Paralel Model Kombinasyonu ve Yerel Öznitelikler Kullanarak Gürbüz Konuşmacı Onaylama
Noise Robust Speaker Verification Using Paralel Model Combination and Local Features

Zekeriya Tüfekci

Elektrik-Elektronik Mühendisliği Bölümü
İzmir Yüksek Teknoloji Enstitüsü,Gülbahçe Köyü, 35430, Urla, İzmir
zekeriyatufekci@iyte.edu.tr

Özetçe

Gürültü, konuşmacı onaylama sisteminin performansını önemli ölçüde düşürür. Paralel model kombinasyonu (PMC) tekniği gürültülü ortamlarda konuşma tanıma için en iyi yöntemlerden biridir. Diğer bir yöntem ise konuşma tanıma için yerel öznitelikler kullanmaktır. Önceki çalışmalarımızda, gürültülü ortamlarda konuşma tanıma için bir yerel (zaman/frekans bölgesinde) öznitelik olan mel frekansı ayrık dalgacık katsayıları (MFDWC) [1] öznitelik vektörü önerildi. Bu bildiride gürültünün konuşma onaylama sistemine etkisini azaltmak için PMC ve yerel özniteliklerin (MFDWC) birlikte kullanımı sunuldu. MFCC ve MFDWC'in performansı değişik gürültü türleri ve sinyal-gürültü oranları için karşılaştırıldı. Deneysel sonuçlar MFDWC'nin MFCC'ye göre önemli oranda hata oranını düşürdüğünü gösterdi. Örnek olarak 12 dB sinyal gürültü oranı için MFDWC MFCC'ye göre hata oranını %38.33 kadar düşürmüştür.

Bildiri konusu: S9 image/Video/Ses Tarama, Bulma

Abstract

Interfering noise severely degrades the performance of a speaker verification system. The Parallel Model Combination (PMC) technique is one of the most efficient techniques for dealing with such noise. Another method is to use features local in the frequency domain. Recently, we proposed Mel-Frequency Discrete Wavelet Coefficients (MFDWCs) [1] as speech features local in frequency domain. In this paper, we discuss using PMC along with MFDWC features to take advantage of both noise compensation and local features (MFDWCs) to decrease the effect of noise on verification performance. We evaluate the performance of MFDWCs for various noise types and noise levels. We also compare the performance of these versus MFCCs and both using PMC for dealing with additive noise. The experimental results show significant performance improvements for MFDWCs versus MFCCs after compensating the HMMs using the PMC technique. For example the MFDWCs gave 6.29 points performance improvement on average over MFCCs for 12 dB. This corresponds to 38.33% error reduction.

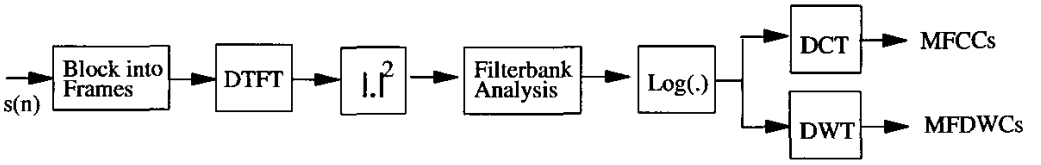
1. Giriş

Konuşmacı tanıma ve konuşmacı onaylama güncel araştırma konularındandır. Konuşmacı tanımanın amacı verilen ses örneğinin bir grup içindeki kişilerden hangisine ait olduğunu bulmaktır. Konuşmacı onaylamada ise amaç, verilen ses örneğinin iddia edilen kişiye ait olup olmadığına karar vermektir. Konuşma tanıma ve onaylama sistemi verilen konuşma sinyalinin metnine bağımlı veya metinden bağımsız olabilir. Metine bağlı konuşmacı onaylama sisteminde konuşmacının önceden belirlenen kelime veya cümleyi söylemesi gerekir. Metinden bağımsız sistemde ise konuşmacı istediği kelime veya cümleyi söyleyebilir. Konuşma tanıma ve onaylama için Gauss karışım modeli (GMM) [2] günümüzde en yaygın olarak kullanılan modellerden biridir. Bu çalışmada metinden bağımsız konuşmacı onaylama sistemindeki modeller için GMM'ler kullanılmıştır.

Gerçek uygulama ortamlarında konuşmacı onaylama sisteminin gürültüye karşı gürbüz olması istenir. Eğitim ve test koşullarının farklı olması durumunda konuşma onaylama sistemlerinin performansı büyük oranda düşer. En iyi performans, test ve eğitim koşullarının aynı olması durumunda elde edilir. Literatürde gürültülü ortamlarda performansın düşmesini azaltmak için birçok yöntem [3] önerilmiştir. Bunların en etkililerinden biri PMC [4] yöntemidir. Bu yöntemde gürültü modeli ve gürültüsüz ortamdaki konuşmacı modeli kullanılarak gürültülü ortam için konuşmacı modeli kestirilmeye çalışılır.

Gürültüye karşı gürbüz diğer bir yöntem de frekans bölgesinde yerel bilgi taşıyan çokbandlı [5–7] ve çokçözünürlüklü [1, 8] öznitelik vektörleri kullanmaktır. Bu bildiride eğer öznitelik vektörünün bazı elemanları frekans bölgesinde yerel bilgi taşıyorsa bu öznitelik vektörü yerel öznitelik vektörü olarak isimlendirilecektir. Yerel öznitelik vektörlerinin MFCC, LPCC gibi yerel olmayan öznitelik vektörlerine göre birçok avantajı [1, 5, 6, 8, 9] vardır. Bunlardan en önemlisi gürültünün öznitelik vektörünü etkilemesidir. Bir frekans bölgesideki gürültü, yerel öznitelik vektöründeki sadece birkaç elemanı (o frekans bandıyla ilgili elemanları) etkiler. Yerel olmayan öznitelik vektöründe ise, bir frekans bandındaki gürültü öznitelik vektörünün tüm elemanlarını etkiler, çünkü yerel olmayan öznitelik vektörünü hesaplarken frekans bandlarının hepsi kullanılmaktadır.

0-7803-8318-4104/$20.00©2004 IEEE



Şekil 1: MFCC ve MFDWC'nin elde edilmesi

Bu çalışmada PMC tekniği MFDWC'ye uygulanarak gürültü denkleştirme tekniği ve yerel öznitelik vektörünün gürültülü ortamın konuşmacı onaylama performansına etkisi azaltılmaya çalışıldı.

2. Dalgacık Dönüşümü ve MFDWC

MFDWC, çerçevelenmiş konuşma sinyalinin mel ölçekli log süzgeç bankası enerjilerininin ayrık dalgacık dönüşümünün (DWT) alınmasıyla elde edilir. Şekil 1 MFCC ve MFDWC'nin nasıl elde edildiğini göstermektedir.

WT, verilen bir sinyalin yüksek frekanslı bileşenini ölçmek için kısa taban fonksiyonları, düşük frekanslı bileşenini ölçmek için uzun taban fonksiyonlarını kullanır. Bu ise WT'yi kısa zaman Fourier dönüşümü (STFT) ve Fourier dönüşümünden (FT) farklı yapan en önemli özelliğidir. Dalgacık $\Psi(t) \in L^2(\mathbb{R})$ (karesinin integrali alınabilen fonksiyonlar uzayı) ortalaması sıfır ve normu bir olan bir fonksiyondur. Diğer bir anlatımla dalgacık aşağıdaki şartları saglayan bir fonksiyondur.

$$\int_{-\infty}^{+\infty} \Psi(t)dt = 0 \tag{1}$$

ve

$$\|\Psi(t)\| = 1. \tag{2}$$

Wavelet dönüşümünün analiz fonksiyonu ölcek s ve kaydırma u'da aşağıdaki gibi verilir.

$$\Psi_{u,s}(t) = \frac{1}{\sqrt{s}}\Psi(\frac{t-u}{s}). \tag{3}$$

$f(t) \in L^2(\mathbf{R})$ fonksiyonunun zaman u ve ölçek s'deki dalgacık dönüşümü aşağıdaki gibidir.

$$F(u,s) = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\Psi^*(\frac{t-u}{s})dt = \int_{-\infty}^{+\infty} f(t)\Psi^*_{u,s}(t)dt. \tag{4}$$

Burada * karmaşık eşleniği göstermektedir. Kuramsal olarak ortalaması sıfır ve sonlu enerjiye sahip olan herhangi bir fonksiyon dalgacık olabilir. Dalgacık seçmek için birçok kriter bulunmaktadır. Sonsuz süreli dalgacıkları pratikte kullanmak olanaklı değildir. Zaman ve frekans bölgesinde iyi çözünürlüğe sahip olmak için dalgacığın frekans ve zaman bölgesinde hızla azalması istenir. Bu çalışmada kullanılan sinyal sonlu olduğundan sinyal sınırlarındaki süreksizlikten dolayı dalgacık katsayılarında istenmeyen büyük değişimler olacaktır. Bunu azaltmak için katlanmış dalgacık (dalgacığın simetrik veya antisimetrik olması gerekir) veya sınır dalgacığı kullanılabilir. Bunlara ek olarak bizim sinyalimiz ayrık sinyal olduğundan dolayı ayrık dalgacık transformu (DTW) kullanmamız gerekir. Bütün bu koşulları düşündüğümüzde kullanılabilecek dalgacık sayısı sınırlı olmaktadır. Bu çalışmada kullanılan dalgacık ve dalgacığın Fourier dönüşümü Şekil 2'de görülmektedir. Dalgacık fonksiyonunun zaman ve frekans bölgesinde belli bir alanda yoğunlaşması konuşma onaylama sisteminin gürültüye karşı gürbüz olması açısından önemlidir. Dalgacık fonksiyonu ne kadar zaman ve frekans bölgesinde yoğunlaşmışsa ilgili dalgacık katsayısının yerel gürültüden etkilenme olasılığı o kadar az olacaktır. Dalgacık dönüşümü (WT) hakkında daha fazla bilgi için şu kaynaklara [10, 11] başvurulabilir.

MFCC ve MFDWC'nin nasıl elde edildiği Şekil 1'de gösterilmektedir. İlk beş adım her iki yöntem için de aynıdır. Sadece son adım farklıdır. Son adımda MFCC'yi hesaplamak için ayrık kosinüs dönüşümü kullanılmakta, MFDWC'yi hesaplamak içinse ayrık dalgacık dönüşümü (DWT) kullanılmaktadır. İlk adımda konuşma sinyali Hamming ve Hanning gibi yumuşak örtüşen pencereler kullanarak öbeklere bölünmektedir. Bir sonraki adımda pencerelenen sinyalin DTFT'si alınmaktadır. Daha sonraki adımda karesi alınıp süzgeç bankalarından geçirilmektedir. Beşinci adımda ise elde edilen sinyalin logaritması alınmaktadır.

3. Paralel Model Kombinasyon Tekniğinin MFCC ve MFDWC'ye Uygulanması

Eğitim ve test koşulları aynı olduğunda konuşma onaylama sistemi en iyi sonucu verecektir. Test koşulu değiştiğinde iyi sonuç almanın en basit yolu sistemi yeni koşul için tekrar eğitmektir. Fakat bu pratik bir çözüm değildir çünkü yeni ortam için tüm eğitim veritabanına ihtiyacımız vardır. Yeni test ortamına ait eğitim veritabanımız olsa dahi sistemi yeni koşullar için eğitmek çok zaman alacaktır.

Yeni test ortamında iyi sonuç almak için Gales ve Young [4, 12–14] paralel model kombinasyonu (PMC) yöntemini önerdi. PMC tekniği, gürültü modeli ve konuşmacı modeli verildiğinde gürültülü ortamdaki konuşmacı modelini kestirmektedir. Bu yöntem gürültülü ortam için modeli tekrar eğitmege göre daha etkili ve az zaman harcayan bir yöntemdir.

Gürültülü konuşmacı modelini kestirmek için üç değişik PMC tekniği vardır: sayısal tümlevleme [4], veri-sürümlü yöntem [15] ve log-normal yaklaşımı [12]. Bu çalışmada, en az işlem gerektirdiği için log-normal yaklaşımı kullanılmıştır.

İyi performans almak için fark öznitelikleri de (delta coefficients) kullanıldı. Bundan dolayı PMC tekniği kullanarak konuşmacı modellerinin fark öznitelikleriyle ilgili parametreleri de kestirildi [16, 17]. PMC tekniğinin MFDWC ve MFCC'ye uygulanışı arasında çok az fark vardır. PMC MFCC'ye uygulanırken DCT ve ters DCT kullanılmaktadır. PMC MFDWC'ye uygulanırken ise DCT ve ters DCT yerine DWT ve ters DWT kullanılmaktadır.

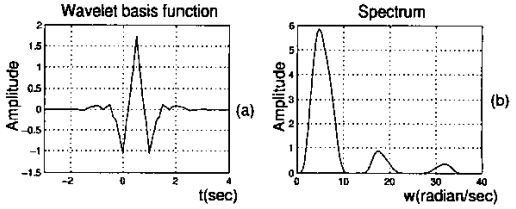
4. Deney Düzeneği ve Sonuçları

MFCC ve MFDWC'lerin konuşmacı tanıma performanslarını karşılaştırmak için NIST 1998 speaker evaluation [18] veritabanı kullanıldı. Bu veritabanı 250 erkek ve 250 kadın olmak üzere toplam 500 kişinin telefon hattı üzerinden kaydedilm iş olan ses kayıtlarını içermektedir. Bu çalışmada sadece erkek konuşmacılar kullanıldı. Modellerin eğitiminde her bir konuşmacı için aynı telefon ahizesini kullanarak iki farklı zamanda kaydedilmiş olan ve her biri ortalama bir dakika süren beş adet konuşma dosyası kullanıldı. Test için her birinin uzunluğu 30 msn olan ve aynı telefon ahizesi kullanılarak oluşturulan 1308 adet konuşma dosyası kullanılmıştır. Bu çalışmanın amacı gürültülü ortamda MFCC ve MFDWC'lerin performansını karşılaştırmak olduğundan dolayı veritabanındaki sadece aynı telefon ahizesiyle yapılan kayıtlar kullanılmıştır.

MFCC ve MFDWC öznitelik vektörleri Şekil 1'de görüldüğü gibi elde edildi. 8 kHz ile örneklenmiş olan ses sinyalleri her on milisaniyede bir otuziki msn'lik Hamming pencereleri ile çerçevelendi. Sinyalin güç spektrumunu hesaplamak için her bir çerçevenin FFT'si hesaplandı. Mel ölçeklenmiş log süzgeç bankası enerjilerini hesaplamak için 26 tane mel ölçekli bant geçiren süzgeç tasarlandı. Daha sonra aradeğerlendirmeyle mel ölçeklenmiş log süzgeç bankası enerji sayısı 26'dan 33'e çıkarıldı.

MFCC'ler mel ölçekli log süzgeç bankası enerjilerinin kosinüs dönüşümü alınarak hesaplandı. MFCC'lerin ilk 16'sı ve sıfırıncısı olmak üzere toplam 17 elemanlı MFCC öznitelik vektörü oluşturuldu. Önceki deney sonuçları [1, 9] simetrik dalgacıkların antisimetrik dalgacıkladan daha iyi sonuç verdiğini göstermiştir. Bundan dolayı bu çalışmada Şekil 2'de gösterilen simetrik dalgacık kullanıldı. Ölçek dörtten 8 öznitelik, ölçek sekizden 4 öznitelik, ölcek onaltıdan 2 öznitelik, ölçek otuzikiden 1 öznitelik ve sıfırıncı öznitelik olmak üzere MFDWC öznitelik vektörü toplam 17 elemandan oluşturuldu. Ayrıca tüm deneylerde fark öznitelikleri de kullanıldı. Sonuç olarak elde edilen öznitelik vektörünün boyutu 34 oldu. Her bir konuşmacı ve background için 64 Gauss dağılımlı köşegen (diagonal) kovaryanslı sürekli GMM kullanıldı. Modelleri eğitmek ve test etmek için HTK [19] yazılımı kullanıldı.

MFCC ve MFDWC'nin performansını gürültülü ortamlarda konuşmacı onaylama konusunda karşılaştırmak için deneyler gerçekleştirildi. Gürültü olarak NOISEX-92[20] veritabanındaki SPEECH, LYNX ve F16 gürültüleri kullanıldı. Bu gürültüler test sinyallerine -6, 0 , +6 ve +12 dB SNR oranlarında eklendi. Deney sonuçları Tablo 1'de görülmektedir. Tablo 1'de görüldüğü gibi konuşmacı onaylama için bu çalışmada kullanılan tüm gürültü türleri ve SNR değerleri için MFDWC MFCC'ye kıyasla daha iyi sonuç verdi. Tablo 1'de temiz



Şekil 2: Dalgacık fonksiyonunun zaman ve frekans bölgesinde yayılımı.

konuşma sinyali icin MFCC ve MFDWC'nin performanslarının birbirine yakın olduğu görülmektedir. Tablo 1'in altıncı satırı her bir SNR değeri icin MFCC ve MFDWC'nin ortalama eşit hata oranlarını (EER) göstermektedir. Görüldüğü gibi -6, 0, 6 ve 12 dB SNR değerleri için, MFDWC MFCC'ye göre eşit hata oranını (EER) sırasıyla %6.84, %20.16, %36.87 ve %38.33 oranında düşürmüştür.

5. Sonuç

Bu makalede bir model denkleştirme yöntemi olan PMC tekniğinin bir yerel öznitelik olan MFDWC'ye uygulanması ve MFDWC'nin konuşmacı tanıma amacıyla kullanılması incelendi. Deneysel sonuçlar bir yerel öznitelik olan MFDWC'in PMC ile kullanıldığında gürültülü ortamlarda MFCC'ye (MFCC'de PMC ile kullanıldığında) göre çok daha iyi sonuçlar verdiğini göstermiştir. MFDWC'nin MFCC'den daya iyi sonuç vermesinin bir nedeni MFDWC'lerin yerel özellik taşımasından dolayı olabilir.

Tablo 1: MFCC ve MFDWC'nin PMC ile birlikte kullanıldığında elde edilen eşit hata oranları (equal eror rates).

Gürültü Tipi	MFCC					**MFDWC**				
	-6 dB	0 dB	6 dB	12 dB	Temiz	**-6 dB**	**0 dB**	**6 dB**	**12 dB**	**Temiz**
Speech	38.46	30.51	22.25	17.43	6.50	**35.40**	**23.01**	**13.84**	**9.94**	**6.88**
Lynx	33.49	25.46	19.11	14.91	6.50	**30.58**	**18.20**	**11.54**	**9.71**	**6.88**
F16	40.83	31.35	23.09	16.90	6.50	**39.07**	**28.52**	**15.29**	**10.70**	**6.88**
Ortalama	37.59	29.11	21.48	16.41	6.50	**35.02**	**23.24**	**13.56**	**10.12**	**6.88**

6. Kaynakça

[1] J. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," in *Proceedings of ICASSP*, 2000.

[2] Douglas A. Reynolds and Richard R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[3] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, 1995.

[4] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, pp. 289–307, 1995.

[5] H. Bourlard and S. Dupont, "A new asr approach based on independent processing and recombination of partial frequency bands," in *Proceedings of ICSLP*, 1996.

[6] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards asr on partially corrupted speech," in *Proceedings of ICSLP*, 1996.

[7] Z. Tufekci and J. Gowdy, "Subband feature extraction using lapped orthogonal transform for speech recognition," in *Proceedings of ICASSP (accepted)*, 2001.

[8] S. Vaseghi, N. Harte, and B. Milner, "Multi resolution phonetic/segmental features and models for hmm based speech recognition," in *Proceedings of ICASSP*, 1997.

[9] Z. Tufekci and J. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," in *Proceedings of SoutheastCon*, 2000.

[10] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[11] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice Hall, 1995.

[12] M. J. F. Gales and S. J. Young, "An improved approach to the hidden markov model decomposition of speech and noise," in *Proceedings of ICASSP*, 1992.

[13] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for hmm recognition in noise," *Speech Communication*, vol. 12, pp. 231–240, 1993.

[14] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model compansation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 4, no. 5, 1996.

[15] M. J. F. Gales and S. J. Young, "A fast and flexible implementation of parallel model combination," in *Proceedings of ICASSP*, 1995, pp. 133–136.

[16] M. J. F. Gales and S. J. Young, "Hmm recognition in noise using parallel model combination," in *Proceedings of EUROSPEECH*, 1993, pp. 837–840.

[17] R. Yang and P. Haavisto, "An improved noise compensation algorithm for speech recognition in noise," in *Proceedings of ICASSP*, 1996, pp. 49–52.

[18] NIST, "The 1998 speaker recognition evaluation plan," *www.nist.gov/speech/tests/spk/1998/current_plan.htm*, 1998.

[19] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory Ltd., version 2.1, 1997.

[20] A. P. Varga, H. J. M. Steenekan, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," Tech. Rep., DRA Speech Research Unit, 1992.