KONUŞMA TANIMA İÇİN NOMA İLE TEK-KANALDA KONUŞMA-MÜZİK AYRIŞTIRMA
SINGLE-CHANNEL SPEECH-MUSIC SEPARATION USING NMF FOR AUTOMATIC SPEECH RECOGNITION

Cemil Demir[1,3], Mehmet Uğur Doğan[1], A. Taylan Cemgil[2], Murat Saraçlar[3]

[1]TÜBİTAK-BİLGEM, Kocaeli, Türkiye
[2]Bilgisayar Mühendisliği, Boğaziçi Üniversitesi, İstanbul,Türkiye
[3]Elektrik-Elektronik Mühendisliği, Boğaziçi Üniversitesi, İstanbul,Türkiye
(cdemir|mugur)@tubitak.uekae.gov.tr, (taylan.cemgil|murat.saraclar)@boun.edu.tr

ÖZETÇE

Bu çalışmada özellikle televizyonda konuşma tanıma uygulamalarında tanıma başarımını önemli oranda düşüren arka plan müziğinin konuşmadan ayrıştırılması için çalışmalar yapılmıştır. Ayrıştırma tek-kanalda yapılacak olduğundan, konuşma ve müzik sinyallerinin eğitim verileri kullanılarak modellenmesi gerekmektedir. Konuşma ve müzik sinyalleri Negatif Olmayan Matris Ayrıştırma (NOMA) yöntemiyle modellenmiştir. Kullback-Leibler (KL) yöntemi kaynakların modellenmesinde kullanılmış ve ayrıştırma performansı incelenmiştir. KL-NOMA yöntemi daha önce konuşma-müzik ayrıştımada kullanılmış olmasına rağmen değişik eğitim kümelerinde NOMA yönteminin performansının karşılaştırılması ilk defa bu çalışmada yapılmıştır. Sinyallerin modellenmesi için kullanılan eğitim verilerinin ayrıştırma performansına etkisini incelemek amacıyla farklı eğitim kümeleri oluşturularak performans analizi yapılmıştır. Kullanılan yöntemlerin performansları ayrıştırma kriterleriyle birlikte konuşma tanıma performansına olan etkileriyle de ölçülmüştür.

ABSTRACT

In this study, single-channel speech source separation is carried out to separate the speech from the background music, which degrades the speech recognition performance especially in broadcast news transcription systems. Since the separation is done using single observation of the source signals, the sources have to be previously modeled using training data. Non-negative Matrix Factorization (NMF) methods are used to model the sources. In order to model the source signals, different training data sets, which contain different music and speech data, are created and the effect of the training data sets are analyzed in this study. The performances of the methods are measured not only using separation performance measure but also with speech recognition performance measures.

1. GİRİŞ

Son zamanlarda haber bültenlerini yazılandırmak için geliştirilen Konuşma Tanıma (KT) uygulamaları popüler hale gelmiştir. Televizyon ve radyodaki haber bültenlerini yazılandırmak için geliştirilen bu uygulamalardaki başlıca problemlerden bir tanesi konuşmanın arkaplanında müzik olduğunda geliştirilen KT sistemlerinin performansının ciddi oranda düşmesidir [1, 2]. Bundan dolayı arkaplan müziğini temizlemek, gürbüz KT sistemleri geliştirmek için çok önemlidir. Gerçek hayatta kullanılacak bir KT sistemi, gelecek olan ses sinyalinde önce konuşma-müzik bölütlemesi yapabilecek; daha sonra bu bölütleme sonucunda konuşma-müzik karışımı olarak etiketlenen kısımlarda konuşma-müzik ayrıştırma yapabilecek yeteneğe sahip bir ön modüle sahip olmalıdır. Daha önce yapılan çalışmada [3] KT sistemleri için geliştirilen konuşma-müzik bölütleme yöntemi anlatılmıştır. Tek-kanalda birden fazla konuşmacıya ait konuşmaların birbirinden ayrıştırılması üzerine yapılan bir çok çalışma [4] olmasına rağmen tek kanalda konuşma-müzik ayrıştırma üzerine pek çalışılmamıştır [5, 6, 7]. Tek-kanalda kaynak ayrıştırmada genel olarak Model-temelli ayrıştırma yöntemleri kullanılmakla beraber şimdiye kadar model-temelli yaklaşımlar, aynı sınıftan kaynakların, örneğin farklı konuşmacılara ait konuşmaların [8] ve müzikteki farklı enstrümanların [9], birbirinden ayrılması için kullanılmıştır.

Bu çalışmada Negatif Olmaya Matris Ayrıştırma (NOMA) yöntemlerinin konuşma-müzik ayrıştırma performanslarının ölçülmesi ve NOMA modellerini eğitmek için kullanılan eğitim kümelerinin ayrıştırma performansına etkisinin incelenmesi amaçlanmıştır. Bu çalışmada Kullback-Leibler NOMA (KL-NOMA) yöntemi konuşma-müzik ayrıştırmada kullanılacaktır. KL-NOMA yöntemi daha önce konuşma-müzik ayrıştırma için kullanılmış olmasına rağmen farklı eğitim kümelerinin ayrıştırma başarımı üzerine olan etkileri ilk defa bu çalışmada incelenmiştir.

Bildirinin içeriği şu şekildedir: 2. bölümde, uygulanacak NOMA yöntemi incelenecek ve bu yöntemle konuşma-müzik ayrıştırmanın nasıl yapılacağı anlatılacaktır. 3. bölümde ayırma ve konuşma tanıma deneyleri için kullanılan düzenekler ve elde edilen sonuçların nicel çözümlemesi yapılacaktır. 4. bölümde bu çalışmayla elde edilen çıkarımlar ve gelecekte yapılabilecek çalışmalara yer verilecektir.

2. YÖNTEM

Tek-kanalda konuşma-müzik ayrıştırma yapmak için konuşma ve müzik kaynaklarının eğitim verileri kullanılarak modellenmesi gerekmektedir. Bu modelleme sırasında kullanılacak özniteliklerin ve modelleme yönteminin seçimi önemli olmaktadır. Birden fazla kaynağın toplamı olan karışım sinyalinin öznitelikleri kaynaklara ait negatif olmayan özniteliklerin toplamına eşit olduğu durumlarda NOMA yöntemlerinin kullanılması uygun olmaktadır. Büyüklük Spektrogramı (BS) bu tür özniteliklerdendir. NOMA yöntemi Lee ve Seung [10] tarafından veri incelemede kullanılması amacıyla k-means ve PCA yöntemlerine alternatif olarak önerilmiştir. NOMA yönteminde verilen negatif olmayan veri matrisi, X, için negatif olmayan bileşen matrisleri bulunmaya çalışılmaktadır. Bu bileşen bulma işlemini matematiksel olarak aşağıdaki gibi gösterebiliriz.

$$\mathbf{X} \approx \mathbf{UV} \tag{1}$$

Bu gösterimde U şablon vektörlerini V ise bu şablon vektörlerine ait uyarım değerlerini temsil etmektedir. BS veri matrisi olarak kullanıldığında şablon vektörleri konuşma yada müziğin karakteristik özelliklerini barındıran vektörleri, uyarım matrisi de her bir zaman için bu karakteristik vektörlerine ait uyarımları içermektedir. Konuşma sinyali için yapılan çalışmalarda şablon vektörlerinin konuşmayı oluşturan fonları temsil ettiği gösterilmiştir.

2.1. KL-NOMA

KL-NOMA yönteminde veriye ait olan BS, X, ile şablon ve uyarım matrislerinin çarpımı arasındaki KL uzaklık ölçütü

$$D(X\|U,V) = -\sum_{u,t} X_{ut} \log \frac{[UV]_{ut}}{X_{ut}} - [UV]_{ut} + X_{ut} \tag{2}$$

en azaltılmaya çalışılmaktadır. Bu gösterimde u ve t sırasıyla frekans ve zaman indekslerini göstermektedirler. Bu uzaklık ölçütünün en azaltılmasını sağlayan çarpımsal güncelleme denklemleri [10] aşağıdaki gibidir:

$$U = U. * (((X./(UV))V^T)./(\mathbf{1}V^T)) \tag{3}$$

$$V = V. * ((U^T(X./(UV)))./(U^T\mathbf{1})) \tag{4}$$

Bu gösterimde $\mathbf{1}$, birlerden oluşan uygun boyutlu matrisi göstermektedir.

2.2. NOMA ile Konuşma-Müzik Ayrıştırma

NOMA ile konuşma-müzik ayrıştırmada, eğitim sırasında konuşma ve müzik sinyallerine ait olan BS matrisleri kullanılarak her bir sinyale ait şablon matrisleri öğrenilmektedir. Bu eğitimi

$$S = U_s V_s \quad \text{and} \quad M = U_m V_m. \tag{5}$$

şeklinde gösterebiliriz. Bu gösterimde U_s veU_m sırasıyla konuşma ve müzik sinyalleri için öğrenilen şablon matrislerini temsil etmektedir. Şablon ve uyarım matrisleri çarpımsal güncelleme denklemleri kullanılarak hesaplanmaktadır. Ayrıştırma sırasında, konuşma ve müzik sinyalleri için eğitilmiş olan şablon matrisleri kullanılarak genel şablon matrisi oluşturulur. Genel şablon matrisi sabitlenerek karışım sinyalinin BS matrisine karşılık gelen genel uyarım matrisi çarpımsal güncelleme denklemleri yardımıyla hesaplanır. Bu ayrıştırmayı

$$X = [U_s^* U_m^*][(V_s^*)^T (V_m^*)^T] \tag{6}$$

şeklinde gösterebiliriz. Konuşma ve müzik sinyaline karşılık gelen uyarım matrisleri ve eğitilmiş olan şablon matrisi yardımıyla karışım içindeki konuşma ve müzik sinyalleri geri çatılır. Geri çatma işlemi elde edilen şablon ve uyarım matrisleri kullanılarak her bir kaynağın sonsal olasılıklarını en büyütecek şekilde yapılmaktadır. Matematiksel olarak şablon ve uyarım matrisleri

$$(U_s^*, V_s^*, U_m^*, V_m^*) = \arg \max_{U_s,V_s,U_m,V_m} p(X|U_s, V_s, U_m, V_m). \tag{7}$$

şeklinde seçilmektedir. Bileşen matrisleri belirlendikten sonra konuşma ve müzik kaynakları, kaynakların birleşik sonsal olabilirliklerini en büyütecek şekilde seçilmektedir. Bu seçimi

$$(\widehat{S}, \widehat{M}) = \arg \max_{S,M} p(S, M|X, U_s^*, V_s^*, U_m^*, V_m^*). \tag{8}$$

şeklinde ifade edebiliriz. Bu sonsal olabilirliği en büyütecek kaynak geri çatımları

$$\widehat{S} = X. * \frac{U_s^* V_s^*}{(U_s^* V_s^* + U_m^* V_m^*)}. \tag{9}$$

$$\widehat{M} = X. * \frac{U_m^* V_m^*}{(U_s^* V_s^* + U_m^* V_m^*)}. \tag{10}$$

şeklinde hesaplanmaktadır.

3. DENEYSEL SONUÇLAR

3.1. Başarım Ölçütleri:

Yaptığımız çalışmada konuşma-müzik ayrıştırma ile amaçlanan KT başarımını arttırmak olduğu için ayrıştırma yöntemlerinin performansları KT başarım ölçütü olan Kelime Doğruluk Oranıyla (KDO) ile ölçülmüştür. Aynı zamanda KT başarımı ile ayrıştırma başarımı arasındaki ilişkiyi incelemek amacıyla yöntemlerin ayrıştırma başarımları da ölçülmüştür. Ayrıştırma başarımlarını ölçmek amacıyla ayrıştırılan konuşma içindeki kalan müzik miktarını ölçmek amacıyla Konuşma-Müzik Oranı (KMO), müzik içinde kalan konuşma miktarını ölçmek amacıyla Müzik-Konuşma Oranı (MKO), konuşmada meydana gelen bozulmayı ölçmek amacıyla Konuşma-Bozulma Oranı (KBO) ve müzikte meydana gelen bozulmayı ölçmek amacıyla Müzik-Bozulma Oranı (MBO) kullanılmıştır.

3.2. Deney Düzeneği:

Bu çalışmada konuşma-müzik ayrıştırmada kullanılan eğitim verilerinin ayrıştırma başarımına etkisini ölçme amacına uygun olarak deney düzenekleri hazırlanmıştır. Deney kümesi; 8 konuşmacıya ait yaklaşık 2 saat uzunluğundaki konuşmaların 4 saniye uzunluğundaki bir cıngıl ile 0, 5, 10, 15 ve 20 dB seviyelerinde yapay olarak karıştırılmasıyla oluşturulmuştur. Kullanılan cıngıllar televizyon haberlerinde kullanılan cıngıllardan seçilmiştir. KL-NOMA için kullanılan BS matrisi

1024 boyutlu pencereleri 512 birim kaydırarak elde edilen çerçevelerin Fourier dönüşümleri alınarak hesaplanmıştır. Eğitim verisi olarak her bir konuşmacı için; kendisine ait başka konuşmalarından oluşan "Kendisi", kendisi dışındaki aynı cinsten olan insanların konuşmalarından oluşan "Diğerleri" ve kendisi ile birlikte kendi cinsinden olan diğer konuşmacılara ait konuşmaların bulunduğu "Tümü" adlı konuşma veritabanları oluşturulmuş ve bu veriler kullanılarak her konuşmacı için KL-NOMA modelleri oluşturulmuştur. Müzik modellerini eğitmek için de benzer bir yaklaşım kullanılmıştır. Ancak müzik modellerinde "Kendisi" veritabanında müziğin orijinal hali kullanılmıştır. Konuşma ve müzik için kullanılan 3 farklı modelin çaprazlanması sonucu konuşma-müzik ayrıştırmada kullanılacak 9 farklı model çeşidi ortaya çıkmıştır. Örneğin diğer konuşmacılara ait verilerin kullanılmasıyla oluşturulan konuşmacı modeliyle birlikte sadece müziğin kendi verileriyle oluşturulan model kullanıldığında; ayrıştırma için kullanılan model Diğerleri-Kendisi (DK) olmaktadır. Bu modellere ait sonuçlar incelenerek konuşma müzik ayrıştırmada konuşma ve müziğe ait eğitim verilerinin ayrıştırma performansına olan etkileri tespit edilmeye çalışılmıştır. Aşağıdaki Tablo 1'de konuşma ve müzik NOMA modellerini eğitmek için kullanılan verilerin özellikleri gösterilmiştir.

Tablo 1: Eğitim Verisi Özellikleri

Özellikler	Konuşma			Müzik		
	Kendisi	Diğerleri	Herkes	Kendisi	Diğerleri	Herkes
Süre(Sn)	120	360	480	4	116	120
Şablon vektör sayısı	200	500	500	50	500	500

3.3. Konuşma Tanıma Sistemi

Geliştirilen KT sistemininde kullanılan cinsiyet-bağımlı akustik modeller yaklaşık 50'şer saatlik konuşma verileri kullanılarak eğitilmiştir. Akustik model eğitim birimi olarak bağlam-bağımlı üçlüsesler kullanılmıştır. Öznitelik olarak 25 ms uzunluğundaki pencerelerin 10 ms kaydırılması sonucu elde edilen çerçevelerin 13 boyutlu MFKK'ları kullanılmıştır. Bu MFKK vektörlerine fark ve fark-fark vektörleri de eklenerek nihai 39 boyutlu öznitelik vektörleri oluşturulmuştur. KT sisteminde kullanılan dil modeli 200 milyon kelime içeren gazete haber metinlerinden 30 bin kelimelik bir sözlük için üç gram olasılıklarının hesaplanması yoluyla elde edilmiştir.

3.4. Eğitim Verilerinin Performans Analizi:

NOMA modellerini eğitmek için kullanılan eğim verilerinin ayrıştırma performansına etkisini incelemek için oluşturulan 9 modelin kullanılmasıyla elde edilen KMO değerleri Tablo 2'de gösterilmiştir. KMO değerleri incelendiğinde müzik için Kendisi modeli kullanıldığında konuşma için kullanılan modelin Tümü veya Diğerleri olmasının KMO değerlerini etkilemediği görülmüştür. Bu gözlem Tablo 3'deki KBO değerleri ve Tablo 4'deki KDO değerleri için de geçerlidir. Konuşma için kullanılan model Kendisi olduğunda ise; Tümü modeli kullanıldığında elde edilen KMO ve KDO değerlerinin Diğerleri modeli kullanıldığında elde edilen değerlere göre daha yüksek olduğu tespit edilmiştir. Müzik için kullanılan Kendisi modeliyle Konuşma için kullanılan Kendisi modellerinin farklılık

Tablo 2: KL-NOMA yöntemiyle elde edilen ortalama çıktı KMO değerleri (dB)

Çıktı KMO (dB)		Girdi KMO (dB)				
Müzik	Konuşma	0dB	5dB	10dB	15dB	20dB
Kendisi	Kendisi	13.9	22.6	31.1	39.2	47.5
	Tümü	10.4	19.8	29.2	37.7	46.5
	Diğerleri	10.6	19.9	29.3	37.6	46.2
Tümü	Kendisi	**13.9**	**22.9**	**31.5**	**40.1**	**48.6**
	Tümü	9.7	19.5	29.0.1	38.3	47.4
	Diğerleri	9.8	19.5	29.3	38.2	47.2
Diğerleri	Kendisi	12.3	21.6	30.3	39.4	48.1
	Tümü	7.8	17.9	27.7	37.2	46.5
	Diğerleri	8.0	18.0	27.9	37.1	46.4

Tablo 3: KL-NOMA yöntemiyle elde edilen ortalama çıktı KBO değerleri (dB)

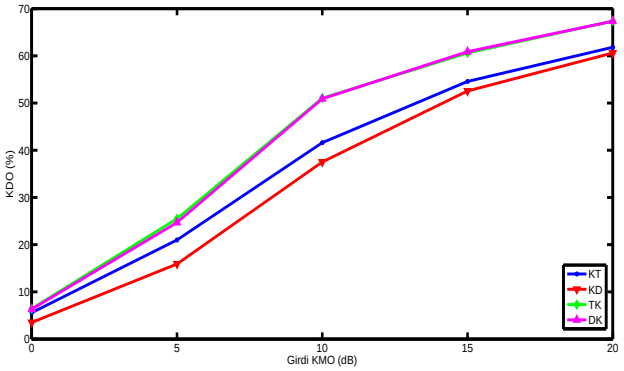
Çıktı KBO (dB)		Girdi KMO (dB)				
Müzik	Konuşma	0dB	5dB	10dB	15dB	20dB
Kendisi	Kendisi	11.6	14.4	16.9	19.7	22.2
	Tümü	**12.1**	**14.8**	**17.5**	**20.3**	**23.1**
	Diğerleri	12.1	14.8	17.5	20.2	23.1
Tümü	Kendisi	7.8	9.2	10.7	11.2	11.9
	Tümü	9.4	11.2	13.1	13.8	14.8
	Diğerleri	9.1	10.8	12.4	13.4	14.3
Diğerleri	Kendisi	7.8	9.3	10.5	11.5	12.3
	Tümü	9.3	11.2	12.7	14.1	15.2
	Diğerleri	9.1	10.9	12.1	13.6	14.7

göstermesinin sebebi çalınan müziğin orijinal halinin NOMA modeli oluştururken kullanılmasına rağmen, konuşma için konuşmacıya ait başka konuşmaların NOMA modelini eğitmek için kullanılmasıdır. Müziğin orijinal hali model eğitmede kullanıldığında konuşmacıya ait konuşmaların konuşma modelini eğitmek için kullanılan kümede bulunup bulunmaması önemini yitirmektedir. Konuşma ve müzik için Kendisi modelleri ile birlikte kullanılan Tümü ve Diğerleri modellerinin KT performanslarının karşılaştırılması Şekil 1'de görülmektedir.

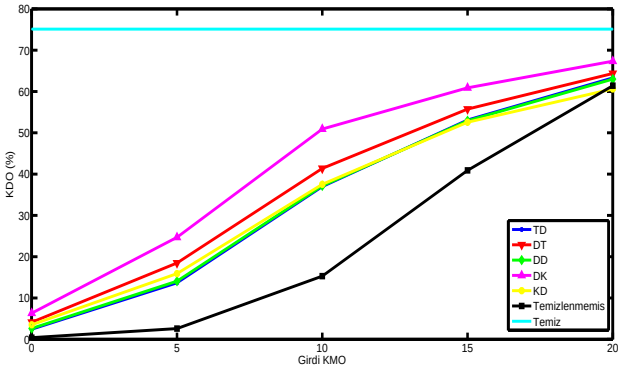
Ayrıştırma ve KT sonuçları incelendiğinde yapılan diğer bir tespit de konuşma yada müziğin kendisinin bulunmadığı eğitim kümeleri kullanılarak eğitilen modellerin kendilerinin bulunmadığı kümeler kadar olmasa da KT sonuçlarını hiç ayrıştırma yapılmadığı duruma göre iyileştirmesidir. Bu iyileştirme Şekil

Tablo 4: KL-NOMA yöntemiyle elde edilen ortalama KDO değerleri (dB)

KDO (%)		Girdi KMO (dB)				
Müzik	Konuşma	0dB	5dB	10dB	15dB	20dB
Referans Sonuçlar	Temiz	75.1	75.1	75.1	75.1	75.1
	Karışım	0.4	2.6	15.3	40.9	61.4
Kendisi	Kendisi	**11.7**	**33.1**	**54.1**	**62.8**	**67.7**
	Tümü	6.5	25.5	51.0	60.6	67.4
	Diğerleri	6.3	24.7	50.9	60.9	67.3
Tümü	Kendisi	5.6	21.0	41.6	54.6	61.9
	Tümü	4.1	17.7	42.0	56.8	64.2
	Diğerleri	4.1	18.5	41.4	55.8	64.4
Diğerleri	Kendisi	3.5	15.9	37.5	52.6	60.6
	Tümü	2.5	13.7	37.0	53.1	63.3
	Diğerleri	2.6	14.1	37.2	52.9	63.0



Şekil 1: Müzik ve Konuşma için kullanılan 'Kendisi' modellerinin KT performanslarının karşılaştırılması.



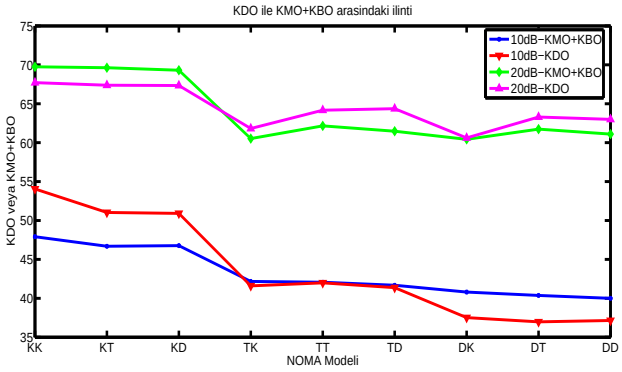
Şekil 2: Konuşma veya müzik için 'Diğerleri' eğitim kümesi kullanıldığında NOMA yönteminin KT performansları.

2'de görülmektedir.

Tablo 2,3 ve 4 incelendiğinde KMO veya KBO değerlerinin KDO değerlerini açıklamak için tek başlarına yeterli olmadıkları görülmüştür. Ancak yapılan incelemede KMO ve KBO değerlerinin toplamları ile KDO değerleri arasındaki ilintinin 0.93 olduğu görülmüştür. Bu ilinti Şekil 3 incelendiğinde görülmektedir.

4. SONUÇ

Bu çalışmada KT performansını arttırmak amacıyla kullanılabilecek NOMA temelli konuşma-müzik ayrıştırma yöntemi geliştirilmiştir. Daha önceki yapılan çalışmalardan farklı olarak bu çalışmada konuşma ve müziğe ait NOMA modellerini eğitmek için farklı eğitim kümeleri oluşturularak bu eğitim kümelerinin ayrıştırma performansına olan etkileri incelenmiştir. Müziğin kendisine ait olan verilerle oluşturulan model kullanıldığında konuşmacıya ait verilerin konuşma eğitim kümesinde bulunup bulunmamasının ayrıştırma performansını etkilemediği görülmüştür. Aynı zamanda konuşmacıya yada çalınan müziğe ait örnekler eğitim kümesinde bulunmadığı durumda da kullanılan NOMA yönteminin KT performansını arttırdığı görülmüştür. Gelecekte yapılacak çalışmalarda konuşma veya müzik için herhangi bir eğitim kümesi kullanılmadığında ayrıştırma performansının nasıl etkilendiği incelenecektir.



Şekil 3: 10 ve 20 dB değerleri için KMO+KBO ile KDO arasındaki ilinti grafiği

5. TEŞEKKÜR

Murat Saraçlar TÜBA-GEBİP tarafından desteklenmektedir. Ali Taylan Cemgil, bu çalışmada, TUBİTAK tarafından 110E292 Bayesci Tensör ayrıştırma (BAYTEN) projesi kapsamında desteklenmektedir.

6. KAYNAKÇA

[1] B. Raj, V.N. Parikh, and R.M. Stern, "The effects of background music on speech recognition accuracy," in *Proc. of ICASSP*, 1997.

[2] E. Arısoy, H. Sak, and M. Saraçlar, "Language modeling for automatic Turkish broadcast news transcription," *Proc. of Interspeech*, 2007.

[3] C. Demir and M. U. Doğan, "Konuşma Tanıma İçin Konuşma-Müzik Bölütleme Sistemi," *Proc. of SIU*, 2009.

[4] M.N. Schmidt and R.K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. of ICSLP*, 2006.

[5] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-Negative Matrix Factorization Based Compensation of Music for Automatic Speech Recognition," in *Proc. of Interspeech*, 2010.

[6] S. Kirbiz and B. Gunsel, "Perceptual single-channel audio source separation by non-negative matrix factorization," in *in proc. of SIU*, 2009, pp. 416–419.

[7] S. Yildirim and M. Saraclar, "Single channel music and speech separation using non-negative matrix factorization," in *in proc. of SIU*, 2009, pp. 301–304.

[8] P. Smaragdis, M. Shashanka, M. Inc, and B. Raj, "A Sparse Non-Parametric Approach for Single Channel Separation of Known Sounds," Proc. of NIPS, 2009.

[9] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on ASLP*, vol. 15, no. 3, pp. 1066–1074, 2007.

[10] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, 1999.