroğlu Erdem[2], A. Tanju Erdem[2]

[1]Elektrik ve Bilgisayar Mühendisliği Bölümü
Koç Üniversitesi, İstanbul
{ebozkurt,eerzin}@ku.edu.tr

[2]Momentum A.Ş.
TÜBİTAK-MAM-TEKSEB, A-205, Gebze, Kocaeli
{cigdem.erdem, terdem}@momentum-dmt.com.tr

Özetçe

Üç boyutlu konuşan kafaların yüz ifadesi canlandırmasını sadece konuşma bilgisini kullanarak otomatik olarak üretmek için bir sistem öneriyoruz. Sistemimiz Almanca dilinde ve yedi duygu içeren Berlin konuşma veritabanı üzerinde eğitildi. İlk olarak konuşma sinyalini bürün (prosody) ve spektral öznitelikler olarak özütlüyoruz. Daha sonra duygu tanıma için iki farklı sınıflandırıcı yapısını inceliyoruz: Gauss bileşen modeli (GBM) ve saklı Markov modeli (SMM) temelli sınıflandırıcılar. Deneysel çalışmalarda, 5-katlı çapraz sağlama yöntemini kullanarak, Mel frekans kepstral katsayıları (MFKK) ve dinamik MFKK özniteliklerine dayanan GBM sınıflandırıcıları ile ortalama %83.42 tanıma oranını elde ediyoruz. Ayrıca MFKK ve doğru spektral frekans katsayılarını (DSF) kullanan iki GBM sınıflandırıcının karar kaynaşımı ortalama %85.30 tanıma oranı sağlıyor. Aynı zamanda bu sonucun bürün SMM sınıflandırıcısı ile ikinci kademe karar kaynaşımı ortalama tanıma oranını %86.45'e yükseltiyor. Otomatik yüz ifadesi canlandırması için önerdiğimiz konuşma bilgisinden duygu tanıma yöntemi, inandırıcı ve ümit verici sonuçlar vermektedir.

Abstract

We present a framework for automatically generating the facial expression animation of 3D talking heads using only the speech information. Our system is trained on the Berlin emotional speech dataset that is in German and includes seven emotions. We first parameterize the speech signal with prosody related features and spectral features. Then, we investigate two different classifier architectures for the emotion recognition: Gaussian mixture model (GMM) and hidden Markov model (HMM) based classifiers. In the experimental studies, we achieve an average emotion recognition rate of 83.42% using 5-fold stratified cross validation (SCV) method with a GMM classifier based on Mel frequency cepstral coefficients (MFCC) and dynamic MFCC features. Moreover, decision fusion of two GMM classifiers based on MFCC and line spectral frequency (LSF) features yields an average recognition rate of 85.30%. Also, a second-stage decision fusion of this result with a prosody-based HMM classifier further advances the average recognition rate up to 86.45%. Experimental results on automatic emotion recognition to drive facial expression animation synthesis are encouraging.

1. Giriş

Bu çalışmada duygu bilgisini taşıyan konuşma parametreleri ve yüz ifadeleri arasında bir ilinti modeli oluşturup, üç boyutlu (3B) yüz ifadesi sentezini konuşma bilgisinden otomatik olarak gerçekleştirmeyi amaçlıyoruz. Literatürdeki bazı çalışmalar, bu amaç için ses-görüntü eşleme modelleri sunmaktadır [1], [2]. Ancak, bu modeller, öncelikle konuşma örüntülerinin yüz hareketi eğrilerine eşlendiği video analizini gerektirmektedir. Bizim amacımız ise, sadece konuşma bilgisini kullanarak, 3B kafa modelinin yüz ifadelerini otomatik olarak canlandırmaktır.

Konuşma sinyalinden duygu tanıma, özellikle bilgisayar-insan etkileşiminde giderek daha fazla önem kazanmaktadır. Buna örnek olarak çağrı merkezi uygulamalarını [3] ve insanlara duygusal öğeler kullanarak karşılık veren oyuncakları [4] gösterebiliriz. Bizim önerdiğimiz yöntemin hedef kullanım alanları bilgisayar oyunları, e-eğitim, oyun tabanlı eğitim ve 3B avatar-yardımcı içeren uygulamalardır.

Konuşma bilgisinden duygu tanıma üzerinde oldukça çalışılmış olmasına rağmen, hala çözümlenmemiş bir problemdir. Araştırmacılar çoğunlukla konuşmada duygu bilgisini taşıyan evrensel öznitelikler ve bunları etkili biçimde modelleyecek sınıflandırıcılar üzerine yoğunlaşmaktadırlar. Ancak, kültürel ve cinsiyet kaynaklı farklılıklar, duygu yüklü konuşmanın algılanmasında etkili olmaktadır. Üstelik duygu yüklü konuşma veri tabanları az sayıda; çoğunlukla tek dilde ve kayıt sayısı ya da duygu sayısı bakımından da yetersizdir. Örneğin, Almanca dilindeki Berlin duygu yüklü konuşma veritabanı [5] konuşmacıların rol yapmasıyla oluşturulmuş herkesin kullanımına açık bir veri tabanıdır.

2. Önerilen Sisteme Genel Bakış

Konuşma sinyalindeki duyguyu otomatik olarak tanıyan ve yüz ifadesi canlandırmasını sentezleyen sistemimizin şeması Şekil 1'de gösterilmektedir. Sistemimiz, iki bölümden oluşmaktadır: *sınıflandırıcı eğitimi* ve *yüz ifadesi canlandırması sentezi*. Duygu yüklü konuşma sinyali sistemimizin tek girdisidir. Tüm sistem, yedi duygu içeren Berlin duygu yüklü konuşma veritabanı üzerinde eğitilmiş ve test edilmiştir. Bu duygular mutluluk, öfke, korku, üzüntü, sıkıntı, iğrenme ve nötr duygularıdır.

Şekil 1'in üst yarısında gösterilen sistemimizin sınıflandırıcı eğitimi aşamasında, öncelikle duygu yüklü konuşma veritabanından kısa-süreli akustik öznitelikler özümlenmektedir. Bu öznitelikler, *Mel frekans kepstral katsayıları* (MFKK) ve *doğru spektral frekans katsayıları* (DSF) gibi spektral öznitelikleri ve bunların dinamik özniteliklerini (birinci ve ikinci türevleri); ayrıca, perde (pitch), perdenin birinci türevi ve enerjiden oluşan bürün özniteliklerini içermektedir. Daha sonra, Gauss bileşen modeli (GBM) sınıflandırıcılarını, spektral özniteliklerin olasılık yoğunluk dağılımını modellemek için kullanıyoruz. Saklı Markov modeli (SMM) sınıflandırıcılarını da zamana bağlı duygu bürün örüntülerini modellemek için kullanıyoruz.

Yüz ifadesi canlandırma sentezi Şekil 1'in alt yarısında betimleniyor. İlk önce, sınıflandırıcı eğitimi kaşamasında eğitilen GBM ve SMM sınıflandırıcılarını kullanarak konuşma sinyalinden duygu bilgisini kestiriyoruz. Daha sonra, tanınan duyguya karşılık gelen ve bir grafik sanatçısının 3B kafa modelinde daha önceden tanımlamış olduğu yüz ifadesini canlandırıyoruz. Konuşma içeriğindeki değişen duyguları ise, doğrusal ara değerlendirme yöntemini kullanarak zamanda örnekliyoruz.

Makalenin içeriği şu şekilde düzenlenmiştir: Bölüm 3'te konuşma sinyalinden duygu tanıma için özniteliklerin özütlenmesini, sınıflandırıcıların eğitimini ve kaynaşımını detaylandırıyoruz. Bölüm 4'te duygu tanıma sonuçlarını, yüz ifadesinin otomatik canlandırılması ve sentezi yöntemlerini sunuyoruz. Son olarak, Bölüm 5'te ise sonuçları değerlendiriyor ve tartışıyoruz.

3. Konuşma Sinyalinden Duygu Tanıma

Bu çalışmada, GBM ve SMM sınıflandırıcıları için bürün ve spektral öznitelikleri irdeliyoruz. Ayrıca, daha iyi duygu tanıma sonuçları elde etmek üzere, farklı sınıflandırıcıların kaynaşımını da inceliyoruz.

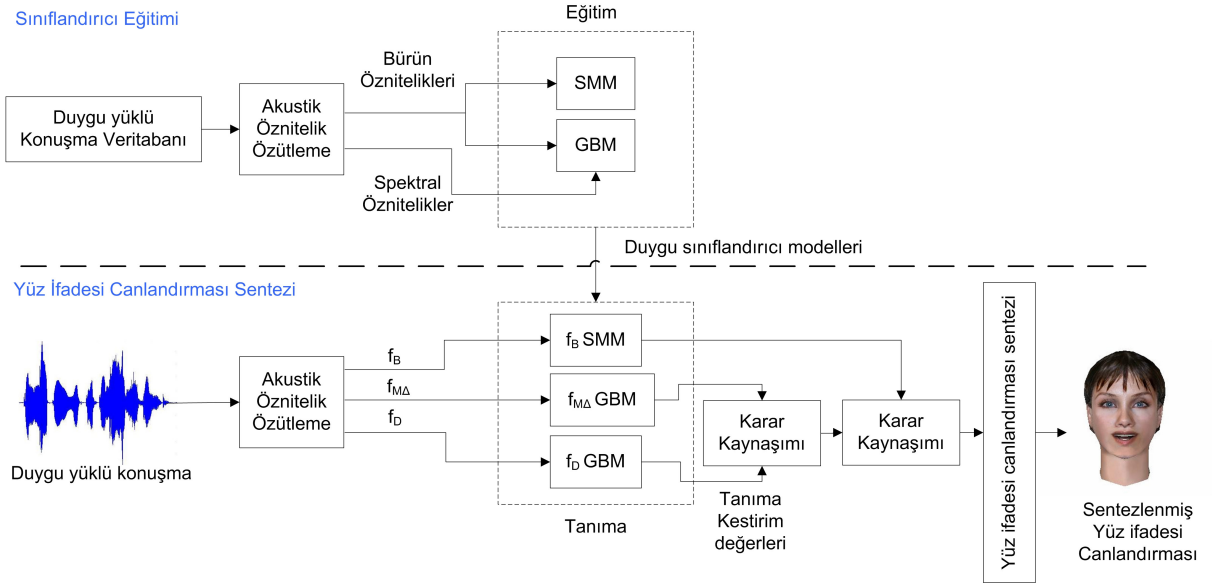
3.1. Öznitelik Özütlemesi

Bir konuşmanın duygusal içeriğini belirleyen iki tür kaynak vardır: sese ve dile ait içerik bilgileri. Bu çalışmada, çok dilli duygu tanıma amacımıza ters düştüğünden ikinci içeriği kullanmaktan kaçındık. Konuşmanın sese ait özniteliklerini tarif etmek için, hem spektral hem de bürün özniteliklerini kullandık.

Konuşmanın farklı duygular için, farklı bürün örüntüleri taşıdığını biliyoruz [6]. Bundan ötürü perde,perdenin birinci türevi ve enerji gibi bürün öznitelikleri, farklı duyguları modellemek için kullanılabilir. Örneğin, yüksek perde değerleri mutluluk, öfke ve korku ile ilintiliyken, düşük perde değerleri üzüntü, sıkıntı gibi duygularla ilintilidir [6].

Konuşmanın perde özniteliklerini öz ilinti yöntemi ile hesaplıyoruz [7]. Perde değerleri cinsiyete göre farklılık gösterdiğinden, konuşmacılar arası eşitliği sağlamak için, sıfırdan farklı perde değerlerinin ortalamasını, her perde değerinden çıkarıyoruz. Kullandığımız MFKK spektral özniteliklerinin, farklı duygularda değişen konuşma tayflarını modellemesini bekliyoruz. Ayrıca konuşma sinyalindeki zamana bağlı değişimleri takip edebilmek için, dinamik öznitelikleri (birinci ve ikinci türevleri) de kullanıyoruz.

Kullandığımız bir diğer spektral öznitelik ise, Itakura [8] tarafından önerilen DSF katsayıları. DSF öznitelikleri, biçimlendirici (formant) frekansları ile yakından ilintili olduğundan, konuşma tayfındaki bürün bilgisini modellemek için uygundurlar.



Şekil 1: Otomatik yüz ifadesi canlandırma sentezi için konuşmadan duygu tanıma sistemi

Ses tayfının zamanla değişimi, insanın konuşmayı algılamasında önemli rol oynar. Bu bilgiyi ifade etmenin bir yolu, kısa süreli tayf değişimini ölçen dinamik öznitelikleri kullanmaktır. Her *i* analiz penceresi için dinamik öznitelikleri aşağıdaki bağlanım (regression) formülü ile hesaplayabiliriz:

$$\Delta f(i) = \frac{\sum_{k=1}^{K} [f(i+k) - f(i-k)]k}{2\sum_{k=1}^{K} k^2}. \quad (1)$$

Analiz penceresi sayısını ise, $2K + 1 = 5$ olarak belirledik.

Kullanılan tüm akustik öznitelikleri Tablo 1'de özetliyoruz. Toplam 12 kepstral katsayı ve enerji teriminden oluşan 13 boyutlu MFKK vektörlerini f_M ile ifade ettik. Bu vektörleri birinci ve ikinci türevleri de içerecek şekilde genişleterek $f_{M\Delta}$ özniteliklerini hesapladık. 16 dereceli DSF özniteliklerini f_D, dinamik öznitelikleri ile $f_{D\Delta}$ olarak simgeledik. Bürün özniteliklerini f_B ile, bunların MFKK öznitelikleri ile kaynaşımını f_{BM} ile temsil ettik. Dinamik öznitelikleri içeren $f_{BM\Delta}$ özniteliğini de kullandık.

Akustik Öznitelikler	
Sembol	**Açıklama**
f_M	Mel frekans kepstral katsayıları
$f_{M\Delta}$	f_M ve dinamik öznitelikleri
f_D	Doğru spektral frekans katsayıları
$f_{D\Delta}$	f_D ve dinamik öznitelikleri
f_B	Bürün (perde, perdenin türevi, enerji)
f_{BM}	f_B ve f_M öznitelikleri kaynaşımı
$f_{BM\Delta}$	f_{BM} ve dinamik öznitelikleri

Tablo 1: Kullanılan öznitelikerin betimlemeleri.

3.2. GBM Sınıflandıcısı

Öznitelik olasılık yoğunlukları her duygu için, köşegen kovaryans matrisi kullanan GBM ile modellendi. GBM ile tanımlanan olasılık yoğunluk fonksiyonu

$$p(f) = \sum_{k=1}^{K} w_k p(f|k) \quad (2)$$

ile ifade edilebilir ve *K* bileşenin ağırlıklı birleşimidir. Formülde *f* gözlem öznitelik vektörüne, w_k *k*'ıncı Gauss bileşeninin birleşimdeki ağırlığına karşılık gelmektedir. Ağırlıklar,

$$0 \le w_k \le 1 \text{ ve } \sum_{k=1}^{K} w_k = 1 \quad (3)$$

koşullarını sağlamalıdır. Koşullu olasılık *p(f|k)* bileşen ortalama vektörü μ_k olan ve köşegen kovaryans matrisi Σ_k olan Gauss dağılımı ile modellendi.

Her duygunun GBM modeli, duyguyu betimleyen eğitim öznitelik vektörleri kullanılarak ve beklenen değerin en büyütmesi yöntemiyle kestirildi. Duygu tanıma aşamasında ise, verilen konuşmanın özniteliklerinin ardıl olasılıkları tüm GBM yoğunlukları üzerinden en büyütüldü.

3.3. SMM Sınıflandıcısı

Duygu bürün çizgelerinin zamanla değişimini SMM sınıflandıcısı ile modelledik. SMM sınıflandırıcısını paralel iki-dallı ve her dal sol-sağ yönlü *N* durum içerecek şekilde yapılandırdık. Her daldaki durum sayısını ve her durumdaki Gauss bileşeni sayısını yaptığımız deneyler sonucu belirledik. Dallardan birinin duyguya bağımlı öznitelikleri, diğerinin duygudan bağımsız öznitelikleri modellediğini varsaydık. SMM için perde, perdenin birinci türevi ve enerjiden oluşan bürün özniteliklerini kullandık. Her daldaki durum sayısını ve her durumun Gauss bileşeni sayısını, deney sonuçlarında en yüksek tanıma oranını verecek şekilde saptıyoruz.

3.4. Karar Kaynaşımı

Karar kaynaşımı için ağırlıklı toplama metodunu kullandık [9]. GBM ve SMM sınıflandırıcılarının her duygu ve konuşma için ürettiği tanıma kestirim değerlerini, karar kaynaşımı öncesi [0,1] aralığına indirgiyoruz [9]. Her *e* duygusu için *log* olabilirlik değerlerini sırasıyla $\overline{\rho}_{\gamma_e}$ ve $\overline{\rho}_{\lambda_e}$ ile ifade edelim. İki sınıflandırıcının birbirinden bağımsız olduğunu varsayarak, ağırlıklı toplamayı,

$$\rho_e = \alpha\overline{\rho}_{\gamma_e} + (1-\alpha)\overline{\rho}_{\lambda_e} \quad (4)$$

olarak tanımlıyoruz. Bu toplamda α katsayısını GBM sınıflandırıcısının ağırlığı olarak ve [0,1] aralığında seçtik.

4. Deneysel Sonuçlar

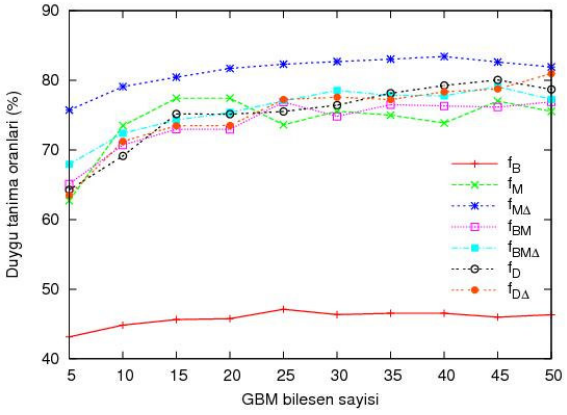
Sistemimizi Almanca dilindeki Berlin duygu yüklü konuşma veritabanında test ettik. Bu veritabanında 5 kadın ve 5 erkek konuşmacıya ait toplam 535 duygu yüklü konuşma kaydı bulunuyor.

DSF özniteliklerini 30 ms'lik pencerelerde ortalanmış 20 ms kaydırmalı çerçeveler üzerinden, diğer öznitelikleri ise 25 ms'lik pencerelerde ortalanmış 10 ms kaydırmalı çerçeveler üzerinden hesapladık.

Tanıma oranlarını 5-katmanlı çapraz sağlama yöntemini kullanarak elde ettik. Veritabanını beş bölüme ayırıp sırasıyla dördünü GBM ve SMM sınıflandırıcılarını eğitmek, kalan birini ise test etmek için kullandık. Bu şekilde, her ifade için bir tanıma sonucu elde etmiş olduk. Beş parçanın tanıma oranlarının ortalamasını nihai tanıma oranı olarak kullandık.

4.1. GBM Sınıflandıcısı Tanıma Sonuçları

Tablo 1'de belirtilen tüm öznitelikleri GBM sınıflandırıcısı ile modelledik. GBM bileşen sayısına göre değişen tanıma oranlarının grafiğinde de görüldüğü gibi (Şekil 2) $f_{M\Delta}$ öznitelikleri % 83.42 ile en yüksek tanıma oranına sahiptir.



Şekil 2: GBM bileşen sayısına göre duygu tanıma oranının değişimi

4.2. SMM Sınıflandıcısı Tanıma Sonuçları

Her duygunun bürün özniteliklerinin zamanda değişimini modellemek için iki-dallı SMM yapısını kullandık. Bu yapının duyguya bağlı ve duygudan bağımsız bürün özniteliklerini beklenen değerin en büyütülmesi yöntemi ile güdümsüz sınıflandırdık. SMM yapısını her dalda 5–30 arası durum ve her durumda en fazla 50 Gauss bileşeni olacak şekilde test ettik. Her dalda 5 durum içeren 48 Gauss bileşenli SMM yapısı, en yüksek tanıma oranını % 50.45 olarak verdi.

4.3. Karar Kaynaşımı Sonuçları

Karar kaynaşımı sonuçlarını denklem 4'te anlatıldığı gibi elde ettik. Test edilen GBM ve SMM sınıflandırıcıları arasında karar kaynaşımlarından en yüksek tanıma oranını verenler, 40 Gauss bileşenli $f_{M\Delta}$ GBM ile 5 durumlu, her durumda 48 Gauss bileşenli f_B SMM modeller oldu. Karar kaynaşım katsayısı da (α) 0.65 seçildiğinde duygu tanıma oranı % 84.73 oldu (Tablo 2).

Diğer taraftan, $f_{M\Delta}$ ve f_D GBM sınıflandırıcılarının karar kaynaşımı sonucunda ise %85.30 tanıma oranı elde ettik. Bu karar kaynaşımına f_B SMM ile ikinci kez karar kaynaşımını uyguladığımızda ise duygu tanıma %86.45'ya yükseldi. Tablo 2, tek aşamalı ve iki aşamalı karar kaynaşım tanıma sonuçlarını özetlemektedir.

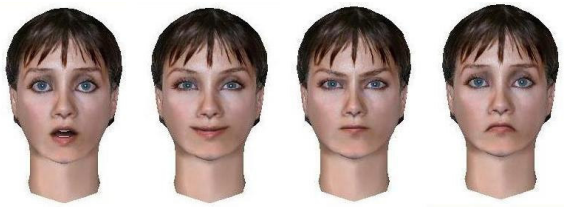
Karar Kaynaşımı	Tanıma Oranı (%)	α
$\gamma(f_{M\Delta}) \oplus \lambda(f_B)$	84.73	0.65
$\gamma(f_{M\Delta}) \oplus \gamma(f_D)$	85.30	0.57
$(\gamma(f_{M\Delta}) \oplus \gamma(f_D)) \oplus \lambda(f_B)$	86.45	0.57, 0.13

Tablo 2: Karar kaynaşımı sonrası duygu tanıma oranları ve sınıflandırıcı ağırlık değerleri. Burada, γ ve λ sembolleri sırasıyla GBM ve SMM sınıflandırıcılarına karşılık geliyor.

4.4. Yüz İfadesinin Canlandırılması

Gerçek hayatta konuşma esnasında duygu içeriği değişebilir. Konuşma sinyalinden yüz ifadesi otomatik olarak canlandırılırken, duygu değişikliklerini yakalayacak kadar kısa, fakat yüksek tanıma oranı verecek kadar uzun karar pencereleri kullanılmalıdır. Karar penceresinin uzunluğunu 100ms ile 3s arasında değiştirerek, tanıma sonuçlarını gözlemledik. İki saniyeden uzun pencereler için tanıma oranı fazla değişmediğinden, karar penceresi boyutunu 2s olarak belirledik. Ancak, her duygu tam olarak 2s sürmeyebileceğinden, karar penceresini konuşma boyunca 40ms kaydırıp, elde ettiğimiz toplam 50 pencerenin tanıma sonuçlarını ortanca süzgecinden geçirip, duygu süresini hesapladık. Konuşma bilgisinden elde edilen duygu tanıma sonuçlarına karşılık gelen yüz ifadelerini, üç boyutlu yüz modelleri üzerinde [10] canlandırdık. Canladırma sırasında, 100 ms geçiş zamanı ve ardışık olarak doğrusal aradeğerleme yöntemini kullandık.

Korku, mutluluk, öfke ve üzüntü duygularına ait sentezlenen yüz ifadeleri Şekil 3'te gösterilmektedir. Ayrıca, yüz ifadesi canlandırmasına ait bazı video örneklerine http://www.momentum-dmt.com/data/FaceAnim adresinden ulaşılabilir.



Şekil 3: Canlandırılan yüz ifadelerine örnekler: soldan sağa korku, mutluluk, öfke ve üzüntü

5. Sonuç

Bu çalışmada, konuşma sinyalinden duygu tanıma ve otomatik yüz ifadesi canlandırması için bir sistem önerdik. Konuşma sinyalindeki duygu bilgisini, spektral ve bürün özniteliklerini kullanarak, GBM ve SMM sınıflandırıcıları ile modelledik. Karar kaynaşımı metodunu kullanarak, duygu tanıma oranını %86.45'e kadar yükselttik. Duygulara karşılık gelen yüz ifadelerini doğrusal aradeğerleyerek canlandırdık. Deneysel sonuçlar oldukça inandırıcı ve ümit vericidir.

6. Teşekkür

Bu çalışma TÜBİTAK-TEYDEB 3070796, TÜBİTAK 106E201 numaralı projeler kapsamında ve COST2102 tarafından desteklenmiştir.

7. Kaynakça

[1] M. Brand, "Voice puppetry," in *Proceedings of SIGGRAPH*, 1999, pp. 21–28.

[2] J. Cassell, T. Bickmore, L. Campbell, K. Chang, H. Vilhjlmsson, and H. Yan, "Requirements for an architecture for embodied conversational characters," in *Proceedings of Computer Animation and Simulation*, 1999, pp. 109–120.

[3] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans.Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[4] P. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, pp. 157–183, Jul. 2003.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of Interspeech*, Lisbon, Portugal, 2005, pp. 1517–1520.

[6] K. R. Scherer, "How emotion is expressed in speech and singing," in *Proceedings of XIIIth International Congress of Phonetic Sciences*.

[7] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*. New York: Macmillan Publishing Company, 1993.

[8] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of the Acoustical Society of America*, vol. 57, no. Suppl. 1, p. S35, 1975.

[9] E. Erzin, Y. Yemez, and A. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality realiability," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 840–852, Oct. 2005.

[10] A. Erdem, "A new method for generating 3d face models for personalized user interaction," in *Proceedings of 13th European Signal Processing Conference*, Antalya, Turkey, Sept. 4-8 2005.