

# Değişimli Oto-Kodlayıcılar kullanarak Diyalog Geliştirme<sup>1</sup>

## Dialogue Enhancement using Variational Autoencoders

Serap Kırbız

Mühendislik Fakültesi, Elektrik Elektronik Mühendisliği Bölümü

MEF Üniversitesi

kirbiz@mef.edu.tr

ORCID 0000-0001-7718-3683

### Özet

Bu makalede, kaynak ayırma algoritmalarından faydalanarak birden fazla kaynaktan oluşan ses kayıtlarında konuşma işaretlerini güçlendirmek amacıyla bir yöntem sunulmaktadır. Konuşma sesleri ve diğer sesler arasındaki doğru dengeyi sağlamak, dinleyici şikayetleri arasında sıkça dile getirilen önemli bir sorun olarak öne çıkmaktadır. Bu çalışmada, diyalog içeren ses kayıtlarından diyalogların ayırılmasında negatif olmayan gürültü giderici oto kodlayıcı modelleri kullanılmakta ve bu diyaloglar, diğer seslerle farklı oranlarda yeniden birleştirilerek, kullanıcı tercihlerine uygun bir dinleme deneyimi sunulmaktadır. Önerilen yöntem, akan veri üzerinde çalışabilme özelliğine sahip olup, televizyon programları gibi gerçek zamanlı uygulamalara da uygulanabilmektedir. Anahtar kelimeler: kaynak ayırma, diyalog geliştirme, derin öğrenme

### Abstract

The paper proposes a method to enhance speech signals in audio recordings consisting of multiple sources by using source separation algorithms. Achieving the right balance between speech sounds and other sounds is a frequently mentioned issue among listener complaints. In this study, non-negative denoising variational autoencoder models are used to separate dialogues from audio recordings containing dialogues, and these dialogues are remixed with other sounds at different rates to provide a listening experience that suits the user's preferences. The proposed method has the ability to work on streaming data and can also be adapted to real-time applications such as television programs.

Keywords: source separation, dialogue enhancement, deep learning

### 1. Giriş

Bu makalede, gürültülü karışım işaretlerindeki diyalogları ayırmak için derin öğrenme tabanlı modeller kullanılarak, kaynaklar istatistiksel olarak incelenmekte ve kaynak ayırma için yeni yöntemler geliştirilmektedir. Geleneksel kaynak ayırma yöntemleri genellikle karışım işaretinin genlik spektrogramını giriş olarak kullanır ve bu karışımı, kendisini oluşturan kaynak işaretlerine ayırmayı hedefler. Ayırılacak işaretler negatif olmayan matrisler olarak ele alındığından, bu alanda yaygın bir yaklaşım olan Negatif

Olmayan Matris Ayırma (NOMA) [1], önemli bir rol oynamaktadır. Matris ayırmanın temel amacı,  $\mathbf{X} \in \mathbb{R}_+^{M \times N}$  şeklindeki giriş matrisini  $\mathbf{W} \in \mathbb{R}_+^{M \times K}$  ve  $\mathbf{H} \in \mathbb{R}_+^{K \times N}$  matrislerine ayırmaktır. Bu işlemde,  $\mathbf{H}$ , “katkılar”,  $\mathbf{W}$  ise “sözlükler” olarak adlandırılır ve hedef,  $\mathbf{X}$  matrisinin doğru bir şekilde geri çatılmasıdır:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{W}\mathbf{H}. \quad (1)$$

NOMA ile  $\mathbf{X}$  karışım işaretinin temsili için,  $K$  kertesinde  $\mathbf{H}$  katkıları ve  $\mathbf{W}$  şablonlarının sözlüğü elde edilmektedir. NOMA [2] kullanılarak gerçekleştirilen eğitilmiş kaynak ayırmada ise, öncelikle her bir kaynak işaretinin temsili için ayrı bir  $\mathbf{W}$  sözlüğü öğrenilmektedir. Sonrasında, bir karışım işareti kaynaklarına ayırılmak istendiğinde, sözlükler bir araya getirilerek, her bir kaynağın katkısı kestirilmektedir.

Önerilen çalışmada, tek kanallı kaynak ayırma problemini çözmek için zayıf etiket denetimine dayalı “negatif olmayan gürültü değişimli oto-kodlayıcılar ve oto kod-çözümler” kullanılmaktadır. Karışımı oluşturan kaynakları kestirmek için toplamsal bileşenlerin kullanılması, yöntemin NOMA’ya benzerliğini ortaya koysa da önerilen model, doğrusal şablonlar yerine doğrusal olmayan şablonlar ve etkili sinir ağı modelleri kullanarak daha esnek bir çözüm sunmaktadır.

Bunun yanı sıra, önerilen kaynak ayırma yöntemi, hedeflenen işlevlerin gereksinimlerine uygun kaynak modellerini birleştirmek ve eşleştirmek amacıyla derin üretici modellerini kullanmaktadır [3]. Bu yaklaşımla, her kaynak için zayıf denetimli eğitimle değişimli oto-kodlayıcılar [4, 5] öğrenilmiş ve değişimli oto-kod çözümler [6] gibi derin saklı-değişken modeller süreçte kullanılmıştır. [4] ile verilen çalışmamızda hem eğitim hem de test aşamalarında kaynakların sınıf etiketleri bilinirken, [3] ile verilen çalışmamızda ise, kaynak ayırmanın yanı sıra karışımı oluşturan kaynakların sınıf etiketlerinin tahmini de gerçekleştirilmiştir.

Bu makalede önerilen yöntem, iki temel işlevi yerine getirmek için tasarlanmıştır: kaynak ayırma, diyalog geliştirme. Deneyler, farklı eğitim stratejileri ve farklı sınıf sayıları için farklı veri tabanlarında gerçekleştirilmiştir.

Bu çalışmadaki katkılar aşağıdaki şekilde özetlenebilir:

- Kaynak işaretlerine erişim sağlanmadığı ve yalnızca sınıf etiketlerine sahip olduğu bir senaryo için negatif olmayan bir Değişimli Oto Kodlayıcı (VAE) modeli ile kaynak ayırma önerilmektedir.
- Kaynak işaretleriyle ilgili sınıf etiketlerine ihtiyaç duyulmadan, Değişimli Oto Kod Çözümleri (VAD)

<sup>1</sup> Bu çalışma EEEAG/215E076 numaralı araştırma projesi kapsamında TÜBİTAK tarafından desteklenmektedir.

modeli ile sınıf etiketi kestirimi ve kaynak ayrıştırma eş zamanlı gerçekleştirilebilir. Modelde, yaklaşık dağılım parametreleri rasgele başlatılmakta ve gradyan-temelli yöntemlerle optimizasyon sağlanmaktadır.

- Zaman bölgesinde ses kalitesini ölçen Ortalama Karesel Hata (MSE), Ölçek-Değişmez İşaret-Gürültü-Oranı (SI-SNR) ve İşaret-Gürültü-Oranı (SNR) gibi başarımlı ölçütlerinin eğitim esnasında kayıp fonksiyonu olarak kullanılması sağlanmıştır.
- Geliştirilen yöntem, gerçek zamanlı olarak çalışabilme yeteneğine sahiptir.
- Önerilen modellerin iyi bir başarımlı elde edebilmesi için en az üç sınıftan gözleme ihtiyaç olduğu gösterilmektedir.

## 2. Ön Bilgi

Kaynak ayrıştırma problemlerinde, derin öğrenme yaklaşımları yaygın olarak uygulanmaktadır [7]. Bu yaklaşımlar, genellikle karışım işaretlerini alarak, bu işaretlerden kaynak işaretlerini ayırmayı hedefleyen gürültü ayırıcı değişimli oto-kodlayıcılar (VAE) gibi derin öğrenme modellerini kullanmaktadır. Derin öğrenme modellerinin başarılı bir şekilde eğitilebilmesi için büyük miktarda etiketli veri gereklidir. Ancak, zayıf denetleme, etiketli verilerin eksik olduğu veya yalnızca sınıf bilgisi ile çalışılan senaryolarda, yüksek soyutlama düzeyine sahip bilgi ve gürültü içerdiği için etkili bir yaklaşım olarak öne çıkmaktadır.

Bu çalışmada, bilgilendirilmiş kaynak ayrıştırma [8] ile ilişkili derin öğrenme yöntemleri kullanılmıştır. Önerilen çalışma, eğitim için kaynak işaretlerine erişimin olmadığı, sadece kaynakların ait olduğu sınıf bilgilerine erişimin olduğu bir senaryoyu ele almaktadır. Bu durum, birçok karışım işareti ve bu karışımların içerdiği kaynakların etiketleri gözlemlendiğinde, karışım işaretinde her bir kaynağın hangi bölümlerde yer aldığına dair sezgisel tahminlerin yapılmasına olanak tanır.

### 2.1. Değişimli Oto Kodlayıcılar

Değişimli Oto-Kodlayıcılar (VAE) [5], olasılıksal üretici modellerinin oluşturulmasında yaygın bir şekilde kullanılmaktadır. VAE' nin temel yapısında, gizli değişken  $\mathbf{z}$  için  $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$  Gauss önsel dağılımı kullanılır ve parametresi  $\theta$  olan  $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$  olasılıksal kodlayıcı modeli ile parametresi  $\phi$  olan olasılıksal  $q_\phi(\mathbf{z}|\mathbf{x})$  kod çözücü modeli tanımlanabilir. Sinir ağı modellerine dayalı kodlayıcı-kod çözücü çifti için, genel amaçlı bir fonksiyon tahmin modeli kullanılabilir.  $N$  gözlemden oluşan  $\mathcal{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$  eğitim veri kümesi, bileşen olabirliğinin değişimli alt sınırının olasılıksal bayır çıkışı ile en büyüklenmesi sonucunda modelin  $\phi$  ve  $\theta$  parametrelerini kestirebilir:

$$\mathcal{L}(\theta, \phi|\mathbf{x}^{(i)}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})]. \quad (2)$$

(2) ile verilen ifade,  $D_{KL}$ , Kullback-Leibler (KL) ıraksayıdır. Önsel  $p_\theta(\mathbf{z})$  dağılımı ve sonsal  $q_\phi(\mathbf{z}|\mathbf{x})$  kestiriminin Gauss varsayımı altında, analitik yöntemlerle değişimli bir alt sınır elde edilebilmektedir. Bununla birlikte, KL ıraksayı,  $q_\phi(\mathbf{z}|\mathbf{x})$  sonsal kestirimi üzerinde  $p_\theta(\mathbf{z})$  önsel dağılımını zorlayan bir düzenleyici işlevi görmektedir. Geri çatma hatasını içeren

$\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})$  dağılımının beklentisi, yeterince büyük bir veri kümesi mevcutsa,  $q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$  sonsal dağılımından tek bir örnek kullanılarak hesaplanabilir. Yeniden parametrelendirme ile  $q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$  dağılımından örnekleme yapılabilir. Kodlayıcı fonksiyonunun çıktıları,  $\mu(i)$  ve  $\sigma(i)$  olacak şekilde, izotropik bir Gauss dağılımı varsayılmaktadır:

$$q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) = \mathcal{N}(\mathbf{z}|\mu(i), \sigma^2(i)\mathbf{I}). \quad (3)$$

(3) ile verilen ifade,  $\epsilon \sim \mathcal{N}(\epsilon|\mathbf{0}, \mathbf{I})$  yardımcı gürültü değişkeni,  $\odot$  ise eleman eleman çarpma işlemidir. Yeniden parametrelendirme  $\mathbf{z} = \mu + \sigma \odot \epsilon$  şeklinde yapılabilir.

Önsel  $p_\theta(\mathbf{z})$  dağılımının izotropik Gauss varsayımı altında, ayrı ve bağımsız gizli birimler teşvik edilmektedir.  $\beta$  -VAE [6], orijinal VAE' nin bir uzantısı olup,  $\mathbf{z}$  gizli değişkeninin sınırlandırılmasıyla, daha ayrıştırılmış temsiller elde edilmesini sağlar.  $\beta$  -VAE' nin orijinal VAE' den tek farkı, alt sınırdaki kullanılan  $\beta$  katsayısıdır:

$$\mathcal{L}_\beta(\theta, \phi|\mathbf{x}^{(i)}) = -\beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})||p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})]. \quad (4)$$

(4) ile verilen ifade  $\beta = 1$ , (2) eşitliğindeki alt sınır iken,  $\beta > 1$  durumunda gizli değişken daha fazla kısıtlanmakta ve geri çatma kalitesine daha az ağırlık verilmektedir.

Oto-kodlayıcı kullanılarak gerçekleştirilen kaynak ayrıştırmada  $\mathbf{x}^{(i)}$  karışımında bulunan,  $\mathbf{s}_k^{(i)}$  orijinal kaynak işareti, gürültü giderici oto-kodlayıcı aracılığıyla tahmin edilebilmektedir:

$$\hat{\mathbf{s}}_k^{(i)} = f_{\theta_k}(g_{\phi_k}(\mathbf{x}^{(i)})), \quad f_{\theta_k}(\cdot) \geq 0. \quad (5)$$

(5) ile verilen ifade  $g_{\phi_k}(\cdot)$  ve  $f_{\theta_k}(\cdot)$ ,  $k$ . kaynak sınıfı için kodlayıcı ve kod çözücü çiftini temsil eder.  $\mathbf{x}^{(i)}$  ve  $\mathbf{s}^{(i)}$ , sırasıyla karışım işaretinin ve karışımı oluşturan orijinal işaretlerin genlik spektrogramları olmak üzere  $\mathcal{X}_s = \{(\mathbf{x}^{(i)}, \mathbf{s}^{(i)})\}_{i=1}^N$  eğitim veri kümesinde, (6) eşitliği ile verilen maliyet fonksiyonu kullanıldığında [9] yaklaşımına ulaşılabilir. Ancak, kayıp fonksiyonu olarak kare hata değil de genelleştirilmiş KL ıraksayı seçilmelidir.

$$D_{GKL}(\mathbf{s}_k^{(i)}||\hat{\mathbf{s}}_k^{(i)}) = \sum_j \mathbf{s}_{k_j}^{(i)} \log \frac{\mathbf{s}_{k_j}^{(i)}}{\hat{\mathbf{s}}_{k_j}^{(i)}} - \mathbf{s}_{k_j}^{(i)} + \hat{\mathbf{s}}_{k_j}^{(i)}. \quad (6)$$

(6) eşitliğinde,  $j$ ,  $\mathbf{s}_k^{(i)}$  kaynak işaretinin elemanlarının indeksidir. Eğitim kümesi  $\mathcal{X}_h = \{(\mathbf{x}^{(i)}, \mathbf{h}^{(i)})\}_{i=1}^N$  olup,  $K$  sınıf için kaynak sınıf etiketleri  $\mathbf{h}^{(i)} \in \{0,1\}^K$  ile ifade edilmektedir. Önerilen çalışmada,  $\hat{\mathbf{x}}^{(i)}$  kaynaklarını kestirmek için  $\mathbf{s}_k^{(i)}$  orijinal kaynakları yerine sadece  $h_k^{(i)}$  kaynak etiketleri kullanılmaktadır:

$$\hat{\mathbf{x}}^{(i)} = \sum_{k=1}^K \hat{\mathbf{s}}_k^{(i)} h_k^{(i)}. \quad (7)$$

$\hat{\mathbf{s}}_k^{(i)}$  kaynaklarını kestirmek için, kayıp fonksiyonu olarak (6) ile verilen  $D_{GKL}(\mathbf{s}_k^{(i)}||\hat{\mathbf{s}}_k^{(i)})$  kullanılmaktadır. Böylece,  $\mathbf{x}^{(i)}$  karışımının geri çatılması için sadece kaynak etiketleriyle ilişkili oto-kodlayıcıların kullanılması hedeflenmektedir. Geliştirilen model, NOMA yöntemine benzer şekilde karışımı ayrıştırmak amacıyla toplamsal bileşenleri kullanır. Ancak, bileşenleri temsil etmek için NOMA' da sadece doğrusal şablonlar alınabilirken, önerilen çalışmada doğrusal olmayan

şablonlar ve daha etkili sinir ağı yapılarıyla esnek bir çözüm sunulmaktadır. Ayrıca, modelde her bir kodlayıcı, diğer kaynaklara ilişkin gürültüyü ihmal edip, yalnızca kendi sınıfına ait öznelikleri öğrenirken; kod çözümler de ilişkili kaynak işaretlerini geri çatmaktadır.

Şimdiye kadar, modelin standart oto-kodlayıcıları baz aldığı varsayılmıştır. Bu çalışma,  $\beta$  -VAE'nin, Poisson kod çözümleri ile kullanılması önerilmektedir:

$$p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z}) = \mathcal{PO}(\mathbf{x}^{(i)}|\mathbf{z}) = \hat{\mathbf{x}}^{(i)}. \quad (8)$$

(8) ile verilen ifade  $\mathbf{z}$ , model çıktısıdır. Poisson dağılımının genlik spektrogramını modellemek için kullanılması, KL ıraksayının normalize edilmeden kullanımına karşılık gelmektedir [10].

Ayrıştırılmış kaynakları zaman bölgesinde elde etmek için karışımın işaretinin fazı ile yumuşak Wiener filtresi kullanılır:

$$\hat{\mathbf{y}}_t^{(i)} = \text{STFT}^{-1} \left( \frac{(\hat{s}_t^{(i)})^2}{\sum_c (\hat{s}_c^{(i)})^2} \odot \mathbf{x}^{(i)} \odot e^{j\Phi} \right). \quad (9)$$

(9) eşitliğinde  $\Phi$ , karışım spektrogramının fazı,  $c$  bileşen indeksi,  $t$  hedef bileşen indeksidir.  $\text{STFT}^{-1}$ , kısa zamanlı ters Fourier dönüşümüdür.

### 3. Önerilen Metot

#### 3.1. Model

Oto-Kodlayıcı Değişimli Bayes (AEVB), üretici modeller arasında güçlü ve etkili bir yöntem olarak kabul edilmektedir. AEVB algoritması, sonsal dağılımın yaklaşık parametrelerini kodlayıcı kullanarak kestirir.

Ancak, tüm veriye erişimin olmadığı durumlar için AEVB algoritmasının alternatif bir uygulaması önerilmektedir. Önerilen yöntem, değişimli oto kod çözümleri (VAD) olarak adlandırılmaktadır [11]. VAD ile, yaklaşık sonsal dağılımın parametreleri başlangıçta rasgele seçilir ve gradyan-temelli yöntemler ile bu parametreler güncellenir.

#### 3.2. Değişimli Oto-Kodlayıcı

Değişimli oto-kodlayıcı (VAE) [5], üretici modelleri geliştirmek için kullanılan bir çerçevedir.  $\mathcal{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ , eğitim veri kümesinin  $N$  örneğini içersin. Verinin  $z$  gizli değişkeninin rasgele bir süreci tarafından üretildiği varsayımı altında  $\mathbf{z}$  için  $p_{\theta}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$  önsel dağılımı,  $\phi$  ile parametrelendirilmiş kodlayıcı modeli,  $\theta$  ile parametrelendirilmiş bir  $p_{\theta}(\mathbf{x}|\mathbf{z})$  kod çözümleri tanımlanabilir.  $\mathcal{X}$  eğitim veri kümesi kullanıldığında,  $\phi$  ve  $\theta$  model parametreleri, stokastik gradyan çıkışı ve (2) numaralı denklemde ifade edilen değişimli alt sınır en büyüklere tahmin edilmektedir.

Gerçekleştirilen VAE modeli, [4] ile önerilen VAE modelini kullanmakla birlikte, önerilen çalışmada bir adet  $g(\cdot)$  kodlayıcısı kullanılır ve bu kodlayıcı ile  $z_k$  değişkenleri ve  $h_k$  sınıf etiketleri kestirilir:

$$(\mathbf{z}, \mathbf{h}) = g(\mathbf{x}^{(i)}). \quad (10)$$

Kaynak ayrıştırma işlevi için oto-kodlayıcı kullanıldığında  $\mathbf{s}_k^{(i)}$  orijinal kaynak işaretini,  $\mathbf{x}^{(i)}$  karışım işaretinden ayrıştırmak için bir oto-kod çözümleri kullanılmaktadır:

$$\hat{\mathbf{s}}_k^{(i)} = f_{\theta_k}(\mathbf{z}_k), \quad f_{\theta_k}(\cdot) \geq 0. \quad (11)$$

Sinir ağının tek bir ileri geçişiyle, yukarıdaki model sayesinde test aşamasında hızlı çıkarım yapılabilmektedir. Ancak bu yaklaşımda, eğitilmiş kodlayıcı yalnızca eğitim kümesindeki sınıf kombinasyonları üzerine uzmanlaşmıştır. Eğitim sırasında gözlemlenmeyen bir sınıf kombinasyonuna sahip karışım işareti test sırasında modele verildiğinde, model bu karışımı ayrıştırmada ve doğru sınıflandırmada zorlanmaktadır. Öte yandan, kod çözümleri kaynak sınıflarının üretken modellerini öğrenme kapasitesine sahiptir ve eğitim kümesindeki verinin dağılımına büyük ölçüde bağlı değildir. Böylece, daha uyarlanabilir ve eğitilebilir yöntemler için yalnızca eğitimli oto-kod çözümleri kullanabilmemizi sağlamaktadır.

#### 3.3. Değişimli Oto-Kod Çözümleri

Önerilen yöntemde, eğitim bittikten sonra  $g(\cdot)$  kodlayıcı atılmakta ve test aşamasında kestirim için yalnızca  $f_k(\cdot)$  kod çözümleri kullanılmaktadır. Bu esnada amaçlanan,  $z_k$  ve  $h_k$  tahmininde, kodlayıcının tahminleri yerine optimizasyon kullanılmaktadır. Bu süreçte eğitilmiş kod çözümler parametreleri sabit tutulmaktadır. [11] çalışmasındaki VAD benzeri bir yaklaşım benimsenmektedir.

Orijinal VAE modelinde olduğu gibi  $z_k = \mu_k + \sigma_k \odot \epsilon$  yeniden parametrelendirilmesi yapılır ve KL ıraksayı kullanılır. Bu yöntemle,  $z_k$  doğrudan tahmin edilmemekte; bunun yerine normal dağılıma sahip gizli  $z_k$  değişkeni için ortalama  $\mu_k$  ve varyans  $\sigma_k^2$  tahmin edilmektedir. Yalnızca kod çözümleri kullanan çıkarımda iki farklı yöntem uygulanmaktadır: tam kapsamlı ve birleşik.

Tam kapsamlı ters-VAE çıkarımında, tüm sınıf kombinasyonları değerlendirilerek, doğrudan birleştirme problemi çözümleri çalışılır.  $h_k^{(i)}$  için uygun değerler ayarlanır ve aşağıda verilen model bütün olabilir sınıf kombinasyonları için en iyilenir:

$$\hat{\mathbf{x}}^{(i)} = \sum_{k=1}^K \hat{s}_k^{(i)} h_k^{(i)}. \quad (12)$$

Son aşamada, tüm örnekler için kaybı en düşük olan  $h_k^{(i)}$  değeri seçilir. Seçilmiş  $h_k^{(i)}$  değerleri, tahmin edilen sınıf etiketleridir ve ilgili  $z_k$  değerleri, kaynak işaretinin kestirilmesini sağlar.

Alternatif olarak, yalnızca kod çözümlerinin ortak çıkarımında,  $h_k^{(i)}$ 'nin olası değerleri için  $\{0, 1\}$  yerine  $(0, 1]$  aralığı alınarak, optimizasyon probleminin konveks olmaması sağlanır.

Genelleştirilmiş KL ıraksayı kullanıldığında,  $z_k$  ve  $h_k$  değerleri tahmin edilir ve aşağıdaki ifadeyi en küçüklemek amacıyla Adam en iyileştirme uygulanır:

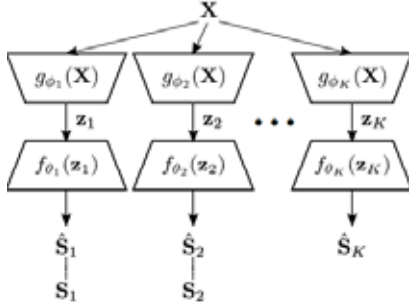
$$\arg \min_{z_k \in \mathbb{R}, h_k \in (0, 1]} D_{GKL}(\mathbf{x}^{(i)} \parallel \hat{\mathbf{x}}^{(i)}). \quad (13)$$

#### 3.4. Eğitim

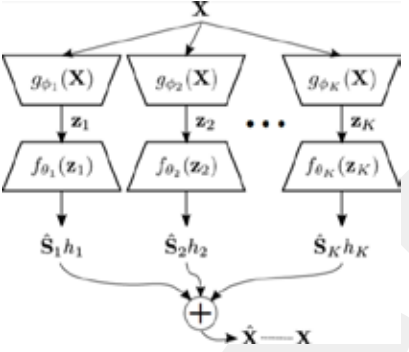
Eğitim esnasında, yalnızca sınıf etiketleri veya kaynak işaretleri kullanılabilir.

##### 3.4.1. Kaynak İşaretiyle Eğitim

Eğitim esnasında kaynak işaretlerinin mevcut olduğu durumlarda, otomatik kodlayıcılar kullanılarak başlanır.  $\mathbf{x}^{(i)}$ , karışım işaretinin genlik spektrogramı olmak üzere, karışım işaretini oluşturan her bir  $\mathbf{s}_k^{(i)}$  orijinal kaynak işareti, (5) eşitliği



Şekil 1: Kaynak İşareti ile eğitim için model mimarisini



Şekil 2: Kaynak sınıfı ile eğitim için model mimarisini

ile verilen oto-kodlayıcıların kullanımı sonucunda tahmin edilebilir.

$\mathbf{x}^{(i)}$  ve  $\mathbf{s}^{(i)}$ , sırasıyla karışım işareti ve orijinal kaynak işaretlerinin genlik spektrogramları olmak üzere;  $\mathcal{X}_s = \{(\mathbf{x}^{(i)}, \mathbf{s}^{(i)})\}_{i=1}^N$  eğitim kümesinden [9] çalışmasında anlatılan yaklaşıma ulaşılabılır. Ancak, kayıp fonksiyonu olarak karesel hata yerine, (6) ile verilen KL ıraksayı kullanılmış olmalıdır.

Kaynak işaretleri kullanılarak eğitilen modelin yapısı Şekil 1' de verilmektedir. Şekil 1' de genelleştirilmiş KL ıraksayı noktalı çizgilerle belirtilmiştir. Daha sade bir gösterim için

$^{(i)}$  örnek indeksleri gösterilmeyip, karışım işareti  $\mathbf{x}'$  in, sadece 1. ve 2. kaynak işaretlerinden oluştuğu varsayılır. Kaynak işaretleri kullanılarak eğitilen modelde kayıp, genelleştirilmiş KL ıraksayı  $D_{GKL}(s_1 || \hat{s}_1) + D_{GKL}(s_2 || \hat{s}_2)$  kullanımı ile orijinal ve kestirilen kaynak işaretleri arasında hesaplanır.

### 3.4.2. Kaynak Sınıfı ile Eğitim

$\mathbf{h}^{(i)} \in \{0, 1\}^K$ , kaynak sınıf etiketleri olmak üzere, eğitim kümesi olarak  $\mathcal{X}_h = \{(\mathbf{x}^{(i)}, \mathbf{h}^{(i)})\}_{i=1}^N$  kullanılmaktadır. Modelde, kaynakları kestirmek için  $\mathbf{s}_k^{(i)}$  orijinal işaretlerine erişim olmayıp, yalnızca  $h_k^{(i)}$  kaynak etiketleri kullanılır:

$$\hat{\mathbf{x}}^{(i)} = \sum_{k=1}^K \hat{\mathbf{s}}_k^{(i)} h_k^{(i)}. \quad (14)$$

$\hat{\mathbf{s}}_k^{(i)}$  kaynaklarını kestirmek için,

$$D_{GKL}(\mathbf{x}^{(i)} || \hat{\mathbf{x}}^{(i)}) = \sum_j \mathbf{x}_j^{(i)} \log \frac{\mathbf{x}_j^{(i)}}{\hat{\mathbf{x}}_j^{(i)}} - \mathbf{x}_j^{(i)} + \hat{\mathbf{x}}_j^{(i)} \quad (15)$$

ile verilen  $D_{GKL}(\mathbf{x}^{(i)} || \hat{\mathbf{x}}^{(i)})$  genelleştirilmiş ıraksayında,  $j$ ,  $\mathbf{x}^{(i)}$ 'nin elemanlarının indeksidir.

Böylelikle,  $\mathbf{x}^{(i)}$  karışımının geri çatılması için sadece kaynak etiketleriyle ilişkili oto-kodlayıcılar kullanılmaktadır. Önerilen model, NOMA' daki gibi toplamsal bileşenler kullanmasına rağmen, NOMA' dan farklı olarak bileşenlerin temsilinde doğrusal yerine doğrusal olmayan şablonlar ile daha etkili sinir ağı modelleri kullanılabilir. Model, her bir kaynak işaretinin geri çatılmasında kod çözücü kullanırken, kodlayıcılar karışımı oluşturan diğer kaynaklara ait gürültü ve girişimleri ihmal eder ve oto-kodlayıcı kullanarak yalnızca ilişkili kaynak sınıfını öğrenir.

Eğitim esnasında kaynak sınıflarını kullanan model, Şekil 2' de görülmektedir. Kaynak sınıfları kullanılarak eğitilen modelde, maliyet fonksiyonu  $D_{GKL}(x || \hat{x} = \hat{s}_1 + \hat{s}_2)$ , karışım işareti  $\mathbf{x}$  ile kestirilen kaynaklar  $\hat{\mathbf{s}}_k$ ,  $k = 1, 2$  arasında hesaplanmaktadır.

### 3.5. Çıkarım

Tek kanallı kaynak ayrıştırma probleminde, kaynak işaretleri ile ilgili yeterince önsel bilgi mevcut değilse, sonsuz sayıda geçerli çözüm vardır. Kaynaklar hakkındaki önsel bilgi, üretici kaynak ayrıştırma yaklaşımlarında bir veya daha fazla üretici model aracılığıyla modele verilir. Bu yüzden, üretici modellerin yeni kaynak işaretleri oluşturmak için yeterince ön bilgisi bulunmaktadır. Çıkarım aşamasında, VAE ve VAD modelleri kullanılır. İleri çözüm, giriş işaretini (10) eşitliği ile gösterildiği gibi kodlayıcıya giriş olarak alarak, daha düşük boyutta gizli değişkenler üretir. Ters çözümde, kod çözücü, kodlayıcı tarafından üretilen gizli değişkenleri giriş olarak alarak, kaynak işaretlerini kestirir. İleri-çözümde VAE modeli kullanılırken, ters-çözümde test aşamasında kodlayıcı yerine VAD modeli kullanılmaktadır.

Ayrıştırılmış kaynakları zaman bölgesinde geri çatmak için karışım fazı ve (9) eşitliği ile verilen Wiener filtresi kullanılmaktadır.

### 3.6. Performans Tabanlı Maliyet Fonksiyonları

Kaynak ayrıştırma yöntemleri, ağırlıklı olarak genlik spektrogramları üzerinde çalışmaktadır. Kaynak ayrıştırma yöntemlerinde maliyet fonksiyonu olarak, genellikle ortalama karesel hata (MSE) [1], KL ıraksayı [1], Itakura-Saito ıraksayı [2], Bregman ıraksayı [12] kullanılır.

Son yıllarda derin öğrenme tabanlı kaynak ayrıştırma yöntemleri, doğrudan zaman bölgesinde çalışmaktadır ve ayrıştırılan işaretleri doğrudan zaman bölgesinde geri çatmaktadır. Kaynak ayrıştırma yöntemlerinde ayrıştırma başarımı zaman bölgesinde ölçüldüğü için, maliyet fonksiyonu olarak doğrudan zaman bölgesindeki başarımlar ölçütleri kullanılabilir. Böylelikle en iyileştirmek istenen kayıp fonksiyonu, kaynak ayrıştırma başarımını da ölçtüğünden hedeflenen uygulamada, ayrıştırma daha başarılı hale gelmektedir.

Önerilen çalışmada KL ıraksayına ek olarak İşaret Gürültü Oranı (SNR), ortalama karesel hata (MSE), ve ölçek-değişmez İşaret Bozulma Oranı (SI-SDR) gibi ölçütler de kullanılmaktadır. Böylece, kaynak ayrıştırma başarımını değerlendiren ölçütler, eğitim esnasında da kullanılmakta ve başarımın yükselmesi amaçlanmaktadır.

Ortalama karesel hata (MSE), aşağıdaki şekilde hesaplanmaktadır:

$$MSE = \frac{1}{L} \sum_{k=1}^L \left( y_t^{(i)}(k) - \hat{y}_t^{(i)}(k) \right)^2 \quad (16)$$

Burada  $k$ , zaman indeksini verirken,  $L$ , işaretin boyutudur.

BSS-Eval ölçütlerinden [13] İşaret Hata Oranı (SAR), İşaret Bozulma Oranı (SDR) ve İşaret Girişim Oranı (SIR), başarıyı ölçmek için konuşma ayırma uygulamalarında yaygın olarak kullanılmaktadır.

İşaret-Gürültü-Oranı (SNR) aşağıdaki gibi hesaplanır:

$$SNR = 10 \log_{10} \left( \frac{\|y_t^{(i)}\|^2}{\|y_t^{(i)} - \hat{y}_t^{(i)}\|^2} \right) \quad (17)$$

SNR, BSS-Eval ölçütlerinden SDR<sup>1</sup> ye karşılık gelmektedir. (17) ile verilen eşitlikte,  $y_t^{(i)}$  orijinal işaret,  $\hat{y}_t^{(i)}$  kestirilen işaret,  $y_t^{(i)} - \hat{y}_t^{(i)}$ , ise gürültüdür. Gürültünün, yani orijinal ve kestirilen işaretler arasındaki farkın, orijinal kaynak işaretine göre ortogonal olması için orijinal veya kestirilen işaretler yeniden ölçeklendirilmelidir [14]. Gürültünün orijinal işarete dik olacağı şekilde orijinal işaretin yeniden ölçeklendirilmesi;  $\hat{y}_t^{(i)}$  kestiriminin,  $y_t^{(i)}$  tarafından kapsanan çizgi üzerine dik izdüşümünün alınmasına ya da bu çizgi boyunca  $\hat{y}_t^{(i)}$  kestirimi için en yakın noktanın bulunmasına karşı gelir.

Ölçek-değişmez işaret-bozulma oranı (SI-SDR), SNR ölçütüne dayalı olarak, işaretin bozulma oranını daha doğru bir şekilde değerlendirmek için tanımlanmış bir ölçüttür. Orijinal işaretin ve kestirilen işaretin yeniden ölçeklendirilmesi ile bu değer, kaynak ayırma başarımını daha iyi temsil etmeyi amaçlar. SI-SDR aşağıdaki şekilde tanımlanmaktadır [14]:

$$SI - SDR = \frac{\|y_t^{(i)}\|^2}{\|y_t^{(i)} - \beta \hat{y}_t^{(i)}\|^2} = \frac{\|\alpha y_t^{(i)}\|^2}{\|\alpha y_t^{(i)} - \hat{y}_t^{(i)}\|^2} \quad (18)$$

(18) ile verilen tanımda,  $\alpha$  ölçeklendirme çarpanı olup, orijinal kaynak işareti için en iyi değer  $\alpha = \frac{y_t^{(i)T} \hat{y}_t^{(i)}}{\|y_t^{(i)}\|^2}$  olacak şekilde elde edilir.

### 3.7. Mimari

$T$  ve  $F$ , sırasıyla zaman ve frekans bileşenlerinin sayısını ifade ederken, önerilen model  $T \times F$  boyutunda spektrogramları giriş işareti olarak alan Evrişimsel sinir ağları (CNN) kullanmaktadır [15]. Kodlayıcı mimarisinin yapısı Tablo 1 ile verilmiştir. Kodlayıcı, üç evrişimsel katman, bir tam bağlı katman ve Gauss gizli katmanından oluşmaktadır. İlk evrişimsel katman,  $1 \times F$  boyutunda 128 filtre kullanarak, tıpkı NOMA' da olduğu gibi frekans eksenini boyunca şablonlar öğrenir. İkinci ve üçüncü katmanlar,  $4 \times 1$  boyutunda filtreler kullanan evrişimsel katmanlar olup, zamansal şablonları öğrenmektedir. Tam Bağımsız katmanda ve Gauss gizli çıktı katmanında sırasıyla 512 ve 128 filtre kullanılmaktadır. Kodlayıcı ve kod çözücünün yapısı simetrik. Kod çözücü, devrik evrişimler kullanır. Kodlayıcı ve kod çözücüde son katmanlar hariç diğer katmanların tümü düzeltilmiş doğrusal birim (ReLU) sapmaları kullanılmaktadır. Son katmanlar dışındaki her katmana toplu normalleştirme uygulanmaktadır. Çıktının negatif olmasını önlemek için kod çözücüde son katman olarak yumuşak artı işlevi (softplus)  $o(x) = \log(1 + e^x)$  kullanılmaktadır.

## 4. Deneyler

Önerilen yöntemin ayırma başarımı, tek gözlem içeren kaynak ayırma problemi çerçevesinde değerlendirilmiştir.

### 4.1. Veri Kümeleri

Geliştirilen modelin başarımı üç veri kümesi üzerinde değerlendirilmektedir. Kullanılan ilk veri kümesi, Sıfırdan Dokuz Konaşma Komutları (SC09) içeren 10 sınıflı bir veri kümesidir [16]. Bu küme, bir saniye süresine sahip rakam ifadelerinden oluşmaktadır. Eğitim kümesi, farklı kayıt koşulları altında farklı konuşmacıları, farklı ses yüksekliği, gürültü, hizalama ve yanlış etiketleme gibi durumlar içerdiğinden zordur. Veri kümesinde tamamı sessizlik olan kısımlar ve farklı dillerde ifade edilen sayılar dahi bulunmaktadır.

Diyalog seslerinin arka plan seslerinden ayrıştırılması hedeflendiği için, ikinci ve üçüncü veri kümelerinde film sesleri kullanılmaktadır. Başarımı nesnel olarak ölçebilmemiz için, tüm kaynaklara ayrı ayrı erişimimiz olmalıdır. Ancak, gerçek film kayıtlarında tüm kaynaklara erişim mümkün olmayacağı için başarımlar, nesnel olarak ölçülemez. Bu nedenle, diyalog ayırma başarımını ölçebilmek için, <sup>1</sup> adresinde bulunan film sesleri ile SC09 [16] ve GRID [17] veri tabanlarından alınan konuşma sesleri kullanılmaktadır.

İkinci veri kümesindeki kaynaklar konuşma ve arka plan sesleri şeklinde iki farklı tür içerdiğinden, sınıf-tabanlı eğitilen modeller için tanımlanabilirlik problemi yaratabilir. Bu problemi aşmak amacıyla üçüncü bir veri kümesi daha oluşturulmuştur. Üçüncü veri kümesi, ikinci veri kümesinde bulunan arka plan sesleri ile GRID veri Kümesinden [17] bir kadın ve bir erkek konuşmacının sesleri birleştirilerek oluşturulmuştur.

Her üç veri kümesi için, kayıtlar, 8,600 Hz ile alt örneklenmekte ve tüm kayıtların Karesel Ortalamalarının Kökü (RMS), hedeflenen RMS değerine normalleştirilmektedir. Bu RMS normalizasyonu, kayıtların ses yüksekliğindeki farklılıkları azaltmak için elzemdir.

İlk veri kümesi, 10 sınıflı SC09 veri kümesinden beş sınıf seçilerek oluşturulmuştur. Karışım işaretlerini elde etmek için  $C(5; 2) = 10$  etiket kombinasyonunun  $\{0,1\}$ ,  $\{0,2\}$ , vb. şekilde bir listesi oluşturulur. Sonrasında, kombinasyondaki sınıf etiketlerinden seçilen kayıtlar 0 dB ile karıştırılır. Veri kümesi sırasıyla 15,000 eğitim, 1,875 sağlama ve 1,875 test verisi içermektedir.

İkinci veri kümesi, iki sınıflı kaynak ayırma deneyi için kullanılmaktadır. Birinci sınıfta konuşma sesleri, ikinci sınıfta film arka plan sesleri bulunmaktadır. Arka plan sesleri, <sup>1</sup> adresinde bulunan film seslerinin konuşma işareti içermeyen kayıtlardan seçilmektedir. İlk veri kümesinde olduğu gibi, tüm sesler bir saniye süresindeki seslere bölünerek toplam 828 arka plan, 15,928 konuşma sesi oluşturulmaktadır. İki sınıftan rasgele seçilen işaretler 0 dB ile karıştırılmaktadır. İkinci veri kümesinde 15,000 eğitim, 1,875 sağlama ve 1,875 test verisi bulunmaktadır.

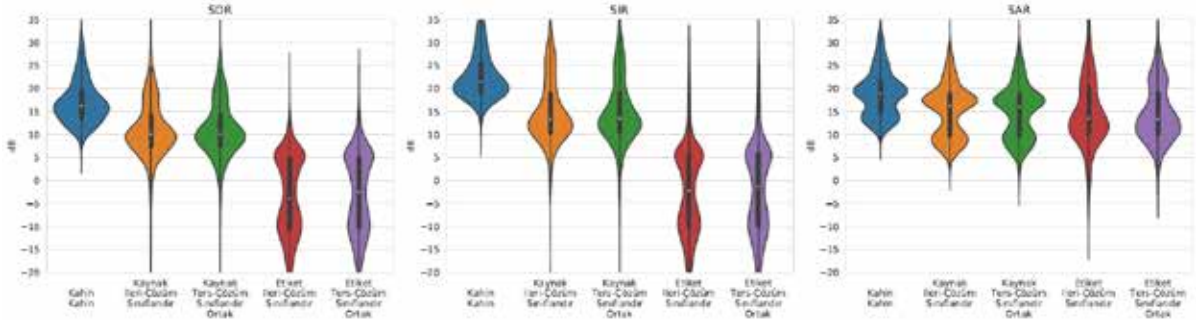
Tüm veri kümelerinde, karışım işaretlerine %50 örtüşme ile 400 noktalı Hanning penceresi kullanılarak kısa zamanlı Fourier Dönüşümü (STFT) uygulanmaktadır. Elde edilen spektrogramlar,  $F = 201$  frekans bileşeni ve  $T = 44$  zaman bileşeni içermektedir.

### 4.2. Değerlendirme Ölçütleri

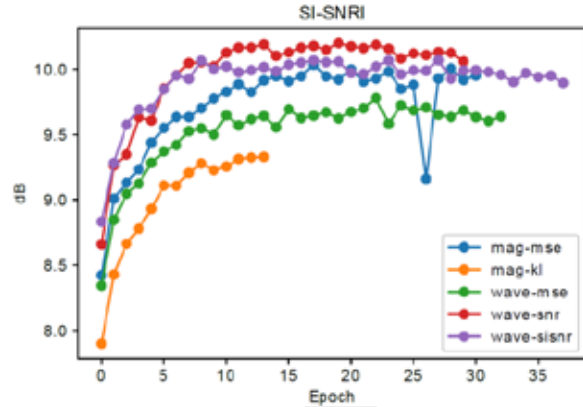
Kaynak ayırma sonuçları BSSEVAL araç kutusundaki [13] İşaret Hata Oranı (SAR), İşaret Bozulma Oranı (SDR) ve İşaret-Girişim Oranı (SIR) ile değerlendirilmektedir. Diğer kaynak işaretlerinden kalan girişim yani gürültü miktarı SIR ile,

<sup>1</sup> <https://soundbible.com/tags-movies.html>





Şekil 4: VAE modeli kullanılarak ikinci veri kümesi üzerinde elde edilen SAR, SDR ve SIR türünden başarımlar.



Şekil 5: Eğitim esnasında kullanılan farklı kayıp fonksiyonları için elde edilen SI-SNRI türünden başarımlar

kestirimi yaptığı yöntemleri ifade eder. Eğer işaret kestirimi için ters-çözüm kullanıldıysa, dördüncü ifadeye hangi ters çözüm yönteminin kullanıldığı belirtilmektedir: Ortak ve Kapsamlı. Yatay ekseninde SAR, SDR ve SIR türünden verilen başarımlarda ilk yöntemin altında sadece Kâhin, Kâhin yazmakta olup, bu durum sınıf etiketlerinin ve kaynak işaretlerinin bilindiği senaryoya karşılık gelmektedir ve önerilen yöntemlerin başarımlarını kıyaslamak için verilmiştir.

Öncelikle, Kaynak eğitilmiş modellerin SAR, SDR ve SIR değerlerinin beklenildiği üzere daha yüksek olduğu görülmektedir. Kaynak eğitilmiş modeller arasında SDR değerleri karşılaştırıldığında Ters-Çözüm işaret kestirimi kullanan modellerin 1-2 dB daha iyi başarımlar sergilediği dikkat çekmektedir. Bu sonuç, üretici modellerin öğrenebildiği senaryolarda, Ters-Çözüm işaret kestiriminin İleri-Çözüm işaret kestirimine kıyasla bir avantaj sunduğunu ortaya çıkarmaktadır. Etiket eğitilmiş modellerde de benzer bir başarımlar farkı gözlemlenmiş olup, yaklaşık 0.5 dB civarındaki bu fark, Kaynak eğitilmiş modellere kıyasla daha düşük seviyededir. Bu farkın, üretici modelin öğrenme kapasitesine bağlı olduğunu düşündüğümüzden, beklentilerimizle uyumlu olduğu görülmektedir.

İkinci deneyde, ikinci veri kümesi üzerinde VAE modeli kullanılarak kaynak işaret eğitimi ile sınıf etiket eğitimi karşılaştırılmıştır. Sonuçlar Şekil 4'te sunulmuştur. SDR ve SIR değerleri açısından değerlendirildiğinde, kaynak eğitilmiş ayrıştırıcıda SDR değeri ortalama 10 dB civarındayken, etiket eğitilmiş ayrıştırıcıda ortalama SDR değeri yaklaşık -5 dB olarak elde edilmiştir. Girişim miktarını ölçen SIR değerlerinde de benzer bir durum gözlemlenmiş ve kaynak eğitilmiş

ayrıştırıcının, etiket eğitime kıyasla ortalama 15 dB daha yüksek performans gösterdiği belirlenmiştir.

Bu veri kümesinde yalnızca iki sınıfın bulunması, yalnızca sınıf etiketleri kullanılarak çözülemeyen bir tanımlanabilirlik sorununa yol açmaktadır. Bu nedenle, sadece iki sınıf içeren durumlarda, etiket eğitimi ile gerekli üretici modelleri öğrenilememektedir. Bu sorunu aşmak amacıyla üçüncü deneyde, üç sınıflı bir diyalog geliştirme senaryosu tasarlanmıştır.

Üçüncü deneyde, Bölüm 3.5'te açıklanan başarımlar ölçütlerinin maliyet fonksiyonu olarak kullanılmasının başarımlara etkisi incelenmiştir. Deneyler, SC09 kümesinden seçilen üç sınıflık bir alt küme üzerinde gerçekleştirilmiştir. Bu kapsamda, 1,500 eğitim ve 1,000 test karışım işareti oluşturulmuştur. Deneylerde, kaynak işareti denetimi ve standart oto-kodlayıcı ile birlikte aşağıdaki maliyet fonksiyonları kullanılmıştır:

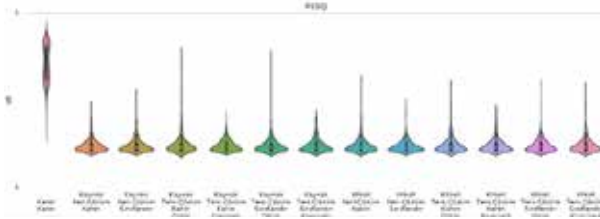
- **mag-mse**: Genlik spektrogramında ortalama karesel hata.
- **mag-KL**: Genlik spektrogramında Kullback-Leibler (KL) ıraksığı.
- **wave-mse**: Sesin zaman bölgesinde ortalama karesel hata.
- **wave-snr**: Sesin zaman bölgesinde İşaret-Gürültü-Oranı (SNR).
- **wave-sisnr**: Sesin zaman bölgesinde ölçek-değişmez İşaret-Gürültü-Oranı (SI-SNR).

Bu makalede önerilen evrişimli yapay ağ mimarilerimiz, genlik spektrogramına uygulanır. Ancak, zaman bölgesinde tanımlanan maliyet fonksiyonlarının kullanılabilmesi ve uçtan uca eğitim [19] gerçekleştirilebilmesi için PyTorch kütüphanesinden türevi alınabilir bir kısa zamanlı ters Fourier Dönüşümü (ISTFT) işlevi kullanılmaktadır. Bu sayede, gradyan bilgisi ISTFT aracılığıyla geri yayılabilmektedir.

Adil bir karşılaştırma sağlamak amacıyla, eğitim sırasında maliyet fonksiyonu değerlendirme kümesinde hesaplanmakta ve bu metrik erken durdurma için kullanılmaktadır. Değerlendirme kümesinde hesaplanan maliyet fonksiyonu 10 ardışık iterasyon boyunca azalma göstermezse, eğitim süreci sonlandırılmaktadır.

Farklı kayıp fonksiyonları kullanıldığında elde edilen SI-SNRI türünden başarımlar sonuçları Şekil 5'te görülmektedir. Kayıp fonksiyonları zaman bölgesinde hesaplandığında, ayrıştırma başarımları daha yüksek olurken, en başarılı ayrıştırma **wave-snr** kayıp fonksiyonu olarak kullanıldığında elde edilmiştir.

Üçüncü veri kümesinde yapılan deneylerde algısal konuşma kalitesi PESQ [20] değerleri ile ölçülmüş olup, sonuçlar Şekil 6 ile raporlanmaktadır. Kaynak ve etiket eğitilmiş



Şekil 6: Üçüncü veri kümesi üzerinde PESQ türünden başarımların kaman grafikleri olarak gösterilmiştir.

modeller için PESQ değeri 1.14 dB civarında olup, tüm eğitim-çıkartım çeşitleri için yakın sonuçlar elde edilmiştir.

Önerilen yöntem, akan veri üzerinde de uygulanabilir. Bir sistemin gecikmesi  $G_{toplam}$ , iki ana bileşenden oluşur:  $G_b$ , ilk çıktıyı üretmek için gereken işlem süresi ve  $G_i$ , her bir çıktının üretilmesi için gerekli işlem zamanıdır. Model, öncelikle, NVIDIA Tesla K80 GPU' suna yüklenmektedir. İşlemler, 400 örneklik zaman çerçeveleri için gerçekleştirilmektedir. Örneklem frekansı olarak 8,600 Hz kullanıldığı için, 400 örneğe sahip bir çerçeve 47 ms' lik bir süreyi kapsamaktadır. Bu koşullarda  $G_b = 3.3 ms$  olup, bir çerçeve,  $G_i = 1 ms$  sürede ayrıştırılmaktadır. Böylece, sistemin toplam gecikmesi  $G_{toplam} = 4.3 ms$  olmaktadır. Bu süre, sistemin filmlerde diyalogları ayırmak amacıyla rahatlıkla kullanılabileceğini gösterir. Buna ek olarak, ortak ve kapsamlı ters çözüm en iyileme yöntemleri de 10.06 s ve 14.52 s sürmektedir. Her iki yöntem de dengeli olmayan veri kümelerinde yüksek başarımlı ayrıştırma yapabilmekte olup, filmlerde diyalog ayrıştırmak için ihtiyaç duyulan işlem süreleri gereksinimlerin üzerindedir.

## 5. Sonuç

Bu makale, kaynak işaretlerine erişim sağlanmadığı ve yalnızca sınıf etiketlerine sahip olduğu ve olunmadığı senaryolar için negatif olmayan VAE ve VAD modelleri önermektedir.

Kaynak işaretleriyle ilgili sınıf etiketlerine ihtiyaç duyulmadan, sınıf etiketi kestirimi ve kaynak ayrıştırma eş zamanlı gerçekleştirilebilmektedir. Ayrıca, yalnızca karışım işaretlerine erişimin olduğu ve sınıf etiketleri ile kaynak işaretlerine erişimin olmadığı bir senaryo için VAD modeli önerilir. VAD ve VAE modelleri kullanılarak, kaynak ayrıştırma başarımları üç farklı veri kümesinde değerlendirilmiştir. Başarımlar, SAR, SDR ve SIR türünden ölçülmüş olup, eğitimde sınıf etiketlerinin ve orijinal kaynak işaretlerinin kullanıldığı senaryolarda elde edilen başarımların benzer olduğu gösterilmiştir. Kaynak işaretlerinin kestirimi için ileri çözüm ve ters çözüm yöntemleri karşılaştırıldığında, ters çözüm ile elde edilen başarımın daha iyi olduğu görülmüştür. Eğitim sırasında kaynak sınıflarına erişim olduğu durum ile kaynak sınıflarının kestirildiği durumlar karşılaştırılmış; kaynak sınıfına erişimin olmadığı durumda da benzer bir başarımla kaynakların ayrıştırılabildiği gözlemlenmiştir.

Ayrıca, zaman bölgesinde, uçtan uca çalışabilen bir kaynak ayrıştırma yöntemi geliştirilmiştir. Uçtan uca çalışan kaynak ayrıştırma ile, zaman bölgesinde ses kalitesini ölçen MSE, SI-SNR ve SNR gibi başarımların ölçülmesinde eğitim esnasında kayıp fonksiyonu olarak kullanılması da sağlanmıştır. Geliştirilen yöntem, gerçek zamanlı olarak çalışabilme yeteneğine sahiptir.

## 6. Kaynaklar

- [1] D. D. Lee, ve H. S. Seung, "Algorithms for non-negative matrix factorization", *Advances in neural information processing systems*, 2000.
- [2] C. Févotte, E. Vincent, ve A. Ozerov. "Single-channel audio source separation with NMF: divergences, constraints and algorithms", *Audio Source Separation*, Springer, 2018, 1-24.
- [3] Ç. Hızlı, E. Karamatlı, A. T. Cemgil, ve S. Kirbiz, "Değişimli Oto-Kodlayıcılar Kullanılarak Birleşik Kaynak Ayrıştırma ve Sınıflandırma-Joint Source Separation and Classification Using Variational Autoencoders", *In 28th IEEE Signal Processing and Communications Applications Conference (SIU)*, 2020.
- [4] E. Karamatlı, A. T. Cemgil, ve S. Kirbiz, "Audio Source Separation Using Variational Autoencoders and Weak Class Supervision", *IEEE Signal Processing Letters*, 2019, 1349-1353.
- [5] D. P. Kingma, ve M. Welling, "Auto-encoding Variational Bayes". *In Proc. ICLR*, 2014.
- [6] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, ve A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework", *In Proc. ICLR*, 2017.
- [7] D. Wang, ve J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview", *IEEE/ACM transactions on audio, speech, and language processing* 26.10 (2018): 1702-1726.
- [8] S. Kirbiz, A. Ozerov, A. Liutkus, ve L. Girin, "Perceptual coding-based Informed Source Separation," *2014 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, 2014, pp. 959-963.
- [9] E. M. Grais, ve M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders". *2017 IEEE global conference on signal and information processing (GlobalSIP)* 2017.
- [10] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models", *Computational intelligence and neuroscience*, 2009(1), 785152
- [11] A. Zadeh, Y. C. Lim, P. P. Liang, ve L. P. Morency, "Variational auto-decoder: A method for neural generative modeling from incomplete data." *arXiv preprint arXiv:1903.00840*, 2019.
- [12] S. Sra, ve I. S. Dhillon, "Generalized nonnegative matrix approximations with bregman divergences". *Advances in neural information processing systems*, 2006, pp. 283–290).
- [13] E. Vincent, R. Gribonval, ve C. Févotte, "Performance measurement in blind audio source separation", *IEEE transactions on audio, speech, and language processing* 14.4 (2006): 1462-1469.
- [14] J. Le Roux, S. Wisdom, H. Erdogan, ve J. R. Hershey, "SDR-Half-Baked or Well Done?", *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, (pp. 626–630).
- [15] W. Hsu, Y. Zhang, ve J. Glass, "Learning Latent Representations for Speech Generation and Transformation", *Interspeech 2016; Sep 8-12; San Francisco, CA. 2016. p. 1770-1774.*
- [16] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition". *arXiv preprint arXiv:1804.03209*, 2018.

- [17] M. Cooke, J. Barker, S. Cunningham, ve X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition”. *The Journal of the Acoustical Society of America* 120.5, 2006: 2421-2424.
- [18] D. P. Kingma, ve J. Ba, “Adam: A method for stochastic optimization *arXiv preprint arXiv:1412.6980*, 2014.
- [19] S. Venkataramani, E. Tzinis, ve P. Smaragdis, “End-to-end Non-Negative Autoencoders for Sound Source Separation”, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. p. 116-120.
- [20] A. W. Rix, J. G. Beerends, M. P. Hollier, ve A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs”. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, (pp. 749–752).

### Özgeçmişler



**Serap Kırbız**, doktora derecesini İstanbul Teknik Üniversitesi Telekomünikasyon Mühendisliği Bölümünden almıştır. Doktora çalışması sırasında Hollanda'nın Eindhoven kentinde bulunan Philips Araştırma Merkezinde ve Amerika Birleşik Devletlerinde bulunan Boston Üniversitesi, Bilişsel ve Sınır Sistemleri Bölümünde misafir araştırmacı olarak çalışmıştır. Fransa'nın Grenoble kentine bulunan Gipsa-Lab'da doktora sonrası araştırmacı olarak görev yapmıştır. Halen MEF Üniversitesi, Elektrik ve Elektronik Mühendisliği Bölümünde doktor öğretim üyesi olarak çalışmaktadır. Araştırma alanları arasında sayısal işaret işleme, örüntü tanıma, ses damgalama, ses kaynak ayrıştırma, matris ve tensör ayrıştırma, yüz ifadelerinden duygusu tanıma ve makine öğrenmesi yer almaktadır.