Microsoft

Tek ve Çok Kanallı Ses Kaynağı Ayırma için Derin Öğrenme

Hakan Erdoğan

Microsoft

Boğazda Yapay Öğrenme İsmail Arı Yaz Okulu 2018, 2-5 Temmuz 2018

Hakan Erdoğan Kısa Biyografi

- Microsoft'ta araştırmacı olarak çalışıyor, 2016'dan beri
- Sabancı Üniversitesi'nde öğretim üyesi, 2002-2016 arası
- Mitsubishi Electric Research Labs (MERL) (sabatik), 2014-2015 arası
- IBM TJ Watson Araştırma Merkezi, 1999-2002 arası (menejer: Michael Picheny)
- Master ve Doktora: University of Michigan, 1993-1999 arası
- Lisans: ODTÜ, 1989-1993
- Araştırma alanları: Konuşma sinyali ayırma, konuşma tanıma, sinyal işleme, biyometri

Konuşma Planı

- Derin öğrenme’ye bir bakış
- Ses ayırma problemleri tanımı
- Sinyal Gösterimi
- Klasik yöntemler
- Tek kanallı sinyaller için ayırma yöntemleri
  - Tek konuşmacı ve arkaplan gürültüsü
  - Birden fazla konuşmacıyı ayırma
- Çok kanallı sinyaller için ayırma yöntemleri
  - Uzamsal bilgi kullanımı

Derin Öğrenme

- Yapay zekanın en önemli amaçlarından: Doğru kestirim yapabilme
- Son zamanların en popüler yapay zeka konusu: derin sinir ağları
- Bir hesaplama kutusu: $y=f_W(x)$:
  - x girdi verisi, y çıktı, w kutunun parametreleri (milyonlarca)
  - kutu içinde çok katmanlı hesaplamalar içerir
- Eğitim sırasında verilen her girdi x için bir hedef/etiket değer t vardır.
- $y=f_W(x)$ değerini hedef değer t'ye yakınlaştırmak isteriz. Sinir ağının w parametrelerini ayarlayarak bunu başarabiliriz.
  - Gerçek kullanımda hedef değer bilinmez ama kutuyu kullanarak y değerini hesaplar ve hedefi kestirebiliriz
- Otomatik türev alma ile w parametrelerini güncelleyen stokastik gradyan algoritması ile eğitim yapılır
- Neden son zamanlarda ilgi arttı?
  - Daha fazla veri, daha fazla hesaplama, daha fazla insan kaynağı

Derin öğrenme ile başarı hikayeleri

- Konuşma tanıma
- İmgelerden nesne tanıma
- İmge altına tanım yazısı yazma
- Otomatik dilden dile tercüme
- Video’lardan nesne tanıma, takip etme
- AlphaGo: Go oynayan yapay zeka
- Diğer birçok alanda potansiyel: tıp, tıbbi görüntüleme, iş dünyası için akıllı çözümler, avukatlık, muhasebe işleri, otomatik soru cevaplama

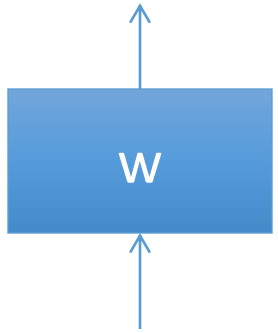
Derin öğrenme çalışmak için ne bilmeliyim?

- Programlama
  - Python (numpy ve scipy kütüphaneleri, araştırma için)
  - C ve C++ (endüstriyel son ürün için)
  - Bir derin öğrenme kütüphanesi: örneğin pytorch, tensorflow, cntk, chainer, Knet, Dynet v.s.
- Calculus (türev, integral)
  - Özellikle çok değişkenli calculus
- Geometri ve doğrusal cebir
  - Temel geometri, vektör, vektör uzayları, matrisler, tensörler, SVD
- Olasılık
  - Temel olasılık teorisi, maksimum olabilirlik, çok değişkenli dağılımları
- Optimizasyon/eniyileme
  - Temel optimizasyon bilgisi, gradyan iniş, Newton yöntemi
- Ses işleme için: Karmaşık sayılar, sinyal işleme

Sinir ağı – bir fonksiyon yakınsayıcı

(x,t) çiftleri eğitim için kullanılır

Ağ çıktısı y'nin t 'ye ne kadar yakın olduğunu ölçecek $L(f_w(x),t)$ kayıp fonksiyonu tanımlanır



$y=f_W(x)$ Ağ çıktısı

W parametrelerine sahip ağ

x Giriş verisi

+ eğitim için bir sürü (x,t) verisi gerekir

Girdi, çıktı ve hedefler

- Girdi verisi genelde bir vektör olarak ifade edilir
  - İmge'nin piksel değerleri
  - Ses sinyalinden çıkarılmış Fourier dönüşümü tabanlı öznitelikler
- Hedef değerleri:
  - Sınıflandırma problemi için sınıf numarası, bir tamsayı
    - Çoğu zaman 0/1 kodlanmış vektör halinde
  - Kestirim (bağlanım) problemleri için bir vektör ya da tensör
  - Bir vektör/tamsayı dizisi, bir vektör/tamsayı kümesi gibi farklı hedefler de tanımlanabilir
- Çıktı verisi:
  - Çıktı verisi de çoğunlukla bir vektördür ve kayıp fonksiyonu ile hedef değerine yakınsanmaya çalışılır

Kayıp fonksiyonu

$$\hat{W} = \arg\min_{W} \sum_{i} \mathcal{L}(f_W(x_i), t_i)$$

Örnek kayıp fonksiyonları

- Sınıflandırma için çapraz entropi ya da maksimum olabilirlik
- Sınıflandırma/tanılama için menteşe kaybı
- Bağlanım problemleri için ortalama karesel hata
- Diğer uygulamaya bağlı kayıp fonksiyonları
- Önemli olan: kayıp fonksiyonunun ağın çıktısına göre türevinin tanımlı ve devamlı olmasıdır

Eğitim - optimizasyon

- Sinir ağlarını eğitme problemini nasıl çözeriz?
  - Kayıp fonksiyonu stokastik gradyan iniş yöntemi ile endüşüklenir
  - Bağıl türevleri hesaplamak için geri-yayılım algoritması kullanılır
  - Stokastik gradyan iniş algoritması
    - Birçok versiyonu bulunur
    - Momentum, RMSPROP, RPROP, ADAM etc.
  - Stokastik gradyan inişte, ilk önce bir mini-yığın veri seçilir, bu veri üzerinden kayıp fonskiyonunun parametrelere göre türevleri G hesaplanır ve W:=W-mG olarak parametreler güncellenir
  - m'ye öğrenme katsayısı denilir ve SGD için önemli bir parametredir

Gri’ye övgü

- Sinir ağlarının bir özelliği:
  - Hata yapmaları, ama hata yaptıklarında dahi doğru karar verdikleri zaman kadar emin olmaları
  - Yani kendi çıktısına bakarak bir ağın verdiği kararın doğru olup olmadığını anlamak pek mümkün olmuyor
- Aynı insan beyni gibi:

Yanny mi Laurel mı?

Elbise altın rengi mi mavi mi?

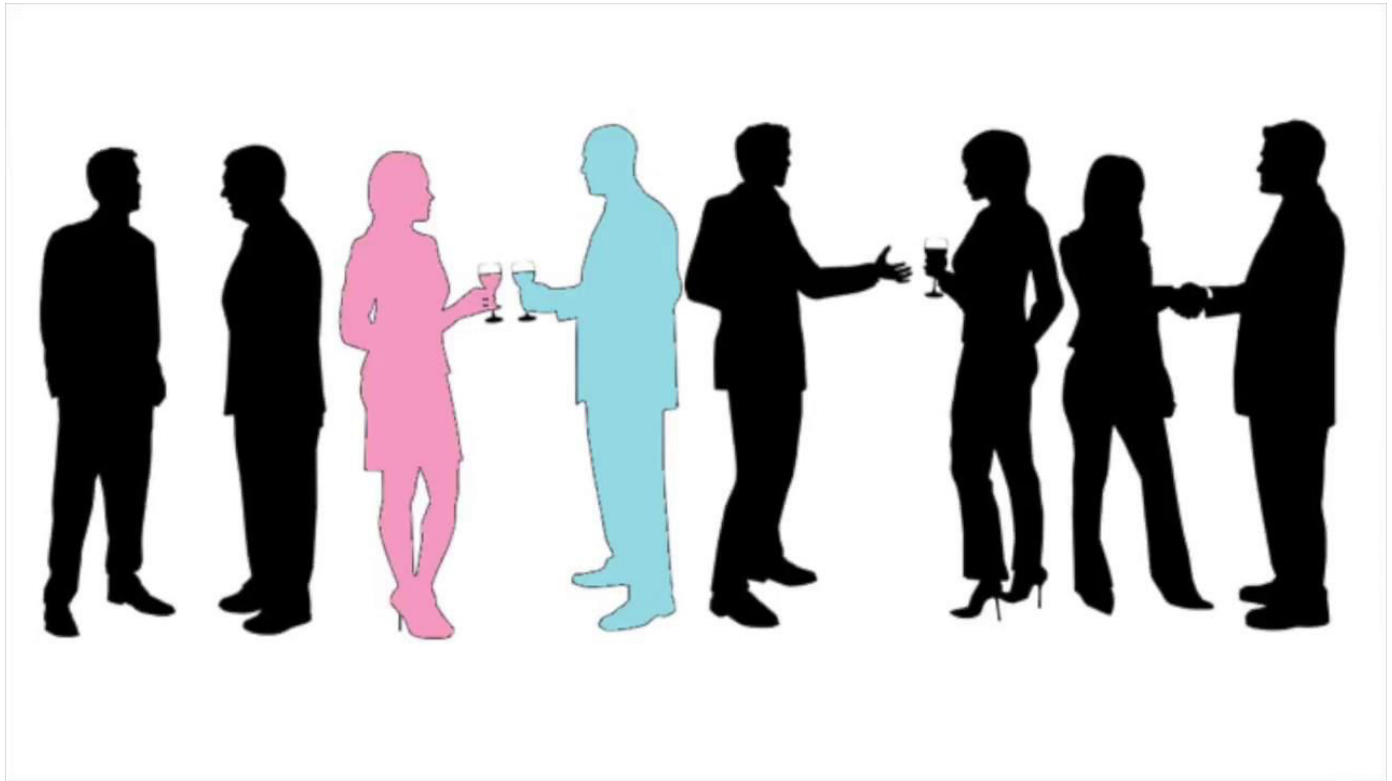
Ne yapabiliriz?

- Eğitirken yanlış payı bırakmak
  - Hedef etiket değerlerini 0-1 olarak kodlamak yerine diğer alternatiflere de düşük de olsa olasılık atanır
    - İnsana denk konuşma tanıma yapan sistemde kullanıldı
- Öğretmen-öğrenci modelleri
  - Etiket değerini 0-1 kesin hedef olarak kullanmak yerine iyi eğitilmiş başka bir modelin çıktısını yumuşak-hedef olarak kullanmak
  - Konuşma tanımada iyi eğitilmiş temiz ses modellerinin çıktılarını gürültülü ses modeli eğitmek için yumuşak hedef olarak kullanmak
- Daha doğru karar vermek için
  - Siyah/beyaz değil gri!

Öğretmen-öğrenci modelinin konuşma tanımada bir uygulaması

- Çok iyi eğitilmiş temiz konuşma modelimiz var
  - Temiz ve yakın mikrofonla konuşmaları çok iyi tanıyor
  - Bu modelleri yüzbinlerce saat veriden eğitmek bazen aylar alıyor (Birden fazla GPU kullanarak)
- Gürültülü ve uzak mikrofon kullandığımızda da iyi çalışan bir model (gürbüz model) elde etmek istiyoruz
  - Temiz ses modeli uzak mikrofonla kayıt yapılmışsa ve gürültü varken iyi çalışmıyor
- Temiz sesi alıyoruz ve temiz konuşma modelinden geçiriyoruz, sonra temiz sesi uzak mikrofon etkisi ile benzetim yapıyor ve gürültü ekliyoruz
- Gürbüz modele girdi olarak benzetilmiş uzak mikrofon ve gürültü eklenmiş sesi veriyoruz ve temiz modelin çıktısı ile aynı çıktıyı vermesini istiyoruz
  - Temiz model öğretmen, gürbüz model öğrenci oluyor
- Bu durumda elimizdeki temiz sesin yazılandırması olmadan da eğitim yapabiliyoruz
- Bu tarz eğitim Microsoft şirketi için oldukça başarılı sonuçlar verdi ve uzak mikrofon modelinde kullanıldı [Li et.al 2018]

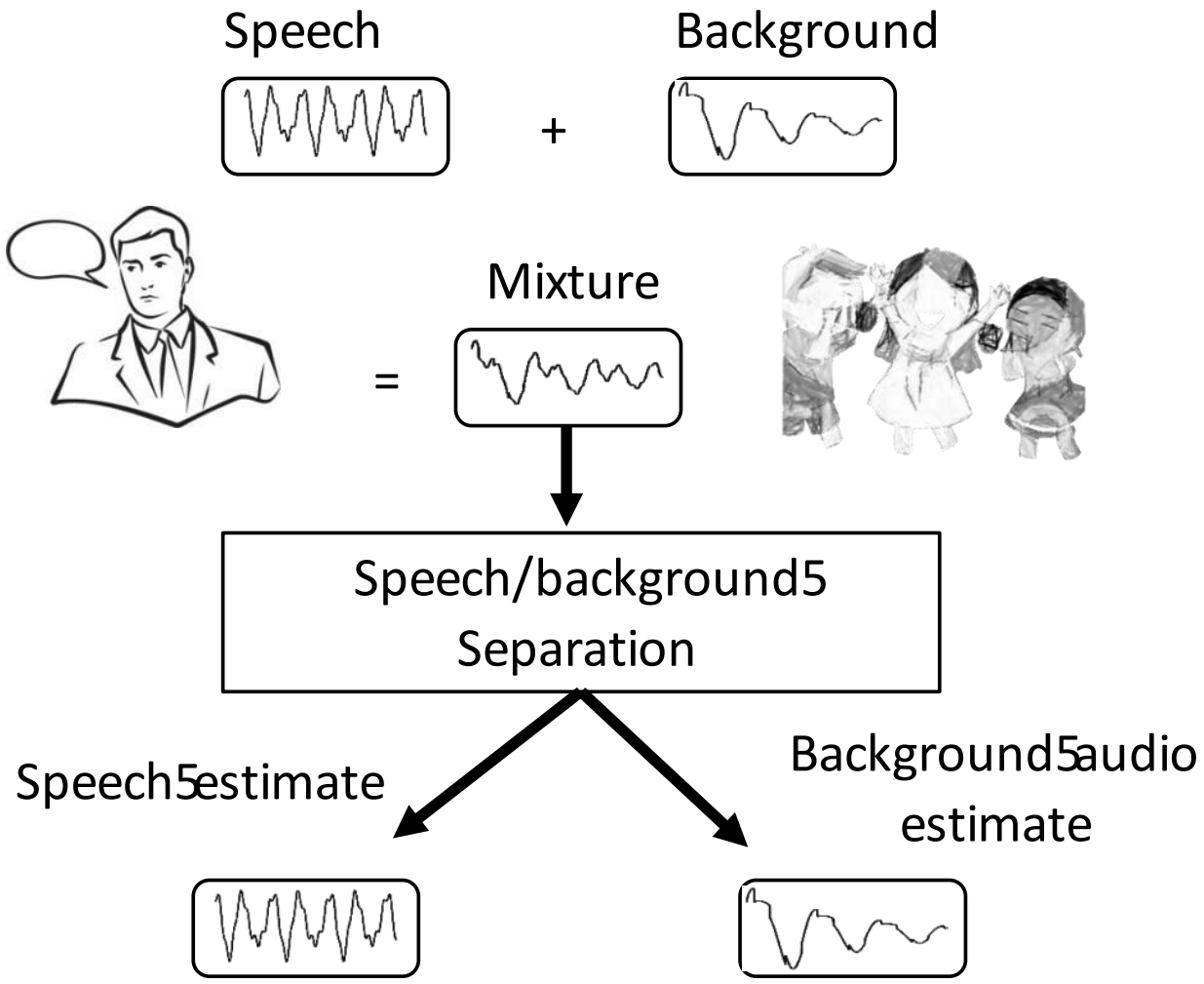
Kokteyl partisi problemi



- İki kulağımız olması nedeniyle mi?
- Sesin yön tayini ve sese odaklanma yeteneği

Ses ayırma problemleri

- Kokteyl parti problemini detaylandıralım
  - Uygulama senaryosuna göre farklı amaçlar olabilir
- Tek (dominant) konuşmacının sesini arka plan seslerinden ayırmak
  - Konuşma iyileştirme
  - Konuşmayı gürültüden arındırmak
  - Konuşma ya da şarkı söyleme ile müziği ayırmak
- Birden fazla konuşmacının konuşmalarını birbirinden ayırmak
  - Aynı anda konuşanlar ya da konuşmalarının bir kısmı üstüste gelen konuşmacılar
  - Birden fazla konuşmacı arasından istenen tek bir konuşmacının sesini çıkarmak
  - İş toplantıları gerçekçi bir senaryo olabilir


Speech
Background
+
Mixture
=
Speech/background5
Separation
Speech5estimate
Background5audio
estimate

Sensör tipine göre

- Tek kanallı:
  - Kayıt için tek mikrofon kullanıldığında
  - Sadece spektro-temporal bilgi (zaman-frekans bilgisi) kullanılabilir
- Çok kanallı:
  - Kayıt için bir mikrofon dizisi kullanıldığında (birden fazla mikrofon)
  - Mikrofonlar tek bir cihaz üzerinde olabilir (örneğin Alexa, Google home, invoke)
  - Ya da farklı cihazlar üzerinde olabilir – senkron problemi
  - Hem spektro-temporal hem de uzamsal bilgi kullanılabilir
  - Mikrofon dizileri ile alınan sinyallere bakarak sesin hangi yönden geldiği kestirilebilir

Sinyal gösterimi

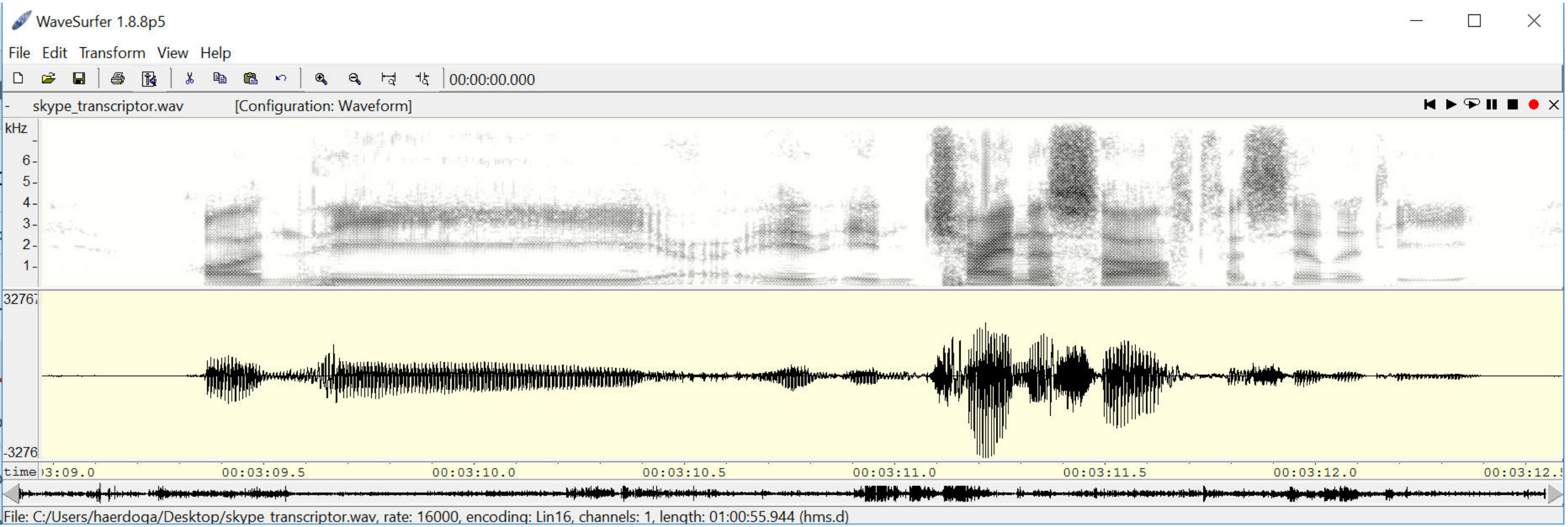
- Ses sinyalleri tek boyutlu sinyallerdir. Zamana bağlı olarak değişen sinyalin seviyesi havadaki ses dalgasının basınç değeri ile orantılıdır. Mikrofon, mekanik titreşimi (basınç) elektrik dalgasına çevirir ve sonrasında zamanda örneklenerek kesik zamanlı sinyal elde edilir.

$$x[n]$$

- Spektro-temporal gösterim: kısa-zamanlı Fourier dönüşümü ve spektrogram
  - Sinyalin belli bir pencere içerisinde Fourier dönüşümü alınır ve analiz çerçevesi kaydırılarak bu işlem tekrar edilir:

$$X(t,f) = \sum_{n=0}^{N-1} x[n+tL] w_a[n] e^{-j2\pi nf/N}$$

Sinyal gösterimi: spektrogram=STFT genliği



Tek kanallı konuşmayı arka plandan ayırma

$$y[n] = s[n] + v[n]$$

- **y, s, n**= karışmış sinyal, konuşma sinyali, gürültü sinyali
- Kısa zamanlı Fourier dönüşümü (STFT) uzayında

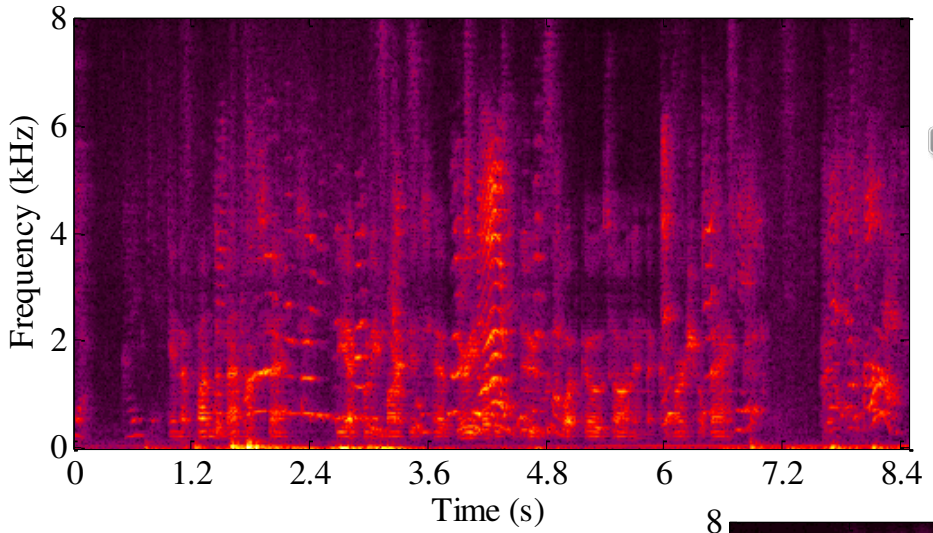
$$Y(t,f) = S(t,f) + V(t,f)$$

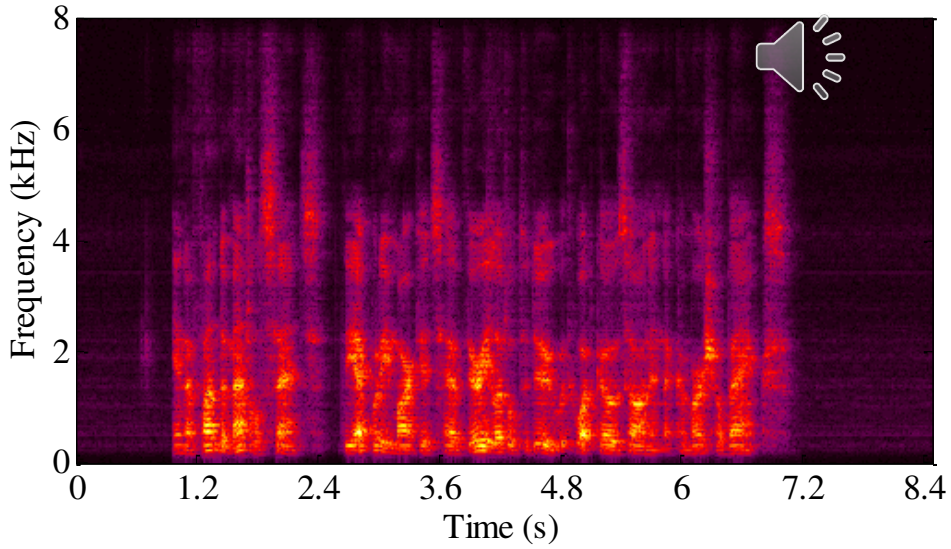
- Çoğu zaman yapılan varsayım:

$$|Y(t,f)| \approx |S(t,f)| + |V(t,f)|$$

- Problem: Karışmış sinyalin STFT'si **Y** verilmişken, konuşma sinyali **S**'nin kestirimi
- Elimizde eğitim verisi olarak temiz konuşma ve gürültü sinyallerinden elde edilmiş karışımlar var

Gürültülü ve temiz konuşma örneği



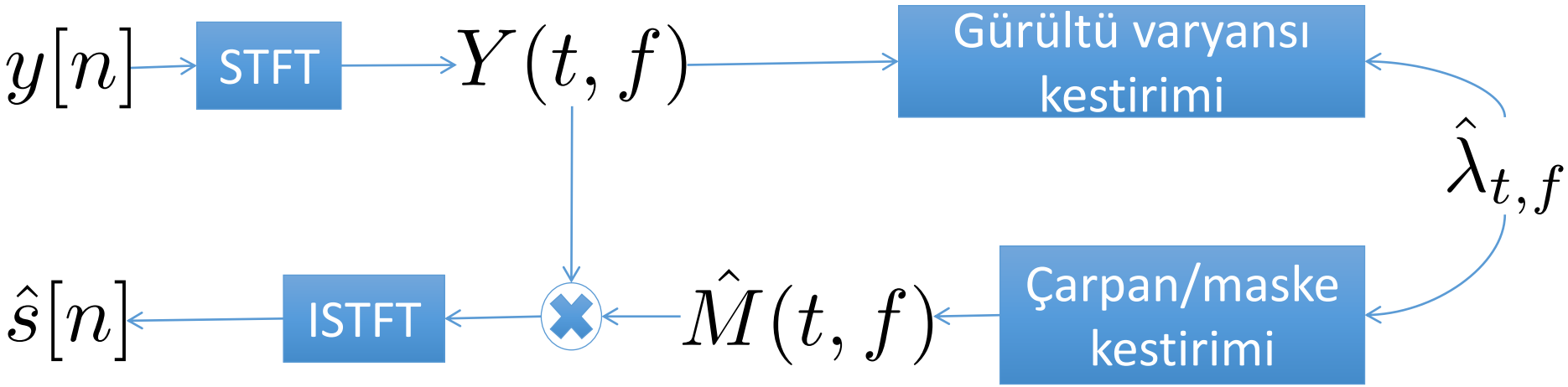


Genlik ve faz

$$s = |s| \exp\{j\theta_s\}$$

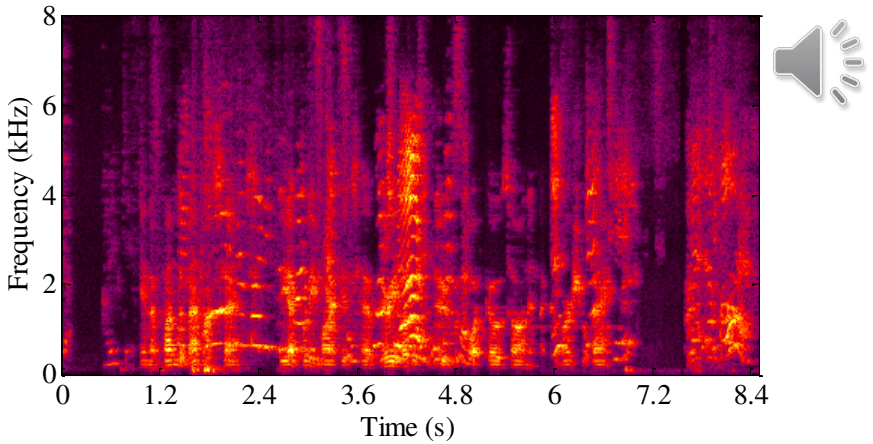
- Genlik $|s|$ ve fazın θ_s ikisini de mi kestirelim?
- Bazı varsayımlar altında θ_s fazının en iyi kestirimi karışık sinyalin fazıdır θ_y [Ephraim&Malah 1984, Cohen&Berdugo 2001]
- Sadece genliğin kestirimi yeterli performans sağlar
- Faz kestirimi sonraki çalışmalara bırakılmıştır

Klasik yöntemler



- Gürültünün durağan olduğu varsayılır ve gürültü varyans parametresi konuşma olmayan bölümlerden tahmin edilir
- Maske parametresi 0 ve 1 arasındadır.

Klasik yöntemlerden OMLSA algoritmasının performansı



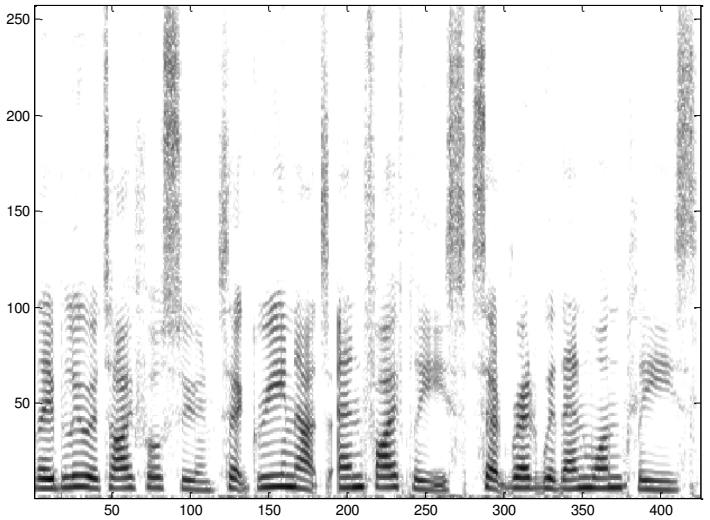
- OMLSA algoritması çıktısı
- Klasik yöntemler iyi çalışmıyor çünkü gürültü durağan değil

Otomatik öğrenme yöntemleri

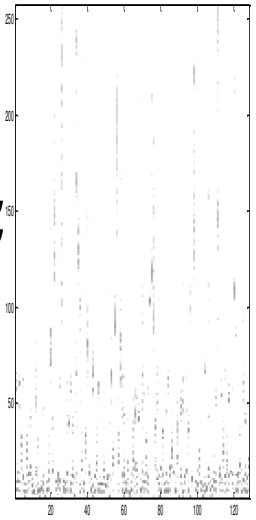
- Ses iyileştirme yerine genelde kaynak ayrıştırma terimi kullanılır
- Büyük veri birçok zaman elimizde var. Eldeki farklı ses kaynaklarını karıştırabiliriz. Otomatik öğrenme kullanabiliriz.
- Otomatik öğrenme yöntemleri
  - Model tabanlı
    - Negatif olmayan matris ayrıştırma (NMF) ve benzer teknikler
    - Diğer olasılıksal teknikler
  - Sinir ağları
    - Çok katmanlı perseptron (MLP) ya da diğer ismi ile (ileri beslemeli) derin sinir ağı
    - Yinelgen sinir ağları (RNN)
    - Uzun kısa-zamanlı-bellek yinelgen ağlar (LSTM-RNN)

Negatif olmayan matris ayrıştırma (NMF)

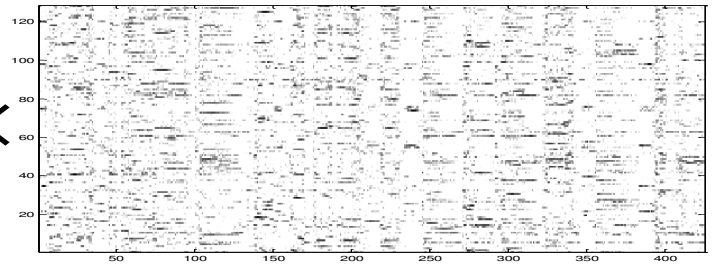
Spectrogram (V)



≈ Dictionary (B)



× Gains (G)



$$\text{V matrisi verildiğinde } B, G > 0 \text{ olmak üzere}$$

$$\arg\min_{B,G} D(V|BG) \text{ probleminin çözümü}$$

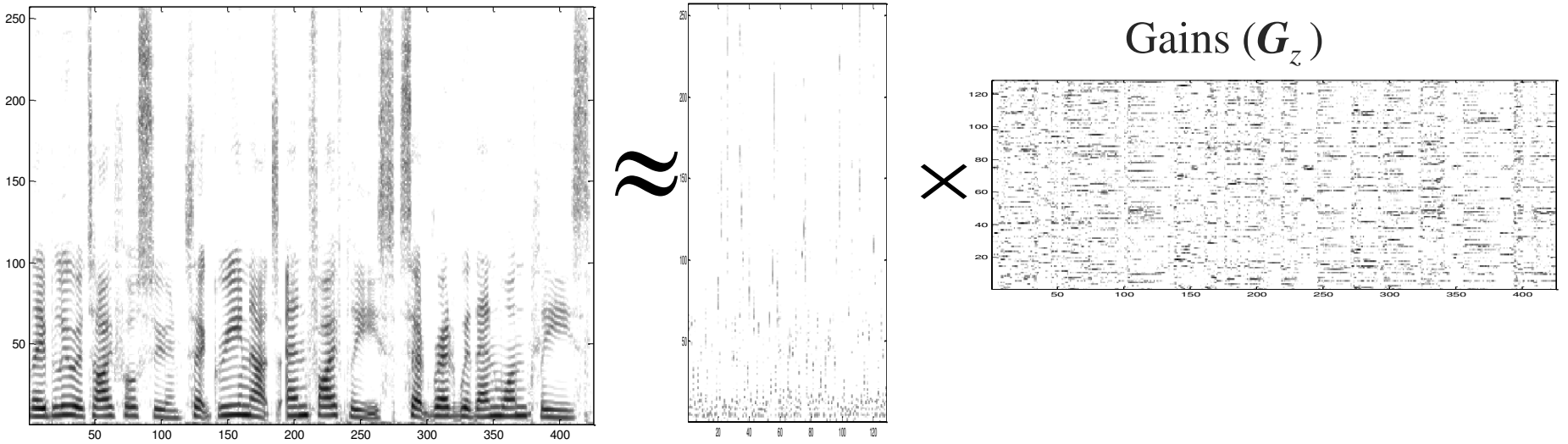
[Lee&Seung 2001, Smaragdis&Brown 2003]

NMF ile kaynak modelleme

Spectrogram ($\boldsymbol{S}_z$)

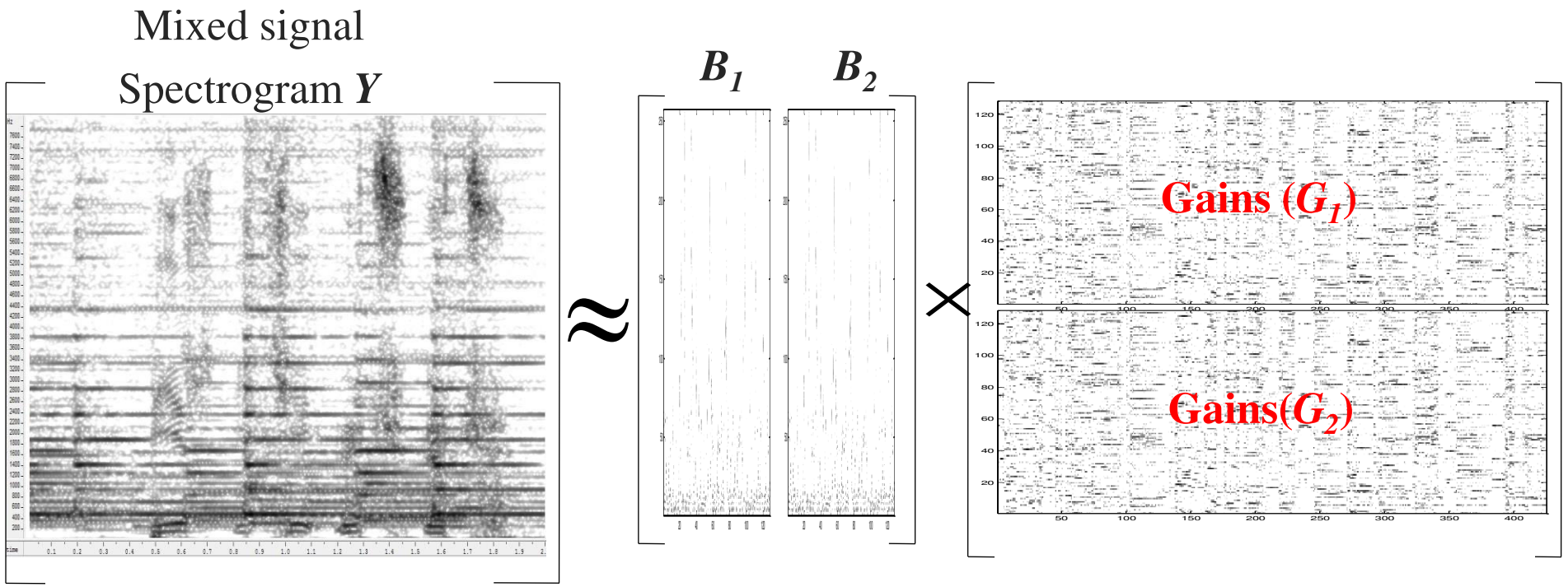
Dictionary matrix ($\boldsymbol{B}_z$)

Gains ($\boldsymbol{G}_z$)



- Eğitilen sözlük matrisi (dictionary) kaynak sinyalinin özeti olarak kullanılır [Schmidt&Olsson 2006, Virtanen&Cemgil&Godsill 2008, Grais&Erdogan 2011]

NMF ile kaynak ayırma



$$\tilde{s}_1 = B_1 G_1 \qquad \tilde{s}_2 = B_2 G_2$$

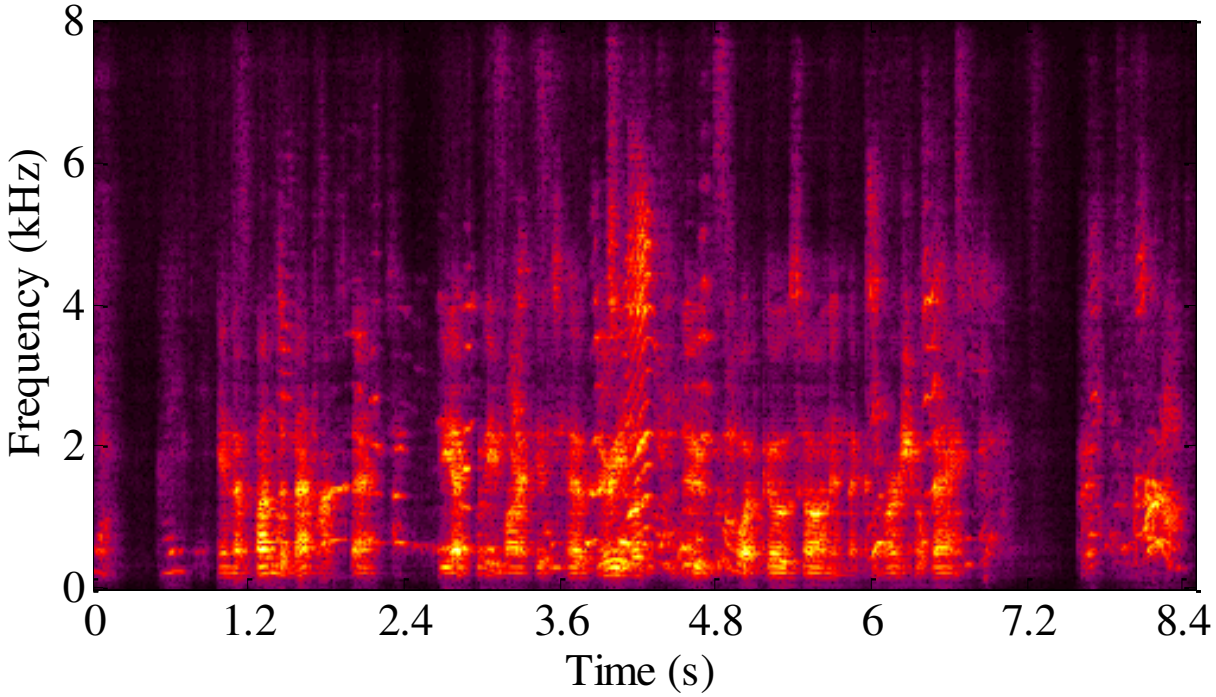
NMF Maske kestirimi

- Her kaynak için kestirilen spektra kullanılarak bir maske kestirilebilir [Grais&Erdogan 2011].

$$\hat{m} = \frac{(\tilde{s}_1)^p}{(\tilde{s}_1)^p + (\tilde{s}_2)^p}$$

- Farklı p değerleri maskenin keskinliğini etkiler
- p>1 ise dominant kaynak daha da etkinleşir.
- p sonsuza gittiğinde 0-1 maskesi elde edilir

NMF başarımı

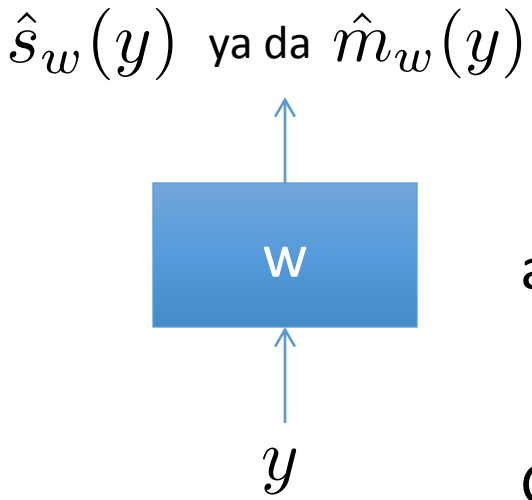


NMF yöntemi sonucu

Derin sinir ağları ile ses temizleme ve kaynak ayırma

- Sinir ağları ile maske kestirimi yapılabilir ya da spektral sinyal kestirilebilir
- **İkili** maske (0 ya da 1) kestirimi yapılırsa problem her spektro-temporal katsayı için ikili bir sınıflandırma problemine döner [Wang&Wang 2013]
- **Yumuşak maske** kestirimi ya da direkt kaynak sinyalin spektral gösteriminin kestirimi ise bir bağlanım problemidir [Xu&Du&Dai&Lee 2014, Huang&Kim&Johnson&Smaragdis 2014, Weninger&Hershey&LeRoux&Schuller 2014, Wang&Narayanan&Wang 2014]

Konuşma iyileştirme ve kaynak ayırma için derin öğrenme kullanımı



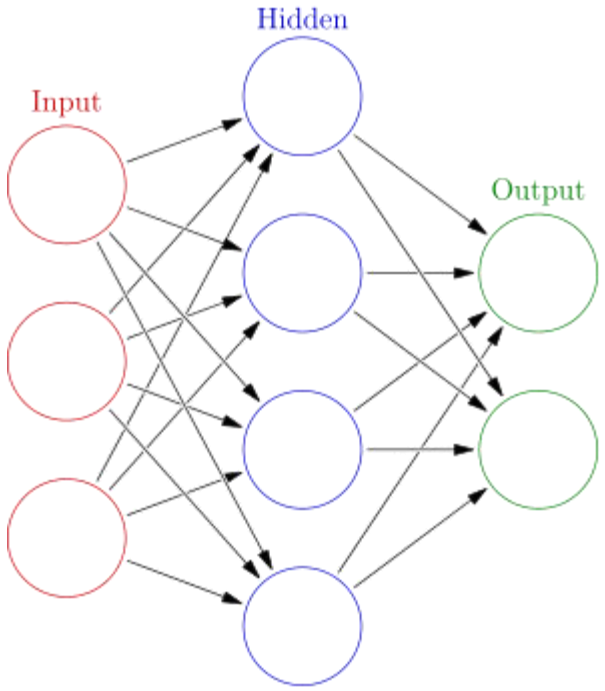
Kestirilmiş sinyal ya da maske

ağ

Gürültülü veri

+ birçok veri ile eğitim

Sinir ağlarının içine hızlıca bakalım



Wikipedia’dan alıntı

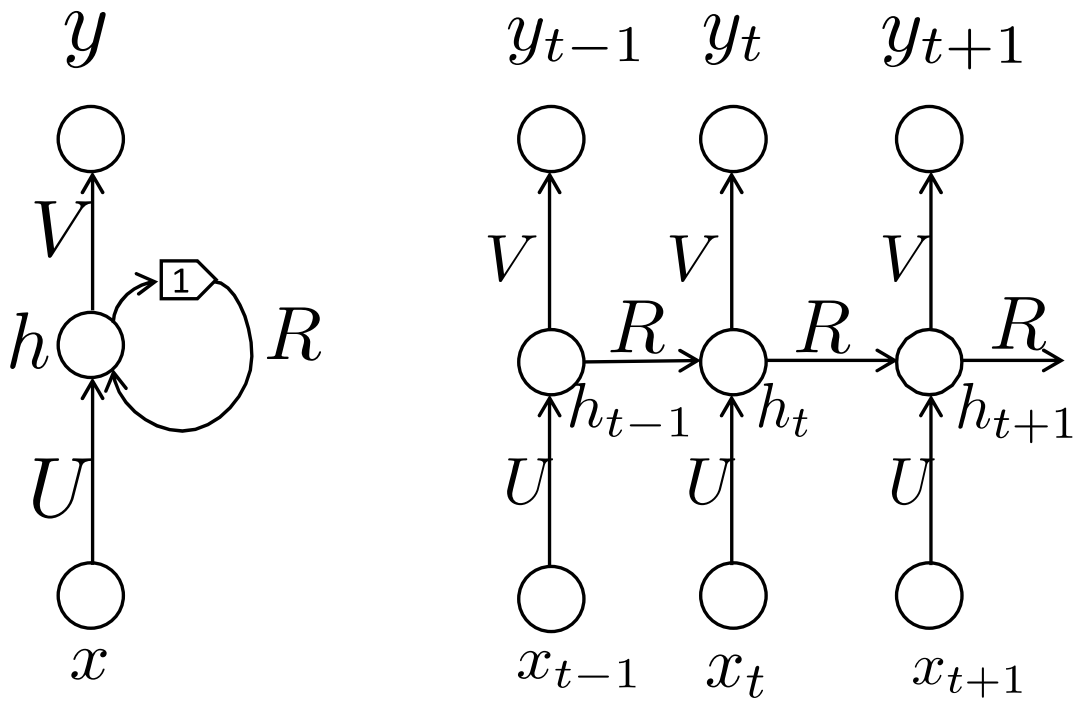
İleri beslemeli sinir ağı, tek katman

$$\mathbf{h} = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$

- Afin bir dönüşüm: Doğrusal dönüşüm artı kaydırma
- Sonrasında doğrusal olmayan bir koordinatsal dönüşüm
  - Sigmoid fonksiyonu
  - Tanh
  - Rektifiye doğrusal (RELU)
  - Diğerleri

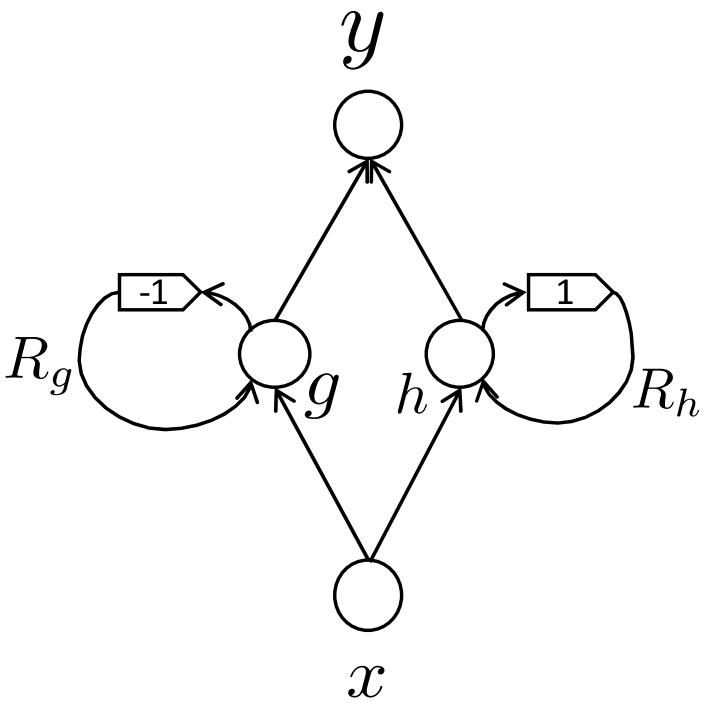
Derin öğrenme için daha fazla katman gerekir ve h vektörel değeri benzeri katmanlardan geçirilir (MLP)

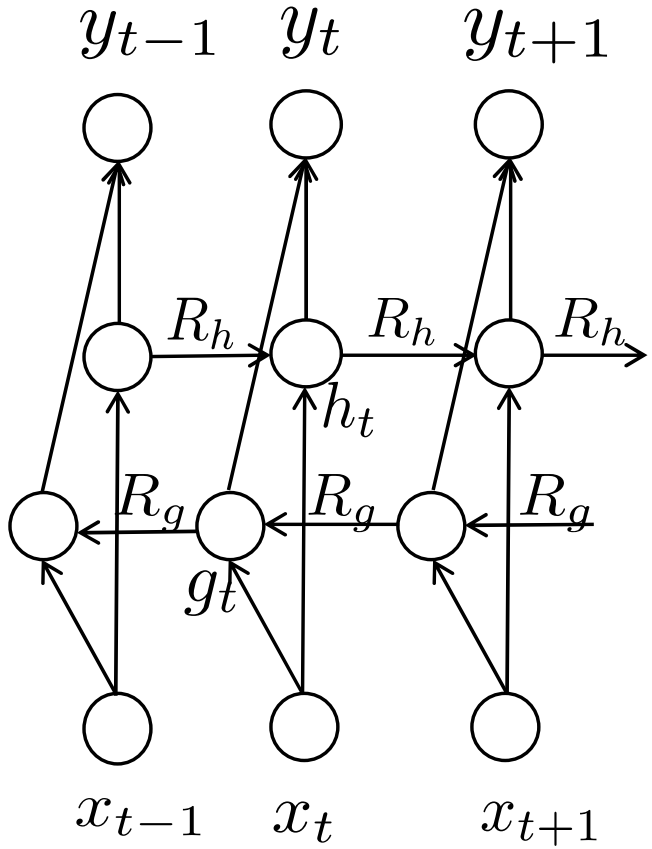
Dizisel veri için yinelgen sinir ağları (RNN)



$$h_t = \sigma(Rh_{t-1} + Ux_t)$$
$$y_t = \sigma(Vh_t)$$

İki yönlü RNN



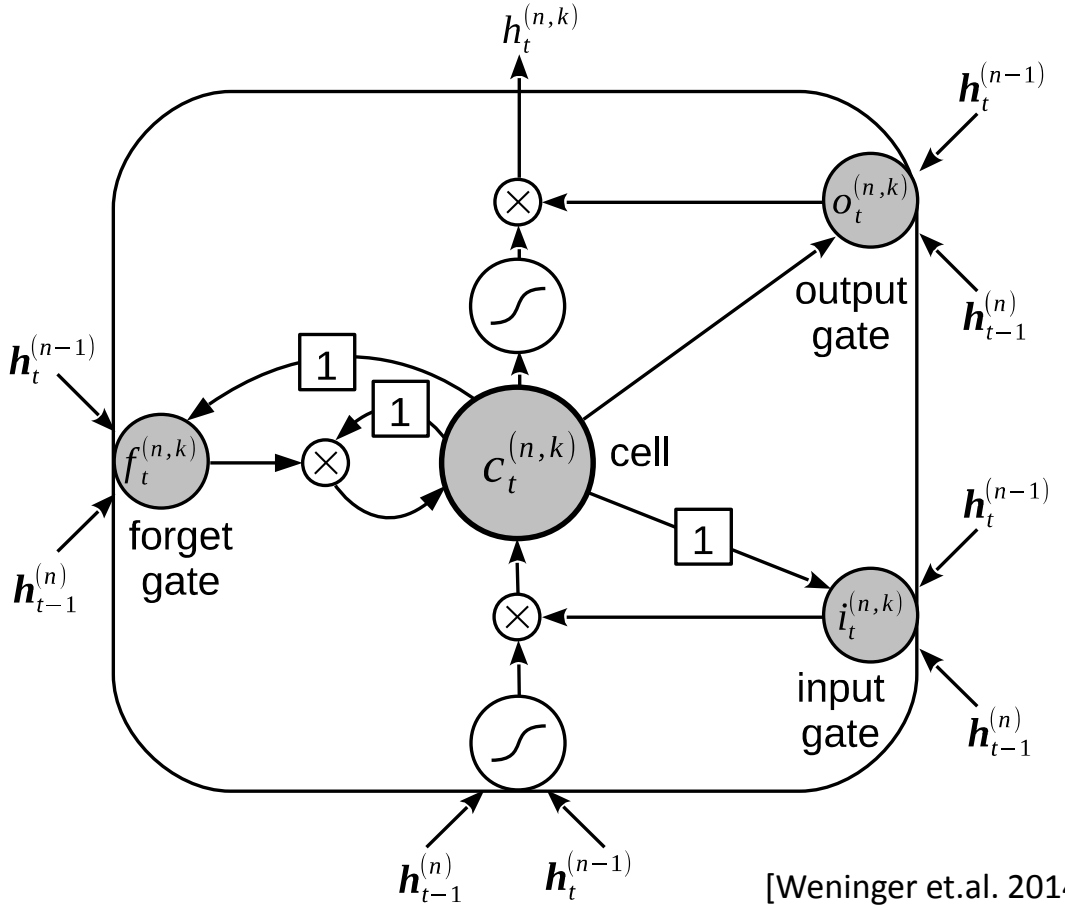


$$
\begin{aligned}
h_t &= \sigma(R_h h_{t-1} + U_h x_t) \\
g_t &= \sigma(R_g g_{t+1} + U_g x_t) \\
y_t &= \sigma(V_h h_t + V_g g_t)
\end{aligned}
$$

RNN’ler neden/nasıl çalışır

- Tarihsel (ve gelecekteki) verileri hatırlayabilirler ve t anında geçmişe ve geleceğe bakarak karar verebilirler
- Bunu teorik olarak yapmaları beklenirken, RNN’leri eğitmek pratikte zordur
- Pratikte eğitim için kullanılan geri-yayılım algoritması ile elde edilen gradyanlar zaman ilerledikçe hızla yükselir ya da hızla azalır
- Bu da RNN’lerin eğitilmesini zorlaştırır
- Bu sorunlardan etkilenmeyen uzun kısa-süreli bellek (LSTM) yinelemeli yapısı geliştirilmiştir [Hochreiter&Schmidhuber 1997]

Uzun kısa-süreli-bellek (LSTM) hafıza hücresi



[Weninger et.al. 2014] alıntı.

LSTM Hesaplamaları

$$
\begin{aligned}
\mathbf{h}_0^{(1,\ldots,N)} &:= \mathbf{0}, \mathbf{c}_0^{(1,\ldots,N)} := \mathbf{0}, \\
\mathbf{h}_t^{(0)} &:= \tilde{\mathbf{x}}_t, \\
\mathbf{f}_t^{(n)} &:= \sigma(\mathbf{W}^{f,(n)}[\mathbf{h}_t^{(n-1)}; \mathbf{h}_{t-1}^{(n)}; \mathbf{c}_{t-1}^{(n)}; 1]) \\
\mathbf{i}_t^{(n)} &:= \sigma(\mathbf{W}^{i,(n)}[\mathbf{h}_t^{(n-1)}; \mathbf{h}_{t-1}^{(n)}; \mathbf{c}_{t-1}^{(n)}; 1]) \\
\mathbf{c}_t^{(n)} &:= \mathbf{f}_t^{(n)} \otimes \mathbf{c}_{t-1}^{(n)} \\
&+ \mathbf{i}_t^{(n)} \otimes \tanh\left(\mathbf{W}^{c,(n)}[\mathbf{h}_t^{(n-1)}; \mathbf{h}_{t-1}^{(n)}; 1]\right), \\
\mathbf{o}_t^{(n)} &:= \sigma(\mathbf{W}^{o,(n)}[\mathbf{h}_t^{(n-1)}; \mathbf{h}_{t-1}^{(n)}; \mathbf{c}_t^{(n)}; 1]) \\
\mathbf{h}_t^{(n)} &:= \mathbf{o}_t^{(n)} \otimes \tanh(\mathbf{c}_t^{(n)}), \\
\tilde{\mathbf{y}}_t &:= \mathbf{W}^{(N+1)}[\mathbf{h}_t^{(N)}; 1].
\end{aligned}
$$

Ağın girdisi ne olmalı?

- Literatürde karışmış sinyalin spektrogramından türemiş öznitelikler girdi olarak denenmiştir
- Log-mel-filtre-bankası öznitelikleri iyi sonuçlar vermiştir
- Konuşma tanıma için 40 mel-filtre katsayısı kullanılırken, ayırma problemi için 100 katsayı daha iyi sonuç vermiştir [Weninger&Hershey&LeRoux&Schuller 2014]
- MLP/DNN için her çerçevenin komşu çerçevelerinden gelen öznitelikler ardarda eklenerek bağlam bilgisi eklenmiş olur.
- RNN/LSTM için ise her çerçevenin kendisi yeterlidir çünkü yinelgen ağlar dizayn itibarıyla bağlamı dikkate alır

Kayıp fonksiyonu ne olabilir?

- Literatürde denenen kayıp fonksiyonlarının hemen hepsini şu şekilde yazabiliriz

$$\min_{w} \sum_{t,f} D(\hat{s}_w(y), |s|)$$

- Karesel hata (çıktı sinyalin SNR'ı ile ilgili) [Huang&Kim&Johnson&Smaragdis 2014, Weninger&Hershey&LeRoux&Schuller 2014]

$$D(\hat{s}_w(y), |s|) = (\hat{s}_w(y) - |s|)^2$$

- Log-spektral mesafe (LSD): log-spektra kestirilir ve gerçek log-spektra ile farkı alınır [Xu&Du&Dai&Lee 2014]

$$D(\hat{l}_w(y), |s|) = (\hat{l}_w(y) - \log |s|)^2$$

Maske mi yoksa spektrum mu?

- Ağ neyi kestirmeli?
- Spektrumu kestirirse: (**SP**)

$$\min_w \sum_{t,f} (\hat{s}_w(y) - |s|)^2$$

- Maskeyi kestirirse (**MA**)

$$\min_w \sum_{t,f} (\hat{a}_w(y) - a^*)^2$$

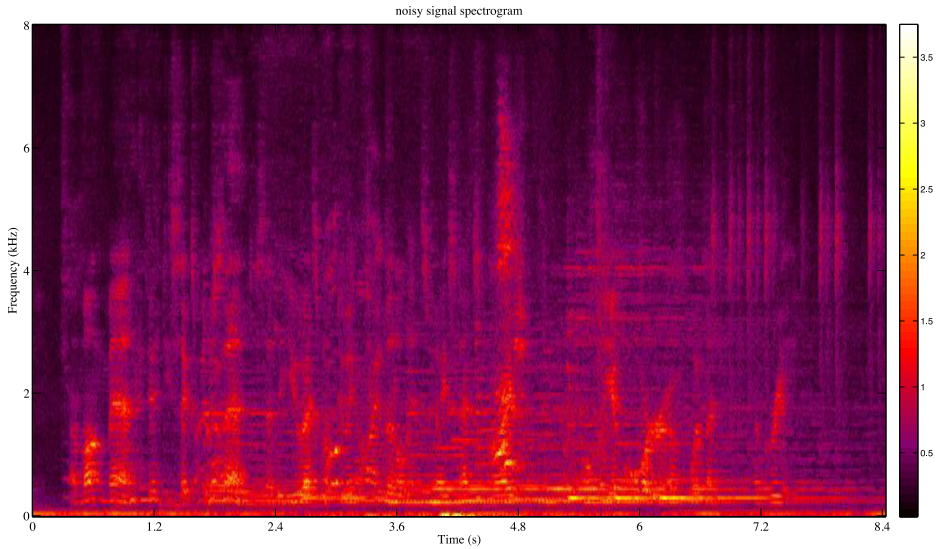
- Maskeyi kestirip hatayı spektrum'da tanımlarsak (**MSA**)

$$\min_w \sum_{t,f} (\hat{a}_w(y)|y| - |s|)^2$$

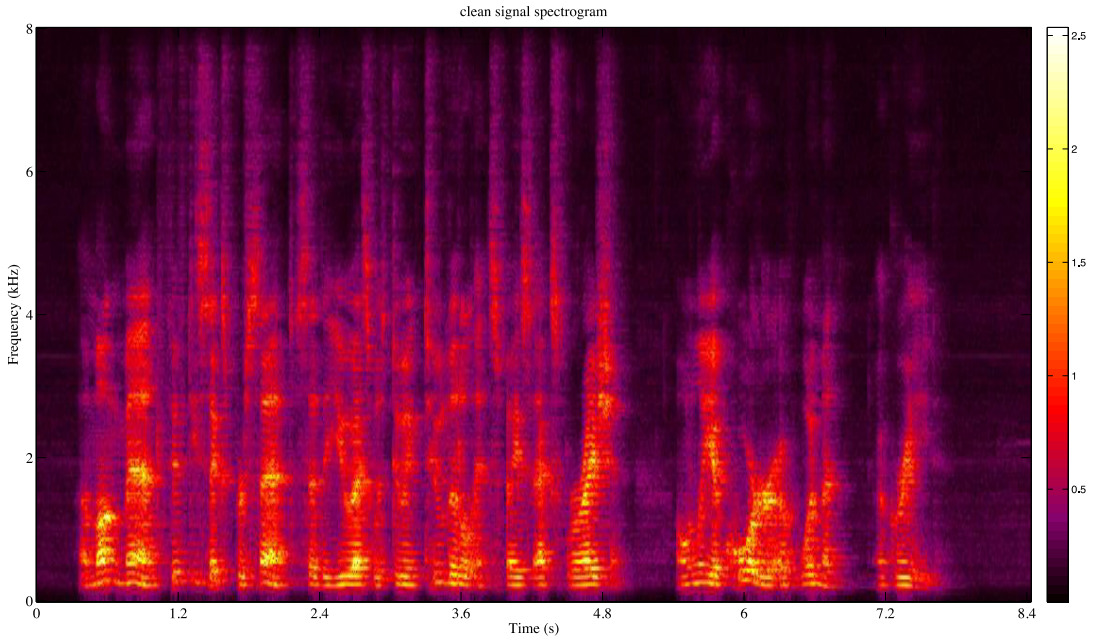
[Weninger&Hershey&LeRoux&Schuller 2014] found MSA is better than MA,
[Wang&Narayanan&Wang 2014] found MSA is better than SP

Neden maskeyi kestirmeli?

- Maske değeri [0,1] arasındadır ve çıktı katmanında sigmoid fonksiyonu kullanılabilir
- Öte yandan direkt spektra kestirimi için sigmoid kullanamayız, doğrusal ya da rektifiye doğrusal çıktı fonksiyonu gerekir
- Sinyale gürültü eklenmemişse, maskenin değeri 1 olacaktır, ve girdi sinyalini çıktıda tekrarlamaya gerek kalmayacak ve daha kolayca 1 değeri kestirilebilecektir

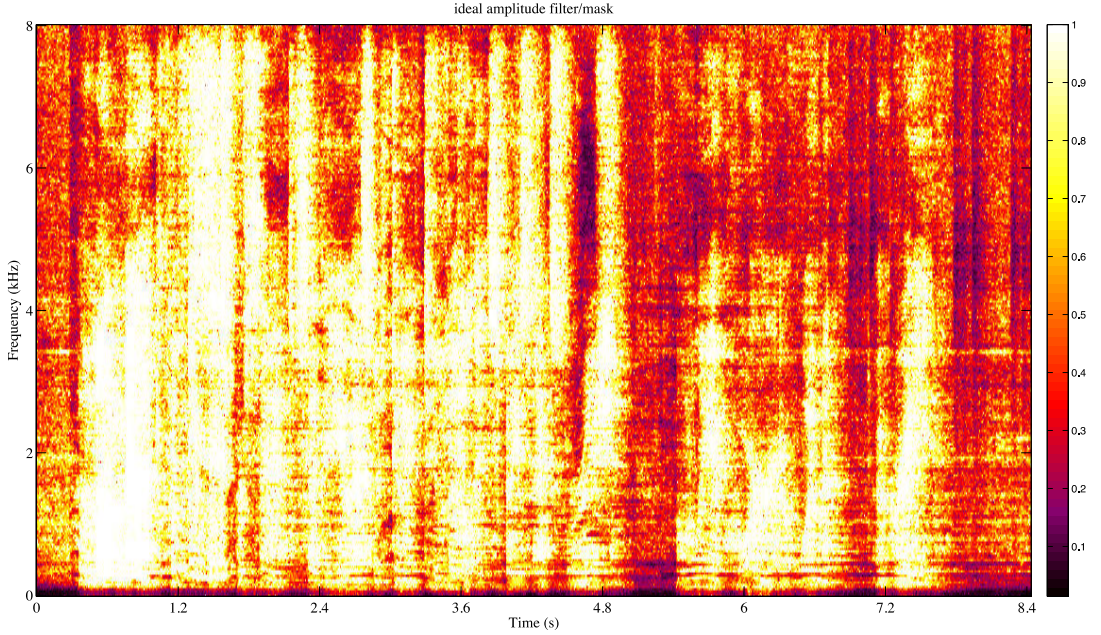


gürültülü



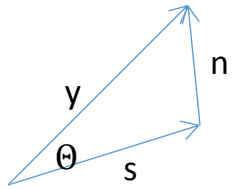
temiz

Ideal maske [0,1] arasi



Hangi ideal maske?

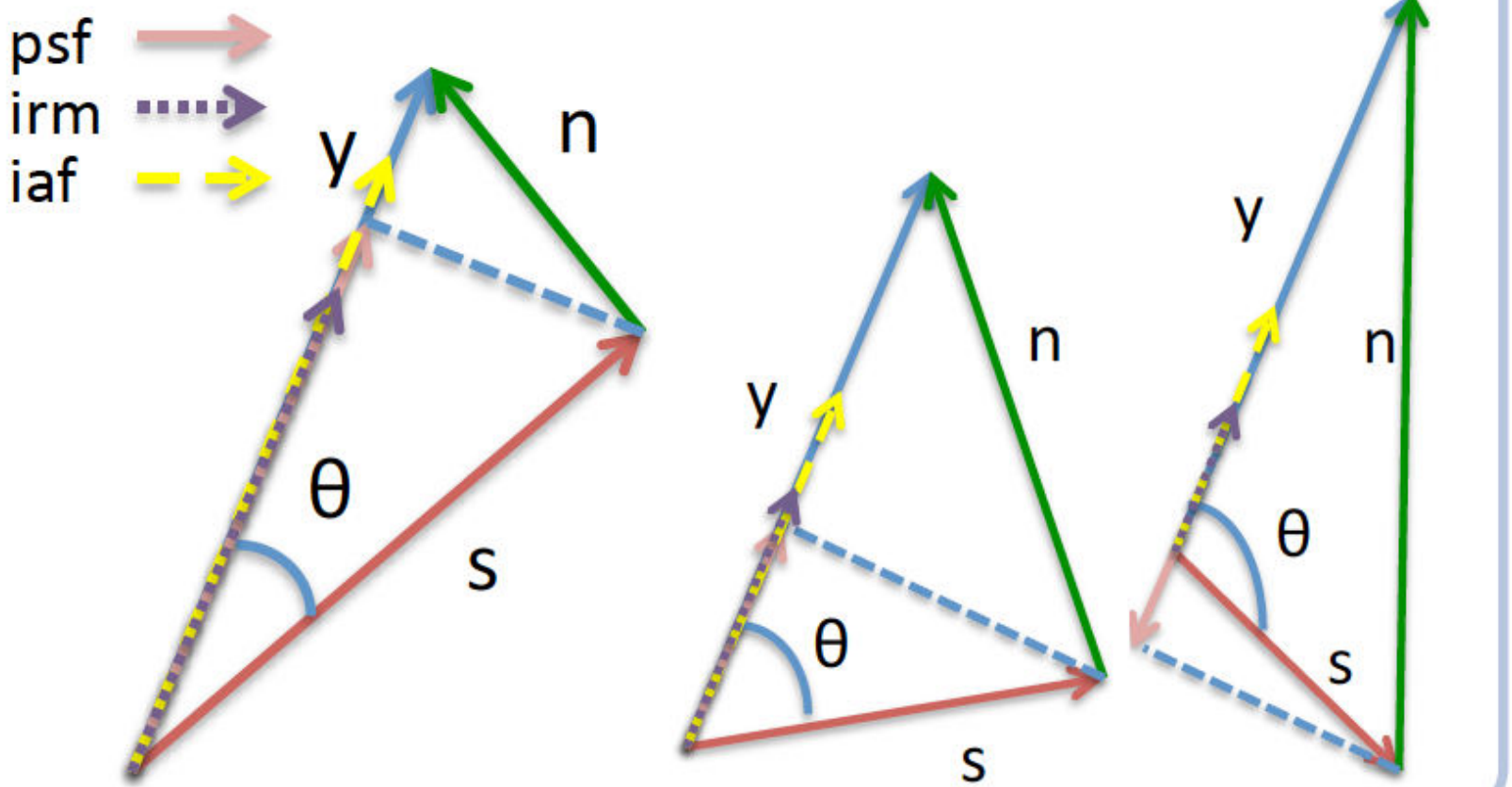
$$y = s + n \quad |y| \approx |s| + |n|$$



IDEAL MASKS

target mask/filter	formula	optimality principle
IBM:	$a^{\text{ibm}} = \delta(\lvert s\rvert > \lvert n\rvert)$,	max SNR given $a \in \{0, 1\}$
IRM:	$a^{\text{irm}} = \frac{\lvert s\rvert}{\lvert s\rvert + \lvert n\rvert}$,	max SNR given $\theta_{\text{s}} = \theta_{\text{n}}$
"Wiener like":	$a^{\text{wf}} = \frac{\lvert s\rvert^2}{\lvert s\rvert^2 + \lvert n\rvert^2}$,	max SNR, expected power
ideal amplitude:	$a^{\text{iaf}} = \lvert s\rvert / \lvert y\rvert$,	exact $\lvert\hat{s}\rvert$, max SNR $\theta_{\text{s}} = \theta_{\text{y}}$
phase-sensitive:	$a^{\text{psf}} = \frac{\lvert s\rvert}{\lvert y\rvert}\cos(\theta)$,	max SNR given $a \in \mathbb{R}$
ideal complex:	$a^{\text{icf}} = s/y$,	max SNR given $a \in \mathbb{C}$

IDEAL MASKS IN COMPLEX DOMAIN



İdeal maske başarımları

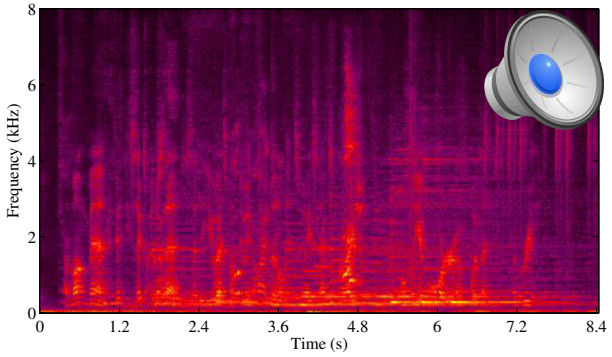
IDEAL MASKS CHiME-2 DEV SET SDR (IN dB)

dt	-6 dB	9 dB	Avg
IBM	14.56	20.89	17.59
IRM	14.13	20.69	17.29
"Wiener-like"	15.20	21.49	18.21
ideal amplitude	13.97	21.35	17.52
phase sensitive filter	17.74	24.09	**20.76**
truncated PSF	16.13	22.49	**19.17**

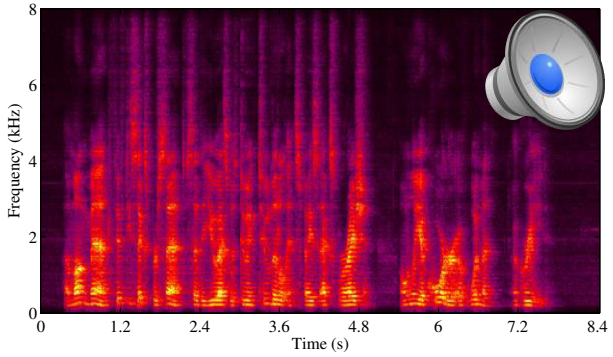
Faza duyarlı maske (Phase-sensitive filter (PSF)) [0,1] arasında sınırlandırılsa bile daha iyi sonuç veriyor

Spectrograms obtained using oracle masks

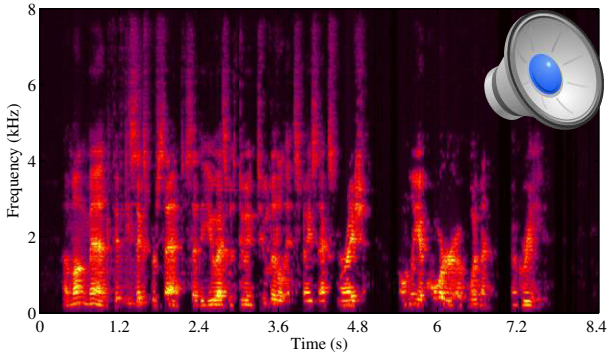
noisy



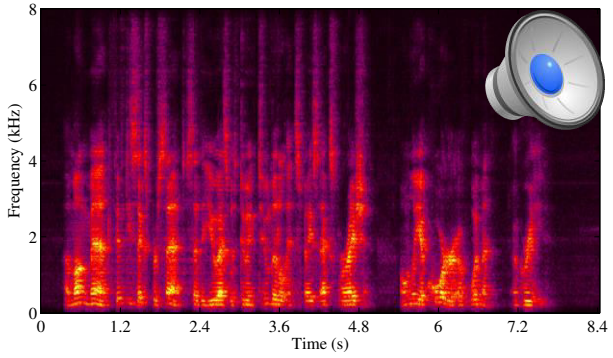
clean



IBM [SDR=8.95 dB]



PSF [SDR=11.86 dB]



PHASE-SENSITIVE APPROXIMATION LOSS

Loss function for training the network:

$$\mathcal{L}(W) = \sum_{t,f} D(\hat{a}_{t,f})$$

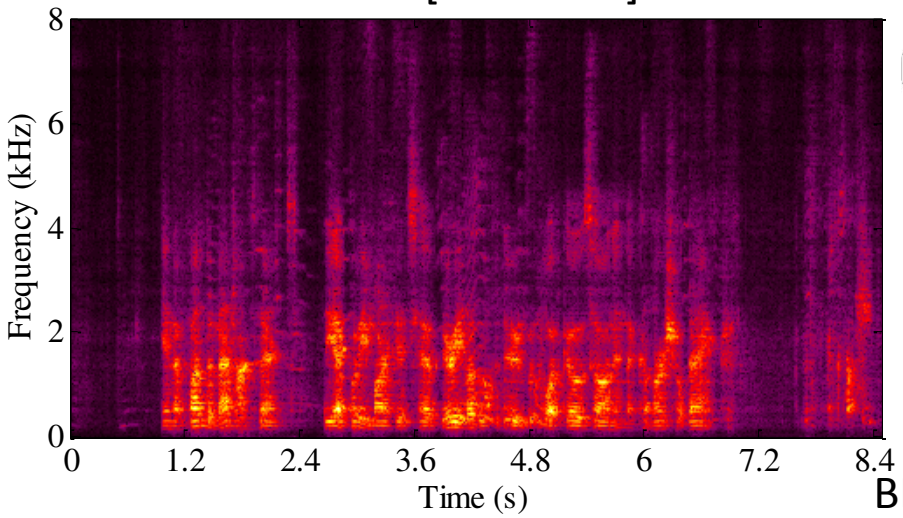
Distortion measures:

- Mask approximation (**MA**): $D_{\text{ma}}(\hat{a}) = |\hat{a} - a^*|^2$
- Magnitude spectrum approximation (**MSA**): $D_{\text{msa}}(\hat{a}) = (\hat{a}|y| - |s|)^2$
- Phase-sensitive spectrum approximation (**PSA**): $D_{\text{psa}}(\hat{a}) = |\hat{a}y - s|^2$
- **PSA** is equivalent to: $D_{\text{psa}}(\hat{a}) = (\hat{a}|y| - |s|\cos(\theta))^2$

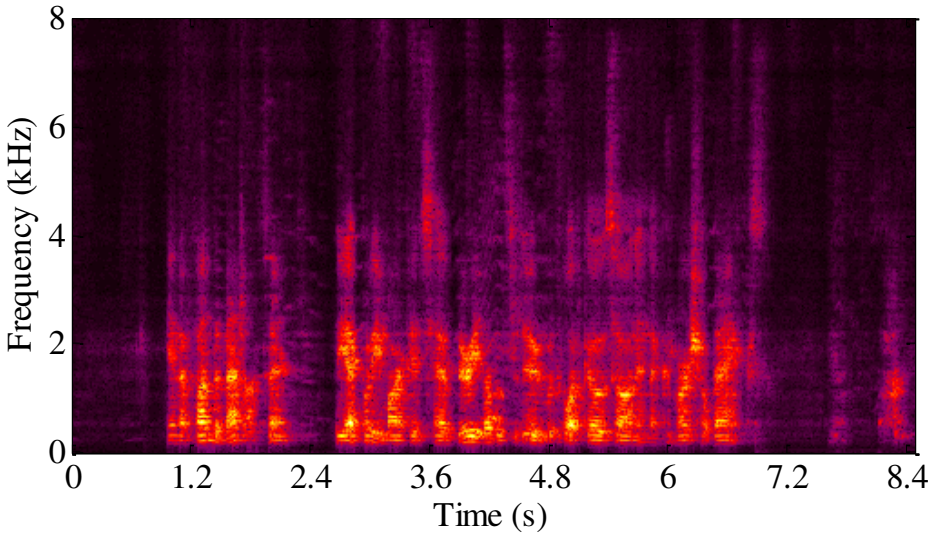
PSA kayıp fonksiyonu ile eğitildiğinde, ağ, eğer girdi verisi gürültülü ise kendi çıktı maskesini doğru miktarda küçültmeyi öğreniyor ve bu daha iyi sonuç veriyor

Spektrogramlar ve işlenmiş sesler

LSTM-MSA [SDR=8.26]



BLSTM-PSA-Align [SDR=10.51]



DNN-HMM konuşma tanıma iki kanallı CHiME-2 verisi kullanıldığında

Iyileştirme yöntemi	WER Dev	WER Eval
BF	25.64	21.12
2ch-NMF	25.13	19.46
BF-LSTM-MSA	19.03	14.82
BF-LSTM-PSA	19.20	14.63
BF-BLSTM-MSA	18.35	14.47
BF+SSA-BLSTM-MSA	18.41	14.25
BF+SSA-BLSTM-PSA	18.19	14.24
BF+ENH+SSA-BLSTM-MSA	**18.16**	**13.95**
BF+ENH+SSA-BLSTM-PSA	18.28	**13.95**

DNN tanıma modeli hedefleri temiz konuşma ile hizalayarak elde edilmiştir

Dizisel eğitim ile eğitilmiş modeller kullanılmıştır

Birden fazla konuşmacının seslerini ayırma

- İki konuşmacı için eğitilecek sinir ağının iki çıktısı olmalıdır
- Fakat hangi konuşmacının hangi çıktıda yer alacağını belirlemek mümkün görünmemektedir, özellikle konuşmacılar ile ilgili ön bilgi yoksa ve herhangi bir şekilde kategorize edilemiyorlarsa (kör ayırma problemi)
- Örneğin A ve B konuşmacısı konuşurken A konuşmacısının sesi birinci çıktıda yer alırken, B ikincide yer alsın diyelim. Şimdi A ve C konuşmacısının karışımında ve B ve C konuşmacısının karışımında hangi konuşmacının hangi çıktıda yer alacağı tutarlı olarak belirlenemez.
- Orjinal fikir: Eğitirken çıktılara hangi kaynağın yerleştirileceğine bırakalım sinir ağı kendisi karar versin!
- Yani modeli eğitirken her A ve B karışımında çıktı sıralamalarının hepsi göz önüne alınır ve en düşük kaybı veren sıralama seçilir
- Sıralamadan bağımsız eğitim (permutation invariant training) bu şekilde kör ayırma sorununu çözer. [Yu, Kolbaek, Tan and Jensen, 2017]

Sıralamadan bağımsız eğitim ve diğerleri

- Birden fazla konuşmacının kör ayrımı için ayrıca aşağıdaki yöntemler de vardır
  - Derin kümeleme (deep clustering)
  - Derin çekici ağı (deep attractor net)
- PIT tekniği daha basit ve anlaşılır bir tekniktir ve başarımı diğerlerine yakındır.
- u-PIT maske uzayı kayıp fonksiyonu şöyle yazılabilir:

$$\sum_{u} \min_{\mathcal{O}} \sum_{i} \sum_{tf \in u} \left(\hat{m}_{\mathcal{O}(i),tf} - m^{*}_{i,tf}\right)^2$$

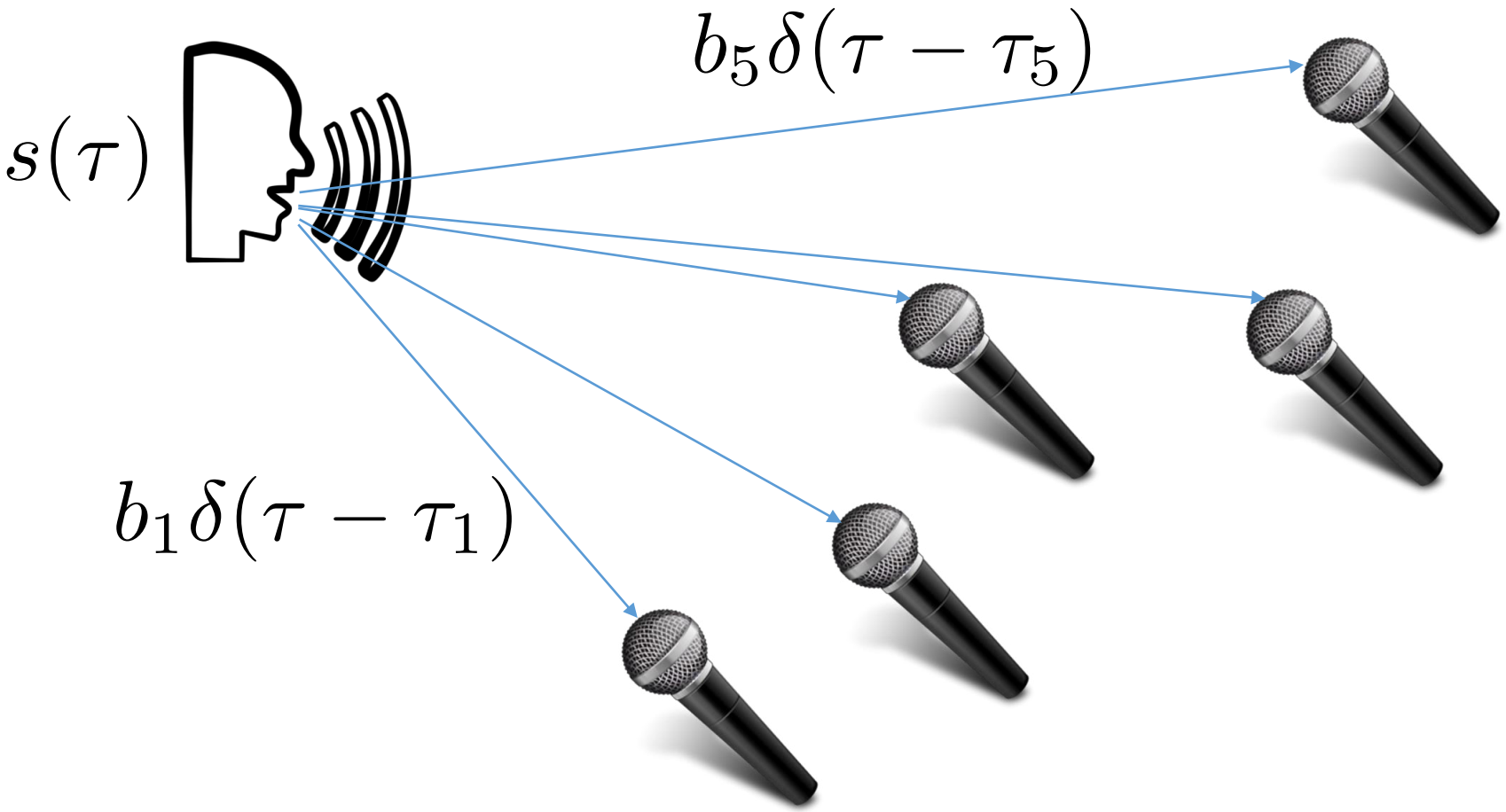
Çok kanallı sinyallerden kaynak ayırma

Şimdilik tek konuşmacı varsayalım

Tek ve Çok Kanal Farkı

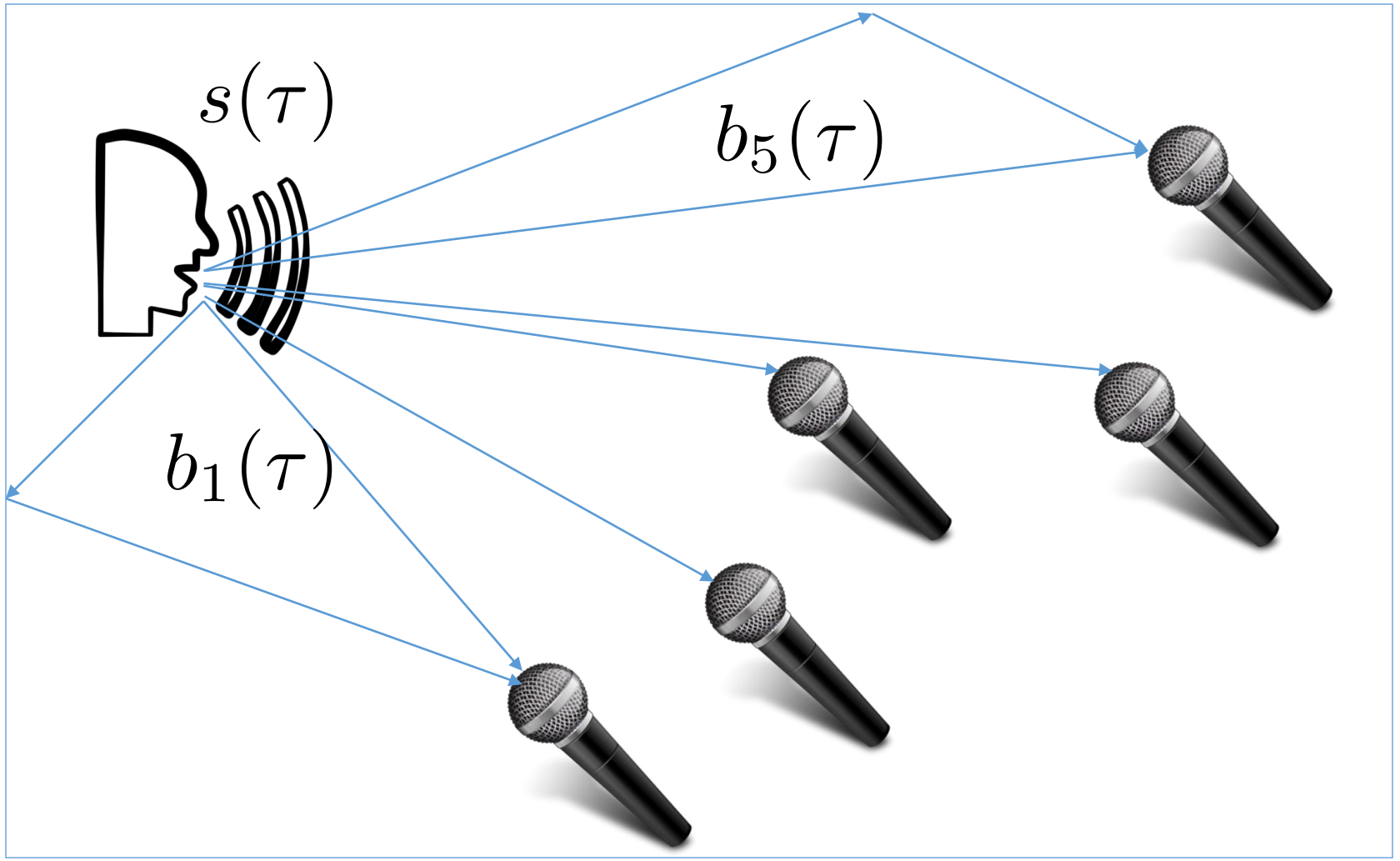
- Tek kanal
  - Karışım sinyalinde kaynakların kendi özellikleri dışında bilgi yok
  - Kaynakların zaman-frekans (spektro-temporal) bilgilerine dayanarak ayırma yapılabilir
  - Uzamsal bilgi bulunmaz
- Çok kanallı
  - Yön ve derinlik anlaşılabilir (insan kulağı iki tane)
  - Sinyallerin geliş yönü kestirilebilir
  - Hüzme oluşturma teknikleri ile belli yönlere bakılabilir
  - Uzamsal öznitelikler kullanılabilir
  - Ek olarak birden fazla kanaldan gelen zaman-frekans bilgileri de kullanılabilir

Çok mikrofonlu yansımasız ortam



Varsayımlar: nokta kaynak, nokta alıcılar, yansıma yok, sabit kaynak ve alıcılar, doğrusallık

Yansımalı ortam



Varsayımlar: sabit kaynak ve alıcılar, doğrusallık

Alınan sinyal modeli

- Yansımasız (anechoic)

$$y_i(\tau) = b_i s(\tau - \tau_i) + v_i(\tau), \quad i = 1, \ldots, M$$

- Yansımalı (reverberated)

$$y_i(\tau) = \sum_k b_i(k) s(\tau - k) + v_i(\tau)$$

$$y_i(\tau) = x_i(\tau) + v_i(\tau)$$

S(t)'nin mikrofon i üzerinde uzamsal imgesi

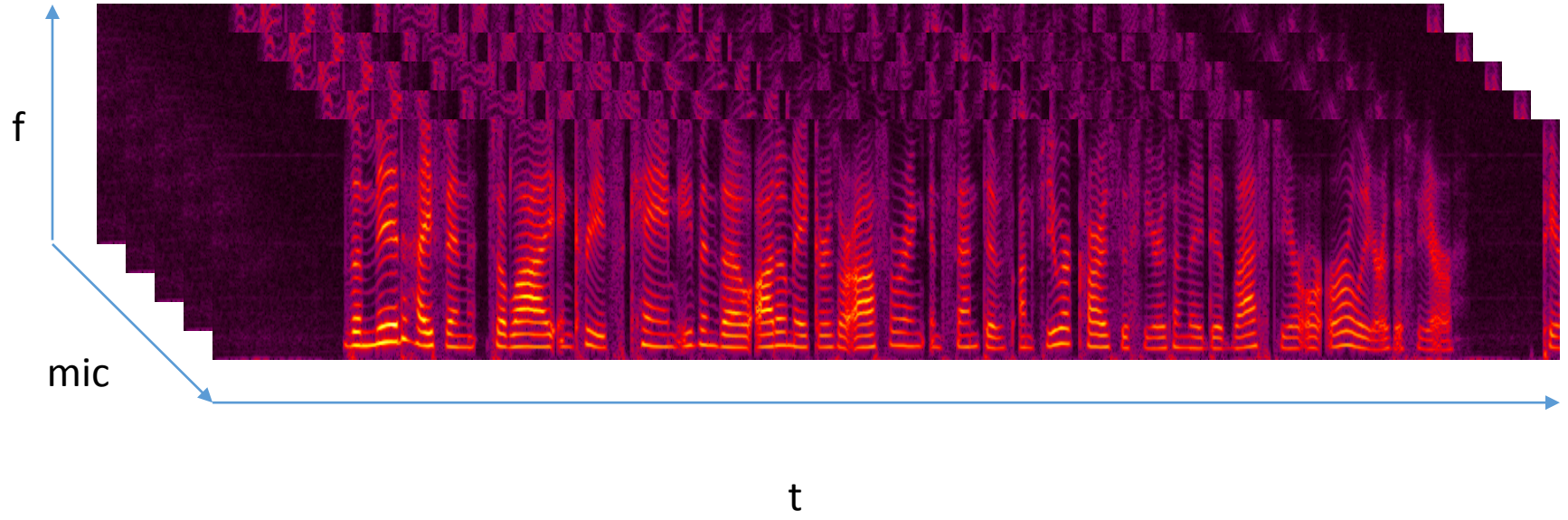
Klasik yöntem: Ağırlıklı geciktir-ve-topla hüzme yapıcısı

- Ağırlıklı geciktir ve topla hüzme yapıcısı- diğer adı beamformit [Anguera&Wooters&Hernando 2007]
  - Bir referans mikrofon seçer
  - Referans mikrofona göre her bir mikrofonun görece gecikme farklarını hesaplar
  - Gecikme farkları (TDOA) GCC-PHAT algoritması ile bulunabilir (çapraz-korelasyondaki tepeler kullanılarak)
  - Yansımasız modelle ilgilidir
- Her zaman bloğu için birden fazla aday gecikme bulunur sonra bloklar arasında tutarlı gecikme seçmek için Viterbi algoritması kullanılır

$$\hat{x}_{\mathrm{ref}}(\tau) = \sum_{i=1}^{M} w_i y_i(\tau - \tau_i)$$

Çok kanallı STFT ve uzamsal filtreleme

- Çok kanallı veri 3 boyutlu tensör olarak görülür
- Uzamsal filtreleme ya da hüzme yapma:
  - Filtrele-ve-topla hüzme yapıcısı
  - STFT uzayında, çarp-ve-topla (mikrofon boyutu boyunca)



MVDR hüzme yapıcısı denklemi

- Denklem [Souden&Benesty&Affes 2010, Benesty&Chen&Huang 2008]
- Çok kanallı karmaşık STFT'yi düşünelim – bir önceki yansı
- Her f değeri için, çarp ve topla işlemi ile sonuç elde edilir

$$\hat{x}_{t,f} = \sum_{i=1}^{M} h_i(f) y_{i,t,f}.$$

Amacımız uzamsal filtreler olan $h_i(f)$ değerlerini bularak kaynak yönünde hüzme oluşturma yani kaynağa doğru bakarak diğer kaynakları baskılamadır

MVDR denklemi için tanımlar

- Note $y_{i,t,f}$ is the STFT at the time-frequency bin (t, f) of microphone i
- Define $M \times 1$ spatial variability vectors for each time frequency bin
  - Noisy signal: $\underline{\mathsf{Y}}_{t,f} = [y_{1,t,f}, \ldots, y_{M,t,f}]^T$
  - Speech signal $\underline{\mathsf{X}}_{t,f} = [x_{1,t,f}, \ldots, x_{M,t,f}]^T$
  - Noise $\underline{\mathsf{V}}_{t,f} = [v_{1,t,f}, \ldots, v_{M,t,f}]^T$
- Define $M \times M$ spatial covariance matrices of each type of signal as $\mathbf{\Phi}_{\text{noisy}}(f) = E_t\left[\underline{\mathsf{Y}}_{t,f}\underline{\mathsf{Y}}_{t,f}^H\right]$ and similarly we can define $\mathbf{\Phi}_{\text{noise}}(f)$ and $\mathbf{\Phi}_{\text{speech}}(f)$
- Forward model channel filters vector

$$\boldsymbol{b}_f = [b_1(f), \ldots, b_M(f)]$$

- Filter-and-sum reconstruction filters vector

$$\boldsymbol{h}_f = [h_1(f), \ldots, h_M(f)]$$

MVDR kriteri

- MVDR criterion tries to minimize noise variance while keeping reconstructed speech “distortionless”
- First, we choose one of the microphones as a reference microphone and refer to its index as ref
- Distortionless constraint specifies that after filtering the source images, we should get the source image at the reference microphone:

$$\boldsymbol{h}_f^H \boldsymbol{x}_{t,f} = x_{\text{ref},t,f}$$

- Since $\boldsymbol{x}_{t,f} = s_{t,f}\boldsymbol{b}_f$, the above constraint is equivalent to $\boldsymbol{h}_f^H \boldsymbol{b}_f = b_{\text{ref}}(f)$
- Under this constraint, we seek to minimize the noise variance after reconstruction

$$\hat{\boldsymbol{h}}_f = \arg\min_{\boldsymbol{h}_f} \boldsymbol{h}_f^H \boldsymbol{\Phi}_{\text{noise}}(f)\boldsymbol{h}_f, \quad \text{such that } \boldsymbol{h}_f^H \boldsymbol{b}_f = b_{\text{ref}}(f)$$

MVDR çözüm

- The solution to the constraint optimization can be given as:

$$\hat{\boldsymbol{h}}_f = b^*_{\text{ref}}(f)\frac{\boldsymbol{\Phi}^{-1}_{\text{noise}}(f)\boldsymbol{b}_f}{\boldsymbol{b}_f^H\boldsymbol{\Phi}^{-1}_{\text{noise}}(f)\boldsymbol{b}_f}$$

- which can be rewritten as

$$\hat{\boldsymbol{h}}_f = \frac{\boldsymbol{\Phi}^{-1}_{\text{noise}}(f)\boldsymbol{b}_f\boldsymbol{b}_f^H\boldsymbol{e}_{\text{ref}}}{\boldsymbol{b}_f^H\boldsymbol{\Phi}^{-1}_{\text{noise}}(f)\boldsymbol{b}_f}$$

  where $\boldsymbol{e}_{\text{ref}}$ is an indicator vector with a 1 in position ref.

- Simplifying further since $\boldsymbol{\Phi}_{\text{speech}}(f) = \sigma^2_{\text{speech}}(f)\boldsymbol{b}_f\boldsymbol{b}_f^H$, multiplying and dividing with σ^2_{speech}, and using a matrix identity for the denominator, we get

$$\hat{\boldsymbol{h}}_f = \frac{\boldsymbol{\Phi}^{-1}_{\text{noise}}(f)\boldsymbol{\Phi}_{\text{speech}}(f)\boldsymbol{e}_{\text{ref}}}{\text{tr}\left\{\boldsymbol{\Phi}^{-1}_{\text{noise}}(f)\boldsymbol{\Phi}_{\text{speech}}(f)\right\}}$$

Genelleşmiş özvektör hüzme yapıcısı

- Generalized eigenvector beamformer aims to maximize SNR at each frequency separately after beamforming

$$\text{SNR}_f = \frac{\boldsymbol{h}^H(f)\boldsymbol{\Phi}_{\text{speech}}(f)\boldsymbol{h}(f)}{\boldsymbol{h}^H(f)\boldsymbol{\Phi}_{\text{noise}}(f)\boldsymbol{h}(f)}$$

- The solution is given by the primary eigenvector of generalized eigenvalue problem

$$\boldsymbol{\Phi}_{\text{speech}}(f)\boldsymbol{h} = \lambda\boldsymbol{\Phi}_{\text{noise}}(f)\boldsymbol{h}$$

- GEV beamformer does not have the property of being distortionless like MVDR, however it works well in practice for ASR
- We did not consider GEV in this study
- We used GEV beamformer for our CHiME-4 ASR system - Tuesday's MERL talk at CHiME-4 workshop

Maske kullanarak uzamsal kovaryans bulma

- This MVDR solution does not need to find direction of arrivals
- It does not require finding steering vectors ($\boldsymbol{b}_f$'s)
- It only requires spatial covariance estimates
- However, these estimates need to be obtained from the utterance at hand
- To estimate the spatial covariances, we need to know which time-frequency bin belongs to speech and noise
- We propose to use single-channel masks obtained using deep recurrent networks for this purpose
- This exact same idea was also proposed by [Heynmann&Drude&Haeb-Umbach 2016] as well for obtaining a generalized eigenvalue (GEV) beam-former

Spektro-temporal maske kullanımı aslında yeni değil

- Daha önceki çalışmalarda da
  - Konuşma olma ihtimali = maske = spektral filtre
- Birçok çalışma, maskeleri spektral-uzamsal öbekleme ile bulur
  - Mikrofonlar arası seviye ve faz farkları kullanan (MESSL [Mandel&Weiss&Ellis 2010])
  - Karmaşık Gauss modeli ile zaman-frekans kutucuklarının öbeklenmesi [Higuchi&Ito&Yoshioka&Nakatani 2016]
  - Direkt olarak zaman-frekans kutucuklarının öbeklenmesi [Sawada&Araki&Makino 2011] [Ito&Araki&Nakatani 2013]
  - Watson karışım modeli [Tran-Vu&Haeb-Umbach 2010]
- Derin öğrenme ile maske çıkarma yukarıdaki diğer yöntemlerin alternatifidir ve uzamsal bilgi kullanan diğerleri ile karşılaştırılmalıdır

Tek kanallı maske kestirim ağlarını eğitmek

- Önceki kısımlarda bu tür ağları eğitmenin yollarına bakmıştık
  - ideal maske (IRM ya da IBM) hedef yapılarak eğitilebilir
  - Sinyal uzayında kayıp fonksiyonu kullanılabilir
  - Biz de MSA kayıp fonksiyonu ile eğittik [Weninger et.al. 2014, Erdogan et.al. 2015]
- Girdi öznitelikler 100 log-Mel-filtre-bankası öznitelikleri
  - 25 ms pencere uzunluğu
  - 10 ms çerçeve kaydırma
- 2 katmanlı LSTM modeli her biri 256 saklı boğum içerir
- CHiME-3 benzetim verisinin 5. kanalı kullanılarak eğitildi

MVDR için uzamsal kovaryans elde etme

- Gürültülü karışım sinyalinin kovaryansı direkt olarak veriden elde edilir

$$\hat{\mathbf{\Phi}}_{\text{noisy}}(f) = \frac{1}{T}\sum_{t=0}^{T-1} Y_{t,f} Y_{t,f}^{H}$$

- Sadece gürültü sinyalinin kovaryansı ise maskelenmiş veriden şu şekilde elde edilir:

$$\hat{v}_{i,t,f} = (1 - \hat{a}_{i,t,f}) y_{i,t,f}$$

$$\hat{V}_{t,f} = \left[\hat{v}_{1,t,f} \ldots \hat{v}_{M,t,f}\right]^{T}$$

$$\hat{\mathbf{\Phi}}_{\text{noise}}(f) = \frac{1}{T}\sum_{t=0}^{T-1} \hat{V}_{t,f} \hat{V}_{t,f}^{H}$$

Ses sinyalinin uzamsal kovaryansı

- Ses sinyalinin uzamsal kovaryansı da aşağıdaki basit işlem ile elde edilebilir

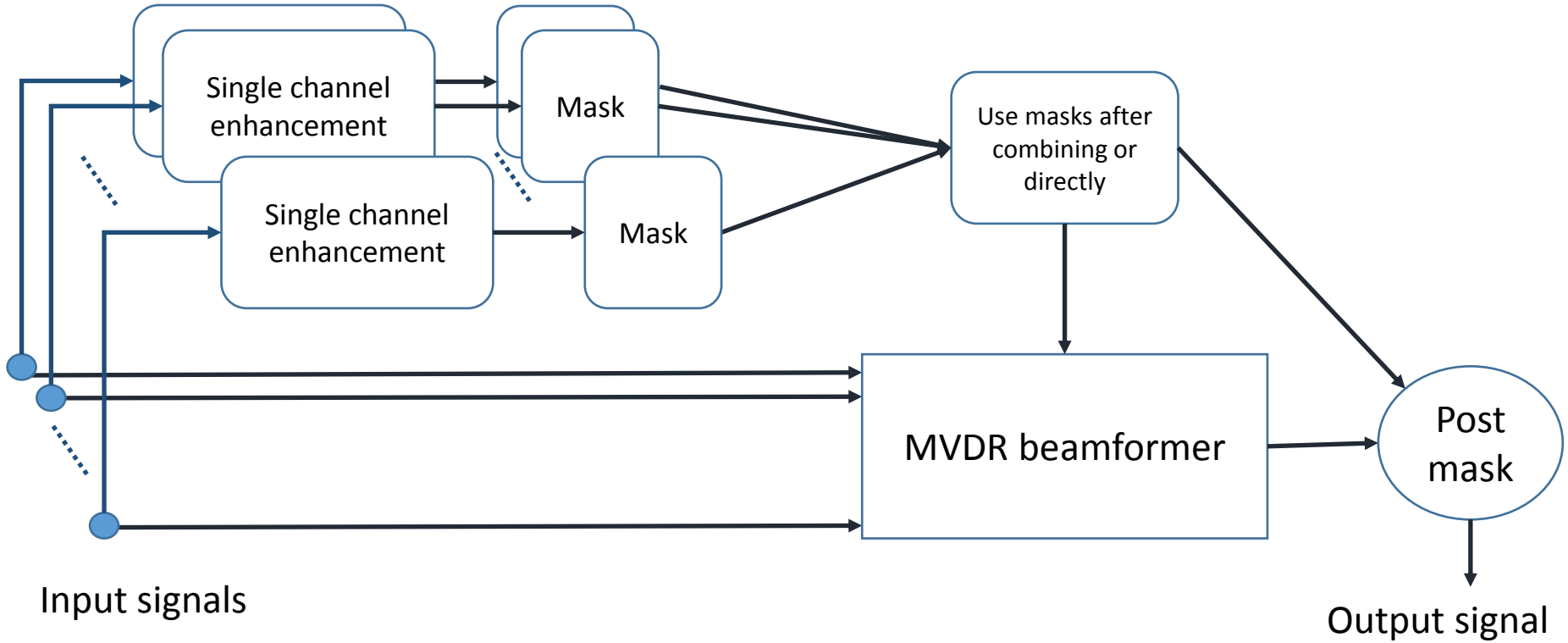
$$\hat{\mathbf{\Phi}}_{\mathrm{speech}}(f) = \hat{\mathbf{\Phi}}_{\mathrm{noisy}}(f) - \hat{\mathbf{\Phi}}_{\mathrm{noise}}(f)$$

- Ya da hem ses hem de gürültü için iki farklı maske bulunarak ses ve gürültü kovaryansları elde edilebilir, mesela [Heynmann&Drude&Haeb-Umbach 2016]

Maskeleme stratejileri

- **Edge mask**: cümlenin ilk ve son 0.5 saniyelik kısmını gürültü kabul etmek
- **Single mask**: her kanaldan elde edilen maskeyi maksimumlarını alarak birleştirmek (ortalama almak da yakın sonuç verdi)
- **Multiple masks**: her kanal kendi maskesini kullanır
- (tercihen) **post-masking**: (sonradan maskeleme)
  - Direct: referans kanalın maskesini uygulama
  - Minfloor: maske ile 0.3 değerinin maksimumunu alarak sert sıfırlayıcı maskelemeye engel olacak şekilde maskelemek

Yöntemin genel şeması



Deneyler

- Deneyleri CHiME-3 geliştirme ve değerlendirme verilerinde yaptık
- Maske kestirimi ağı tek kanaldan eğitildi
- CHiME-3 verisi bir tablet üzerinde 6 mikrofon ile kaydedilmiştir. 5 mikrofon ön tarafta, 1 tanesi arka taraftadır.
- Bir okuyucu cümleleri gürültülü ortamlarda okur. Kafe, cadde, otobüs ve yaya bölgesi gibi dört farklı gürültülü ortam.
- Ses sinyallerini maske-tabanlı MVDR algoritması ile iyileştirdik ve sonra SDR ve PESQ ölçümleri ile başarımı ölçtük.

Sonuçlar – SDR dB, CHiME-3 geliştirme ve değerlendirme kümeleri

mask	post-mask	sim-dev	real-dev	sim-test	real-test
Edge-mask	none	11.78	3.70	12.02	4.20
Single-mask	none	15.04	5.87	14.36	5.02
Single-mask	minfloor	15.79	**6.72**	**15.12**	**5.52**
Single-mask	direct	**15.80**	**6.72**	15.10	5.36
Multi-mask	none	13.42	3.94	13.00	3.75
Multi-mask	minfloor	14.82	5.57	14.22	4.71
CH5 LSTM-enh	n/a	10.44	4.41	10.41	3.11
CH5 noisy	n/a	5.79	1.09	6.50	1.69

Sonuçlar– PESQ, CHiME-3 geliştirme ve değerlendirme verileri

mask	post-mask	sim-dev	real-dev	sim-test	real-test
Edge-mask	none	1.58	1.42	1.67	1.72
Single-mask	none	1.83	1.65	1.91	**1.85**
Single-mask	minfloor	**2.19**	**1.68**	**2.29**	1.79
Single-mask	direct	2.15	1.58	2.27	1.54
Multi-mask	none	1.73	1.50	1.77	1.70
Multi-mask	minfloor	2.13	1.57	2.22	1.70
CH5 LSTM-enh	n/a	1.62	1.35	1.67	1.33
CH5 noisy	n/a	1.27	1.28	1.27	1.45

Çok kanallı PIT ile konuşma ayırma

- Birden fazla konuşmacı varsa
- Aynı tek kanallı PIT'de olduğu gibi konuşmacı sayısı kadar maske kestirilir ve sıralamadan bağımsız bir kayıp fonksiyonu kullanılır
- Sonra maskeler kullanılarak maske-tabanlı MVDR hüzme yapıcı ile her konuşucu için hüzme yapılır
- Çok kanallı PIT için spektral öznitelik yanında çok kanallı sinyalden elde edilen mikrofonlar arası faz farkları (inter-channel phase - IPD) da öznitelik olarak kullanılır. Bu öznitelik uzamsal bilgi içerir. [Yoshioka&Erdogan 2018]

Çok kanallı PIT ses ayırma gerçek veri örneği

Karışmış sinyal (mikrofon 0)

Ayrılmış sinyal 1

Ayrılmış sinyal 0

Sonuçlar

- Ses kaynağı ayrımı konusunda son senelerde çok gelişmeler oldu
- Son ICASSP konferansında yaklaşık 6 oturum bu konuda idi
- Kokteyl parti problemini çözmeye oldukça yaklaşıyoruz
- İnsanlık problemleri çözdükçe daha zor ve çözülmemiş problemlere yöneliyor
  - Tek konuşmacıyı anladık da sıra birden fazlasına mı geldi?
- Genel yapay zeka'dan daha çok uzağız ve oraya giderken çözeceğimiz çok fazla problem var

DİNLEDİĞİNİZ İÇİN
ÇOK TEŞEKKÜRLER

SORULARINIZ?

References-1

- [Ephraim&Malah 1984] Ephraim, Yariv, and David Malah. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 32.6 (1984): 1109-1121.
- [Hochreiter&Schmidhuber 1997] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
- [Cohen&Berdugo 2001] Cohen, Israel. "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging." *Speech and Audio Processing, IEEE Transactions on* 11.5 (2003): 466-475.
- [Lee&Seung 2001] Lee, Daniel D., and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems*. 2001.
- [Smaragdis&Brown 2003] Smaragdis, Paris, and Judith C. Brown. "Non-negative matrix factorization for polyphonic music transcription." *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*. IEEE, 2003.
- [Schmidt&Olsson 2006] Schmidt, Mikkel, and Rasmus Olsson. "Single-channel speech separation using sparse non-negative matrix factorization." (2006).
- [Vincent&Gribonval&Fevotte 2006] E Vincent, R Gribonval, C Févotte, “Performance measurement in blind audio source separation,” Audio, Speech, and Language Processing, IEEE Transactions on 14 (4), 1462-1469, 2006.
- [Hu&Loizou 2008] Hu, Yi, and Philipos C. Loizou. "Evaluation of objective quality measures for speech enhancement." *Audio, Speech, and Language Processing, IEEE Transactions on* 16.1 (2008): 229-238.
- [Virtanen&Cemgil&Godsill 2008] Virtanen, Tuomas, Ali Taylan Cemgil, and Simon Godsill. "Bayesian extensions to non-negative matrix factorisation for audio signal modelling." *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008.

References-2

- [Boldt&Ellis 2009] Boldt, Jesper B., and Daniel PW Ellis. "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation." *EUSIPCO 2009: 17th European Signal Processing Conference, August 24-28, 2009, Glasgow, Scotland*. European Association for Signal, Speech, and Image Processing, 2009.
- [Mohamed&Dahl&Hinton 2009] Abdel-rahman Mohamed, George E. Dahl, Geoffrey E. Hinton. “Deep Belief Networks for Phone Recognition.” *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.
- [Hershey&Rennie&Olsen&Kristjansson 2010] Hershey, J. R., Rennie, S. J., Olsen, P. A., & Kristjansson, T. T. “Super-human multi-talker speech recognition: A graphical modeling approach.” *Computer Speech & Language*, *24*(1), 45-66, 2010.
- [Grais&Erdogan 2011] Grais, Emad M., and Hakan Erdogan. "Single channel speech music separation using nonnegative matrix factorization and spectral masks." *Digital Signal Processing (DSP), 2011 17th International Conference on*. IEEE, 2011.
- [Yu&Deng 2011] Yu, Dong, and Li Deng. "Deep learning and its applications to signal and information processing [exploratory dsp]." *Signal Processing Magazine, IEEE* 28.1 (2011): 145-154.
- [Taal&Hendriks&Heusdens&Jensen 2011] Taal, Cees H., et al. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech." *Audio, Speech, and Language Processing, IEEE Transactions on* 19.7 (2011): 2125-2136.
- [Wang&Wang 2013] Wang, Yuxuan, and DeLiang Wang. "Towards scaling up classification-based speech separation." *Audio, Speech, and Language Processing, IEEE Transactions on* 21.7 (2013): 1381-1390.
- [Huang&Kim&Johnson&Smaragdis 2014] Po-Sen Huang; Minje Kim; Hasegawa-Johnson, M.; Smaragdis, P., "Deep learning for monaural speech separation," *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* , vol., no., pp.1562,1566, 4-9 May 2014.

References-3

- [Grais&Sen&Erdogan 2014] Grais, E.M.; Sen, M.U.; Erdogan, H., "Deep neural networks for single channel source separation," *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* , vol., no., pp.3734,3738, 4-9 May 2014.
- [Weninger&Hershey&LeRoux&Schuller 2014] Felix Weninger, Jonathan Le Roux, John R. Hershey, Björn Schuller, "Discriminatively Trained Recurrent Neural Networks for Single-Channel Speech Separation," to appear in Proc. IEEE GlobalSIP 2014 Symposium on Machine Learning Applications in Speech Processing, Dec 2014.
- [Weninger et.al. 2014] Felix Weninger, Shinji Watanabe, Jonathan Le Roux, John R. Hershey, Yuuki Tachioka, Jürgen Geiger, Björn Schuller, Gerhard Rigoll: *"The MERL/MELCO/TUM system for the REVERB Challenge using Deep Recurrent Neural Network Feature Enhancement"*, Proc. REVERB Workshop held in conjunction with ICASSP 2014 and HSCMA 2014, IEEE, Florence, Italy, 10.05.2013
- [Weninger et.al. 2014] Felix Weninger, Shinji Watanabe, Jonathan Le Roux, John R. Hershey, Yuuki Tachioka, Jürgen Geiger, Björn Schuller, Gerhard Rigoll: *"The MERL/MELCO/TUM system for the REVERB Challenge using Deep Recurrent Neural Network Feature Enhancement"*, Proc. REVERB Workshop held in conjunction with ICASSP 2014 and HSCMA 2014, IEEE, Florence, Italy, 10.05.2013
- [Weninger&LeRoux&Hershey&Watanabe 2014] Felix Weninger, Jonathan Le Roux, John R. Hershey, Shinji Watanabe, "Discriminative NMF and its application to single-channel source separation," *Proc. ISCA Interspeech 2014 (Interspeech 2014),* Sep. 2014.
- [Xu&Du&Dai&Lee 2014] Yong Xu; Jun Du; Li-Rong Dai; Chin-Hui Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *Signal Processing Letters, IEEE* , vol.21, no.1, pp.65,68, Jan. 2014.
- [Wang&Narayanan&Wang 2014] Yuxuan Wang; Narayanan, A.; DeLiang Wang, "On Training Targets for Supervised Speech Separation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on* , vol.22, no.12, pp.1849,1858, Dec. 2014.

References-4

- [Wang&Wang 2015] Yuxuan Wang; DeLiang Wang, “A deep neural network for time-domain signal reconstruction," *ICASSP 2015*.
- [Anguera&Wooters&Hernando 2007] Anguera, Xavier, Chuck Wooters, and Javier Hernando. "Acoustic beamforming for speaker diarization of meetings." *IEEE Transactions on Audio, Speech, and Language Processing* 15.7 (2007): 2011-2022.
- [Souden&Benesty&Affes 2010] Souden, Mehrez, Jacob Benesty, and Sofiène Affes. "On optimal frequency-domain multichannel linear filtering for noise reduction." *IEEE Transactions on audio, speech, and language processing* 18.2 (2010): 260-276.
- [Benesty&Chen&Huang 2008] Benesty, Jacob, Jingdong Chen, and Yiteng Huang. *Microphone array signal processing*. Vol. 1. Springer Science & Business Media, 2008.
- [Mandel&Weiss&Ellis 2010] Mandel, Michael I., Ron J. Weiss, and Daniel PW Ellis. "Model-based expectation-maximization source separation and localization." *IEEE Transactions on Audio, Speech, and Language Processing* 18.2 (2010): 382-394.
- [Higuchi&Ito&Yoshioka&Nakatani 2016] Higuchi, Takuya, et al. "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.
- [Ito&Araki&Nakatani 2013] Ito, Nobutaka, Shoko Araki, and Tomohiro Nakatani. "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.
- [Sawada&Araki&Makino 2011] Sawada, Hiroshi, Shoko Araki, and Shoji Makino. "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment." *IEEE Transactions on Audio, Speech, and Language Processing* 19.3 (2011): 516-527.
- [Tran-Vu&Haeb-Umbach 2010] Vu, Dang Hai Tran, and Reinhold Haeb-Umbach. "Blind speech separation employing directional statistics in an expectation maximization framework." *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.
- [Weninger et.al. 2014] Weninger, Felix, et al. "Discriminatively trained recurrent neural networks for single-channel speech separation." *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014.
- [Erdogan et.al. 2015] Erdogan, Hakan, et al. "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks." *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.
- [Heynmann&Drude&Haeb-Umbach 2016] Heymann, Jahn, Lukas Drude, and Reinhold Haeb-Umbach. "Neural network based spectral mask estimation for acoustic beamforming." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.
- [Xiao et.al. 2016] Xiao, Xiong, et al. "Deep beamforming networks for multi-channel speech recognition." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.