SÜREKLİ SAKLI MARKOV MODELLERİ İLE METİNDEN BAĞIMSIZ KONUŞMACI TANIMA PARAMETRELERİNİN İNCELENMESİ

Cemal HANİLÇİ[*]
Figen ERTAŞ[*]

Özet: Bu çalışmada, ergodik ve soldan-sağa olmak üzere iki farklı saklı Markov modelleri kullanan metinden bağımsız konuşmacı tanıma sistemine ilişkin parametreler (özellik vektörü boyutu, model durum ve karışım sayıları) tanıma başarımına etkisi yönünden karşılaştırmalı olarak incelenmiş ve bunların optimum değerleri belirlenmiştir.

Anahtar Kelimeler: Konuşmacı tanıma, saklı Markov modelleri, özellik vektörü, mel frekansı kepstrum katsayıları.

On the Parameters of Text-Independent Speaker Identification Using Continuous HMMs

Abstract: In this paper, the parameters of text-independent speaker identification system (size of feature vector, number of states and mixtures) using Hidden Markov Models (HMMs) of both ergodic and left-to-right type have been analyzed in relation to identification rate, and their optimum values have been determined.

Key Words: Speaker identification, hidden Markov models, feature vector, mel frequency cepstrum coefficients.

1. GİRİŞ

Konuşmacı tanıma sistemlerinin kullanım alanları oldukça yaygındır. Örneğin son yıllarda, telefon bankacılığı, sesli arama, telefonla alışveriş, veritabanı erişim servisleri, bilgisayarların uzaktan sesle kontrolü ve en önemlilerinden biri de adli uygulamalar gibi birçok alanda kullanılmaya başlanmıştır (Furui, 1997).

Sürekli Saklı Markov Modelleri (SMM) önceleri konuşma tanıma için kullanılmaya başlanmış ve ilerleyen zamanlarda konuşmacı tanıma uygulamalarında da kullanılmıştır (Rosenberg ve diğ., 1991). Örnek olarak, Li ve diğ. (2002) 12 kepstrum katsayısı kullanarak 32 karışım ve 3 durumlu SMM için konuşma tanıma, konuşmacı tanıma ve doğrulama performanslarını, Mirghafori ve diğ. (2005) 8 karışım kullanan soldan-sağa SMM ile 20 *mfcc* kullanarak konuşmacı tanıma performanslarını incelemiştir. Ancak, SMM ile konuşmacı tanıyan sistemlerin kullandığı parametrelerin tanıma başarımına etkileri yönünden karşılaştırmalı bir analizi literatürde henüz yer almamıştır. Bu çalışmada, tanıma başarımına etki eden özellik vektörünün boyutu, durum sayısı ve karışım sayısı gibi parametreler TIMIT veritabanından rastgele seçilen 40 kişilik bir konuşmacı grubu için karşılaştırmalı olarak incelenmiş ve bunların optimum değerleri ergodik ve soldan-sağa SMM için belirlenmiştir.

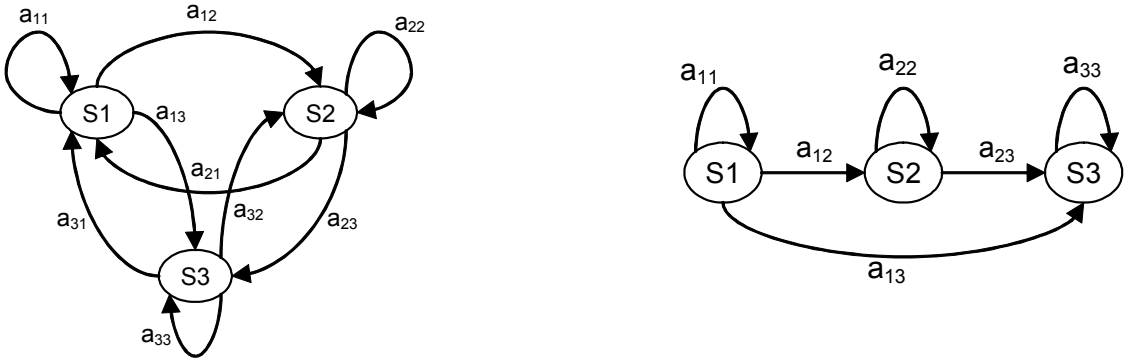
Esas olarak tanıma başarımının, özellik vektör boyutu, durum sayısı ve karışım sayısının bir fonksiyonu olduğu, ancak genelleme yapmak gerekirse, özellik vektör boyutu ve karışım sayısındaki artmanın tanıma başarımını artırdığı sonucuna varılmıştır.

2. SÜREKLİ SAKLI MARKOV MODELLERİ

Saklı markov modelleri (SMM), bir örüntüye ait çerçevelerin spektral özelliklerini modellemede yaygın olarak kullanılan yöntemlerden biridir (Rabiner ve Juang, 1993). SMM (veya diğer istatistiksel yöntemler) ile konuşmacı tanıma sistemlerinin temel dayanak noktası, ses sinyalinin rastsal süreç olarak iyi bir şekilde ifade edilebilmesidir. Bu nedenle SMM tabanlı konuşmacı tanıma yöntemleri diğer bazı yöntemlerden daha iyi performans göstermektedir (Zheng and Yuan,1988; Naik ve diğ., 1989).

[*] Uludağ Üniversitesi, Mühendislik-Mimarlık Fakültesi, Elektronik Mühendisliği Bölümü, Görükle, Bursa.

SMM bugüne kadar konuşmacı tanıma uygulamalarında yaygın olarak kullanılmıştır (Matsui ve Furui,1994; Yu ve diğ., 1995). SMM durum geçiş olasılık matrisine bağlı olarak ergodik ve soldan-sağa SMM olmak üzere iki gruba ayrılmaktadır (Rabiner ve Juang 1986). Şekil 1'de 3 durumlu ergodik ve soldan-sağa saklı markov modelleri görülmektedir.



Şekil 1:
3 durumlu (a) Ergodik Saklı Markov Model, (b) Soldan- Sağa Saklı Markov Model

Şekil 3'den de anlaşılacağı gibi, ergodik SMM'de bir durumdan diğer bütün durumlara geçiş varken soldan-sağa SMM için bu durum söz konusu değildir. Bu da durum geçiş olasılık matrisine yansımaktadır. Ergodik SMM ile soldan-sağa SMM arasındaki diğer bir fark da başlangıç durum olasılıklarında görülmektedir. Soldan-sağa SMM'de durum dizisi birinci durumdan başlamak ve *N*. durumda sonlanmak zorundadır (Rabiner ve Juang, 1986).

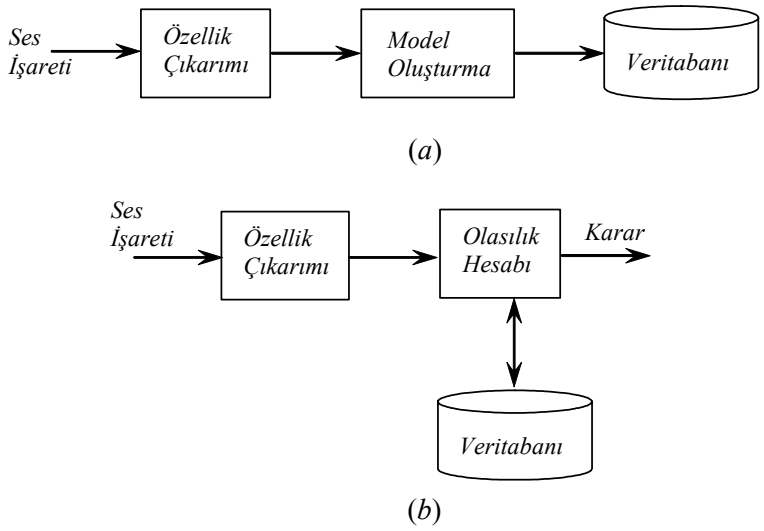
SMM ile konuşmacı tanıma sisteminde özellik vektörleri gözlem sembollerini temsil etmektedir. Gözlem sembol olasılıkları,

$$b_j(o) = \sum_{k=1}^{M} c_{jk} N(o, \mu_{jk}, U_{jk}), \qquad 1 \le j \le N$$

formülü ile hesaplanır. Denklemdeki *o* gözlem sembollerini (özellik vektörü elemanları), c_{jk} *j*. durumdaki *k*. karışım katsayısını, *N* ise ortalama vektörü μ_{jk} ve kovaryans matrisi U_{jk} olan Gauss olasılık yoğunluk fonksiyonunu belirtmektedir. *M* ise kullanılan karışım sayısıdır. Herhangi bir konuşmacıya ait SMM böylece $\lambda(A,b,\pi)$ şeklinde ifade edilmekte olup, *A* durum geçiş olasılık matrisini, *b* gözlem sembol olasılıklarını ve π ise başlangıç durum olasılıklarını temsil etmektedir.

3. SİSTEM TANIMI

SMM ile konuşmacı tanıma sistemi iki aşamadan oluşmaktadır. İlk aşama, her konuşmacıya ait modelin oluşturulduğu eğitim aşaması, ikincisi ise test aşamasıdır. Eğitim aşamasında her konuşmacının eğitim cümleleriyle SMM model parametreleri (A,b,π) hesaplanır. Model parametreleri hesaplanırken Baum-Welch (Forward-Backward) algoritması kullanılır. Test aşamasında ise, konuşmacıların test cümleleri sistemin girişine uygulanır ve veritabanındaki modeller kullanılarak Viterbi algoritması ile giriş cümlesinin olasılığını maksimum yapan durum dizisi bulunur. Bu olasılığı maksimum yapan model, test cümlesinin veritabanındaki kişilerden hangisine ait olduğunu belirler. Şekil 2 SMM ile konuşmacı tanıma sistemine ait eğitim ve test aşamalarını göstermektedir.



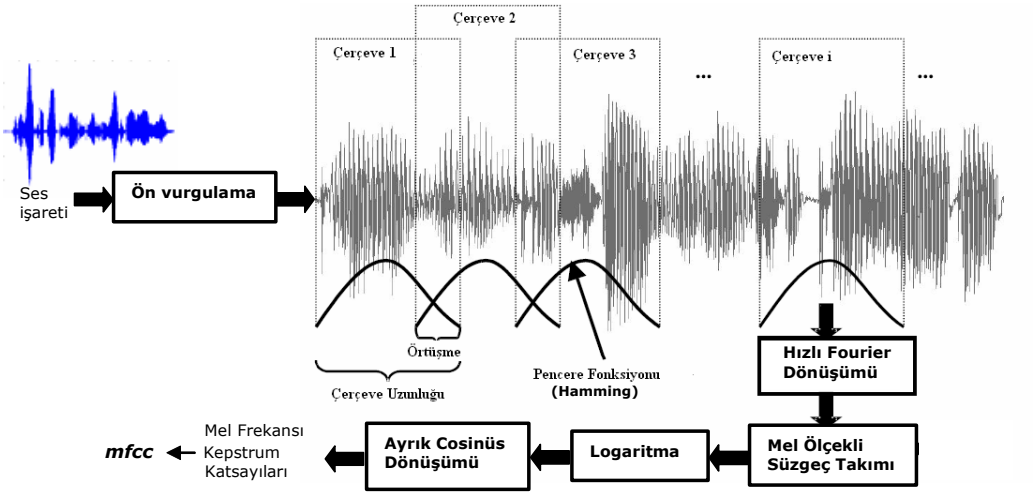
Şekil 2:
Sürekli SMM ile konuşmacı tanıma sistemi (a) Eğitim Aşaması (b) Test Aşaması

3.1. Veritabanı

Bu çalışmada TIMIT (Zue ve diğ., 1990) veritabanı kullanılmıştır. TIMIT veritabanında değişik aksanlara sahip 438'i erkek, 192'si kadın olmak üzere, anadili İngilizce olan toplam 630 konuşmacıya ait 10'ar cümle bulunmaktadır. Yapılan deneylerde TIMIT veritabanının test dizininden 20'si erkek ve 20'si kadın olmak üzere 40 kişilik bir alt grup kullanılmıştır. Her kişinin 7 cümlesi eğitim, 3 cümlesi ise test için kullanılmıştır. Testler ayrık yapılmıştır. Yani her kişinin 3 cümlesi ayrı ayrı test edilerek ortalama başarım hesaplanmıştır (40 kişi için toplam 120 test).

3.2. Özellik Çıkarımı

Çalışmada kişileri temsil etmek üzere mel frekansı kepstrum katsayıları (*mfcc*) kullanılmıştır. Şekil 3'de *mfcc* çıkarma işleminin adımları görülmektedir. Özellik çıkarma işleminde ses sinyali önce transfer fonksiyonu $H(z) = 1 - 0.95z^{-1}$ olan ön-vurgulama süzgecinden geçirilerek yüksek frekans bileşenleri vurgulanır. Sonraki adımda ön-vurgulama işlemi yapılmış ses işareti 10 ms'lik kısımları örtüşen 20 ms uzunluğunda çerçevelere ayrılır. 20 ms uzunluğundaki her bir çerçeve Hamming penceresi ile pencerelenir. Hızlı Fourier dönüşümü ile pencerelenen işaretin genlik spektrumu elde edildikten sonra işaret 0-8000 Hz frekans aralığına yerleştirilmiş 40 adet mel ölçekli süzgeç takımından geçirilir. Elde edilen spektrum genliğinin logaritması alındıktan sonra, ayrık kosinüs dönüşümü ile tekrar zaman ortamına geçilerek *mfcc* katsayıları elde edilir.



Şekil 3:
Mfcc vektörlerinin elde edilmesi

4. DENEYSEL ÇALIŞMA

Değişik boyutta özellik vektörleri ve değişik SMM parametre değerleri için konuşmacı tanıma oranları elde edilmiştir. Tanıma deneyleri 12, 16, 20 ve 24 boyutlu özellik vektörleri ile modelin M=1, 2, 4, 8, 16, 32, 64 karışım ve S=1, 2, 3, 4 durum sayılarının bütün kombinasyonları için tekrarlanmıştır. Yani her bir özellik vektör boyutu için toplam 28 adet test yapılmıştır. Tablo I ergodik ve soldan-sağa sürekli SMM ile 12 *mfcc* kullanılması halinde elde edilen konuşmacı tanıma başarımlarını göstermektedir. Tabloda verilen *S* modelin durum sayısını, *M* ise karışım sayısını belirtmektedir. Ayrıca tablodaki *S-S* soldan-sağa, *E* ise ergodik SMM'i belirtmektedir.

Tablo I.
12 *mfcc* kullanılarak değişik durum ve karışım sayıları için elde edilen ergodik ve soldan-sağa sürekli SMM ile konuşmacı tanıma oranları

		M=1	M=2	M=4	M=8	M=16	M=32	M=64
S=1	E	90.83	95.00	99.17	100	100	100	100
	S-S	90.83	95.83	96.67	100	100	99.17	97.50
S=2	E	90.00	100	99.17	100	100	100	100
	S-S	84.17	95.83	96.67	100	99.17	99.17	97.50
S=3	E	93.33	99.17	100	100	100	99.17	98.33
	S-S	87.50	96.67	99.17	99.17	99.17	98.33	95.83
S=4	E	98.33	99.17	99.17	99.17	100	99.17	100
	S-S	90.00	97.50	95.83	99.17	98.33	95.83	93.33

Aynı şekilde, 16, 20 ve 24 *mfcc* kullanılması durumunda elde edilen konuşmacı tanıma oranları da sırasıyla Tablo II, Tablo III, ve Tablo IV 'de verilmiştir.

Tablo II.
16 *mfcc* kullanılarak değişik durum ve karışım sayıları için elde edilen ergodik ve soldan-sağa sürekli SMM ile konuşmacı tanıma oranları

		M=1	M=2	M=4	M=8	M=16	M=32	M=64
S=1	E	95.00	98.33	100	100	100	100	100
	S-S	95.00	98.33	98.33	100	100	100	100
S=2	E	95.83	99.17	100	100	100	100	100
	S-S	90.83	98.33	98.33	100	100	100	100
S=3	E	96.67	100	100	100	100	100	100
	S-S	93.33	97.50	99.17	100	100	100	97.50
S=4	E	99.17	100	100	100	100	100	99.17
	S-S	93.33	96.67	99.17	100	100	97.50	89.17

Tablo III.
20 *mfcc* kullanılarak değişik durum ve karışım sayıları için elde edilen ergodik ve soldan-sağa sürekli SMM ile konuşmacı tanıma oranları

		M=1	M=2	M=4	M=8	M=16	M=32	M=64
S=1	E	97.50	97.50	100	100	100	100	100
	S-S	97.50	98.33	100	99.17	100	100	100
S=2	E	94.17	98.33	100	100	100	100	100
	S-S	93.33	97.50	100	100	100	100	100
S=3	E	95.83	100	100	100	100	100	100
	S-S	97.50	97.50	100	100	100	100	97.50
S=4	E	100	100	100	100	100	100	97.50
	S-S	96.67	99.17	100	100	100	99.17	89.17

Tablo IV.
24 *mfcc* kullanılarak değişik durum ve karışım sayıları için elde edilen ergodik ve soldan-sağa sürekli SMM ile konuşmacı tanıma oranları

		M=1	M=2	M=4	M=8	M=16	M=32	M=64
S=1	E	95.83	99.17	98.33	100	99.17	100	100
	S-S	95.83	99.17	98.33	100	100	100	100
S=2	E	97.50	98.33	99.17	100	100	100	98.33
	S-S	95.00	96.67	100	99.17	100	100	100
S=3	E	100	100	100	99.17	100	99.17	100
	S-S	95.00	98.33	99.17	100	100	100	96.67
S=4	E	99.17	100	100	100	100	99.17	97.50
	S-S	95.00	98.33	99.17	99.17	99.17	99.17	92.50

Tablolardan görüldüğü üzere tanıma başarımı, özellik vektör boyutu, durum sayısı ve karışım sayısının bir fonksiyonu olarak karşımıza çıkmaktadır. Ancak genel olarak bazı sonuçlara varmak gerekirse, özellik vektör boyutundaki ve karışım sayısındaki artmanın başarımda artış sağladığı, fakat durum sayısındaki artmanın aynı sonucu doğurmadığı söylenebilir. Örneğin karışım sayısı 1 ve 64 için, hem ergodik hem de soldan-sağa SMM'de durum sayısının artması tanıma oranını düşürmektedir. Durum sayısının verdiği bu sonuçlar, (Rabiner ve Juang, 1993) ile de uyumludur. Ayrıca, ergodik SMM ile soldan-sağa SMM'ler karşılaştırıldığında, ergodik SMM'in soldan-sağa SMM'e göre daha iyi başarım verdiği görülmektedir.

Yapılan deneylerden anlaşıldığı gibi TIMIT veritabanından alınan 40 kişilik konuşmacı kümesi için hem ergodik hem de soldan-sağa SMM ile en iyi başarımlar 20 *mfcc* kullanılması durumunda elde edilmiştir. Soldan-sağa SMM ile yapılan bütün deneylerde en düşük başarımların ise karışım sayısının 1 olduğu durumlarda görülmektedir. Hem ergodik hem de soldan-sağa SMM için 32 karışım ve 1 durum kullanılması özellik vektörü boyutu ne olursa olsun %100 konuşmacı tanıma başarımı vermektedir. Bir diğer genel sonuç ise özellik boyutu ne olursa olsun 4 durum kullanıldığında 64 karışımlı soldan-sağa SMM konuşmacı tanıma başarımını ortalama %7 oranında düşürmektedir.

5. SONUÇ

Bu makalede, ergodik ve soldan-sağa olmak üzere iki farklı saklı Markov modelleri kullanan metinden bağımsız konuşmacı tanıma sistemine ilişkin parametreler (özellik vektörü boyutu, model durum ve karışım sayıları) tanıma başarımına etkisi yönünden karşılaştırmalı olarak incelenmiş ve bunların optimum değerleri TIMIT veritabanından rastgele seçilen 40 kişilik bir konuşmacı grubu için belirlenmiştir. Esas olarak, tanıma başarımı özellik vektör boyutu, durum sayısı ve karışım sayısının bir fonksiyonu olarak karşımıza çıkmaktadır. Ancak, genel olarak, özellik vektör boyutunun ve karışım sayısının artması tanıma başarımını artırmaktadır. 32 karışım ve 1 durumlu ergodik SMM özellik vektör boyutundan bağımsız olarak en iyi başarımı verirken, soldan-sağa SMM'de en iyi başarımı veren parametre değerleri 16 karışım ve 1 durumdur. %100 tanıma başarımı en çok 20 boyutlu özellik vektörü ile elde edildiğinden, en iyi başarımı veren özellik boyutu olarak değerlendirilmiştir.

6. KAYNAKLAR

1. Doddington, G. R. (1985) Speaker Recognition-Identifying People by Their Voices, *Proceedings of the IEEE*, 73(11), 1651-1985.
2. Furui, S. (1997) Recent Advances in speaker recognition, *Pattern Recognition Letters*, 18, 859-872.
3. Li, Q., Zheng J., Tsai A. ve Zhou Q. (2002) Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition, *IEEE Trans. On Speech and Audio Processing*, 10(3), 146-157.
4. Matsui, T. ve Furui, S. (1994) Comparison of Text-Independent Speaker Recognition Methods Using VQ-distortion and Discrete/Continuous HMMs, *IEEE Trans. On Speech and Audio Processing*, 456-459.
5. Mirghafori N., Hatch O. A., Stafford S., Boakye K., Gillick D., ve Peksin B. (2005) ISCI's 2005 Speaker Recognition System, *IEEE Workshop on Automatic Speech Recognition and Understanding*, 23-28.
6. Naik, J. M., Netsch, L. P. ve Doddington, G. R. (1989) Speaker Verification Over Long Distance Telephone Lines, *IEEE International Conference on Acoustic, Speech and Signal Processing*, 524-527.

7. O'Shaughnessy, D. (1986) Speaker Recognition, *IEEE Acoustic, Speech and Signal Processing Magazine*, 4-17.
8. Rabiner, L. ve Juang B. H. (1993) *Fundamentals Of Speech Recognition*, Printice Hall, New Jersey.
9. Rabiner, L. R. (1989) A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, 77(2), 257-286.
10. Rabiner, L. R. ve Juang, B. H. (1986) An Introduction to Hidden Markov Models, *IEEE Acoustic, Speech and Signal Processing Magazine*, 4-16.
11. Rosenberg, A. E., Lee, C. H. ve Gökcen S. (1991) Connected Word Talker Verification Using Whole Word Hidden Markov Models*, International Conference on Acoustic, Speech and Signal Processing,* 381-384.
12. Yu. K, Mason. J. ve Oglesby, J. (1995) Speaker recognition using Hidden Markov Models, dynamic time warping and vector quantisation, *IEE Proc.-Vis. Image Signal Process.*, 142(5), 313-318.
13. Zheng. Y. ve Yuan. B. (1988) Text-dependent speaker identification using circular Hidden Markov Models, *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, 580-582.
14. Zue V., Sneff S., ve Glass J. (1990) Speech database development at MIT: TIMIT and beyond, *Speech Communication*, 9, 351-356.

Makale 24.05.2007 tarihinde alınmış, 11.07.2007 tarihinde kabul edilmiştir. İletişim Yazarı: F. Ertaş (fertas@uludag.edu.tr).