

1

KONUŞMA SENTEZİ

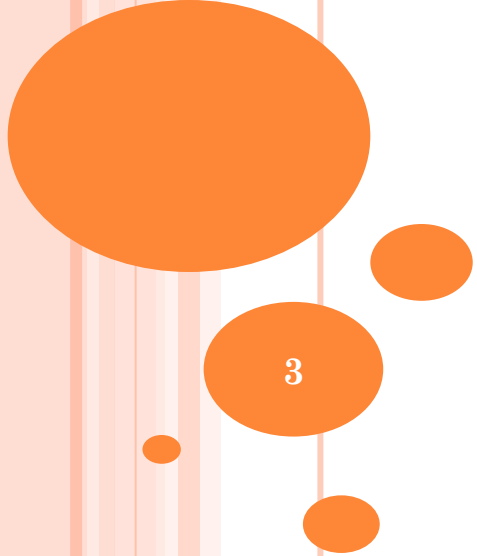
Dr Cenk Demiroğlu
Özyeğin Üniversitesi

SUNUM PLAN₁

- İstatistiksel konuşma sentezi
- Konuşmacıya uyarlanma
- Derin öğrenen ağlarla akustik modelleme
- Postfiltre-tabanlı iyileştirme
- GAN ağları
- Konuşmacıya uyarlanma



İSTATİSTİKSEL KONUŞMA SENTEZİ



İÇERİK

1. Giriş
2. Birim Seçimi tabanlı TTS
3. HMM tabanlı TTS
4. Melez sistemler
 1. UTTS ve HTSS karşılaştırması
 2. Melez yöntemler
5. Önerilen melez sistem
 1. Temel HTTS sistemi
 2. Melez tasarım– Genel değerlendirme
 3. Morfem Seçim Algoritmaları
6. Deneyler ve sonuçlar



İÇERİK

1. Giriş
2. Birim Seçimi tabanlı TTS
3. HMM tabanlı TTS
4. Melez sistemler
 1. UTTS ve HTSS karşılaştırması
 2. Melez yöntemler
5. Önerilen melez sistem
 1. Temel HTTS sistemi
 2. Melez tasarım– Genel değerlendirme
 3. Morfem Seçim Algoritmaları
6. Deneyler ve sonuçlar



1. GİRİŞ

- **TTS (text-to-speech)** yazılı dil sinyalini(metin) konuşma sentezleme tekniklerini kullanarak sözlü dil sinyaline(konuşma) çevirme işlemidir.
- Öne çıkan bazı sentezleme sistemleri
 - Eklemeli Konuşma Sentezi (CSS)
 - Birim Seçimi Tabanlı TTS (UTTS)
 - İstatiksel Konuşma Sentezi (SSS)
 - HMM tabanlı TTS (HTTS)



1. GİRİŞ

- Sistemin karşılaştırmalı olarak artıları ve eksileri
 - UTTS
 - Doğal, büyük veri tabanları gerektiriyor, konuşma karakteristiklerini değiştirmek zor.
 - HTTS
 - Daha az doğal, kaynak gerektiriyor, konuşma karakteristiklerini değiştirme açısından esnek
- Araştırmacılar iki sistemin de avantajlarını kullanabilecek melez tasarımlar üzerinde duruyorlar.
- Morfolojik olarak zengin diller için, temel HTTS sisteminin kalitesini küçük UTTS sistemleri kullanarak geliştirmeyi amaçlayan melez bir sistem önerdik.



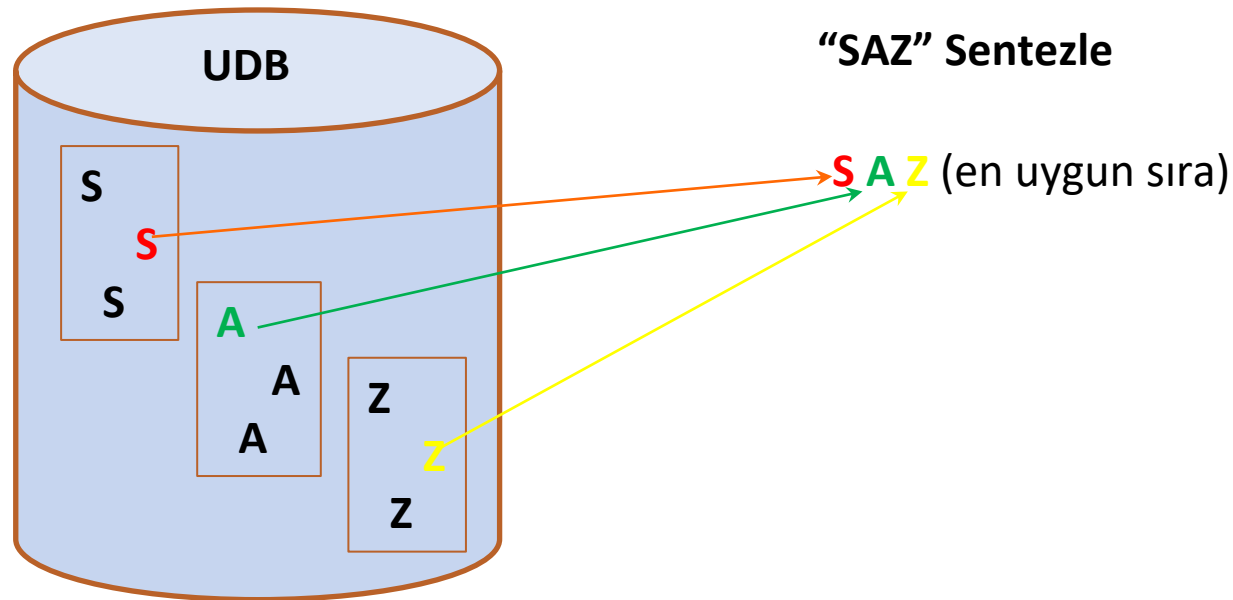
İÇERİK

1. Giriş
2. Birim Seçimi tabanlı TTS
3. HMM tabanlı TTS
4. Melez sistemler
 1. UTTS ve HTSS karşılaştırması
 2. Melez yöntemler
5. Önerilen melez sistem
 1. Temel HTTS sistemi
 2. Melez tasarım– Genel değerlendirme
 3. Morfem Seçim Algoritmaları
6. Deneyler ve sonuçlar



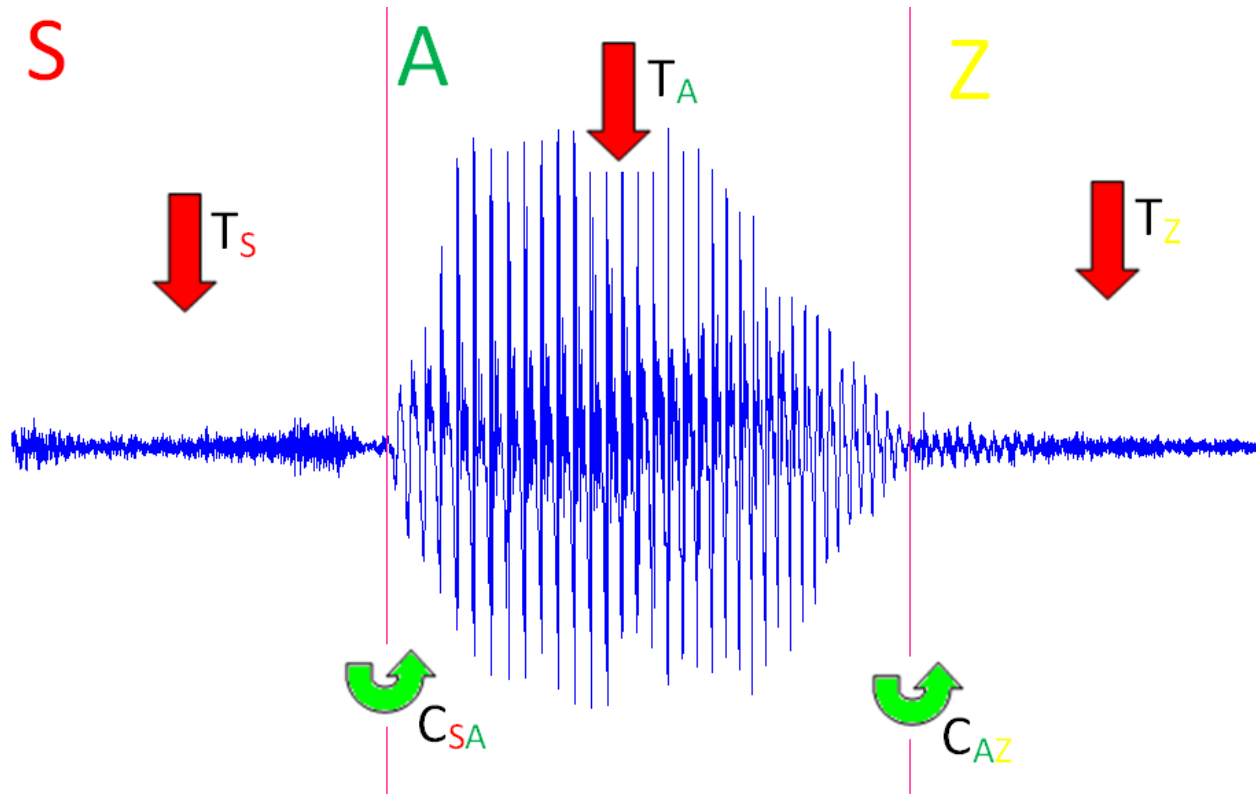
2. BİRİM SEÇİMİ TABANLI TTS (UTTS)

- **Birim Veri Tabanı (UDB):** Kaydedilmiş konuşmalardan bölünmüş konuşma birimleri veri tabanı.
 - Birim: Alt-fonem, fonem, ikili ses, karışık büyüklük, vs
- Verilen bir cümlenin sentezlenebilmesi için konuşma birimlerinin en uygun sırası UDB'den seçilir.



2. BİRİM SEÇİMİ TABANLI TTS (UTTS)

- **En uygun sıra** toplam maliyeti en az sıradır.
 - Toplam Maliyet= $\sum(\text{Hedef Maliyetleri})+\sum(\text{Ekleme Maliyetleri})$



Konuşma en uygun sıradaki birimlerin en az sinyal işleme ile eklenmesiyle oluşturulur.

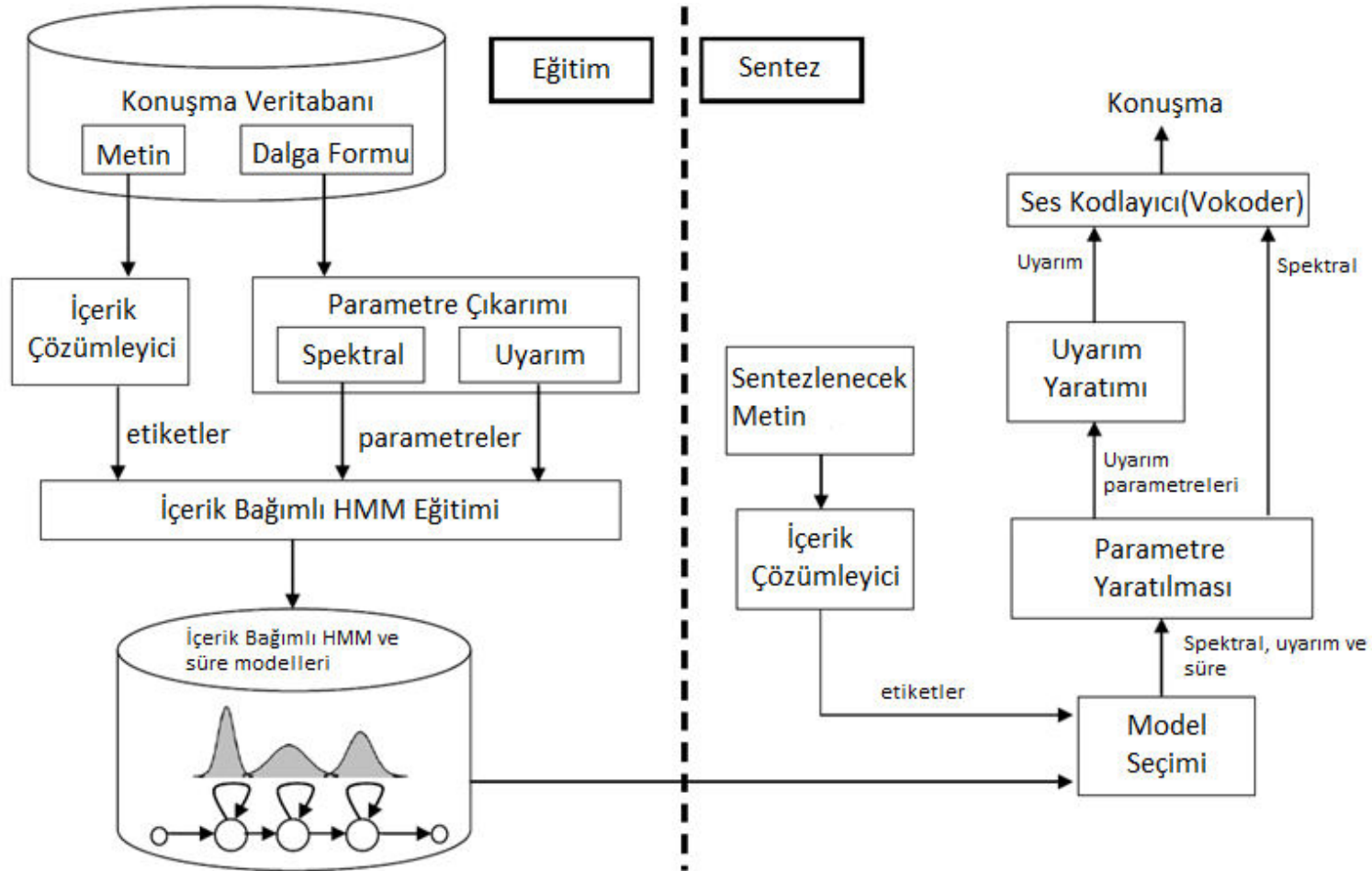


İÇERİK

1. Giriş
2. Birim Seçimi tabanlı TTS
3. **HMM tabanlı TTS**
4. Melez sistemler
 1. UTTS ve HTSS karşılaştırması
 2. Melez yöntemler
5. Önerilen melez sistem
 1. Temel HTTS sistemi
 2. Melez tasarım– Genel değerlendirme
 3. Morfem Seçim Algoritmaları
6. Deneyler ve sonuçlar



3. HMM TABANLI TTS (HTTS)



3. HMM TABANLI TTS (HTTS)

- Tanımlayıcı Etiketler

İçerik	Bilgi
Fonem seviyesi	Önce ve sonra gelen fonemler
Hece seviyesi	Fonem sayısı, önceki şiveler ve vurgular; güncel, sonraki heceler. Güncel kelimedeki sıra.
Kelime seviyesi	Önceki, güncel, sonraki kelimelerin hece sayısı. Güncel deyişteki sıra.
Deyiş seviyesi	Hece sayısı; önceki, güncel, sonraki deyişlerdeki kelimele. Güncel deyişin sözcedeki sırası.
Sözce(utterance) seviyesi	Hece, kelime ve deyişlerin sözcedeki sayısı.
Tonlama	vurgu, accent in syllables TOBI end tones



3.1 PARAMETRE ÜRETİMİ VE SENTEZİ

- Eğitilmiş model parametreleri Λ olarak verilsin. $O = \{o_1, o_2, \dots, o_T\}$ şeklinde bir parametre dizisi aşağıdaki sonsal (posterior) olasılığı en fazla yapacak şekilde üretilebilir.

$$\hat{O} = \operatorname{argmax}_O p(O | \Lambda)$$

$$\hat{O} = \operatorname{argmax}_O \sum_{n=1}^Q p(O, q^{(n)} | \Lambda)$$

$$\hat{O} = \operatorname{argmax}_O \sum_{n=1}^Q p(O | q^{(n)}, \Lambda) p(q^{(n)} | \Lambda)$$



$$\hat{O} = \operatorname{argmax}_O p(O | \hat{q}, \Lambda)$$

$$\hat{O} = \operatorname{argmax}_O \left(\prod_{t=1}^T P(O; \mu_{q_t}, \Sigma_{q_t}) \right)$$

O'yu tüm olası yollardan en fazla yapmaya çalışmak zor bir problem. Bu yüzden onu **en iyi yol** olan \hat{q} üzerinden en fazla yapmaya çalışmak daha basit olacaktır. \hat{q} **durum sürelerinin dağılımını en fazla** yapacak şekilde seçilebilir.

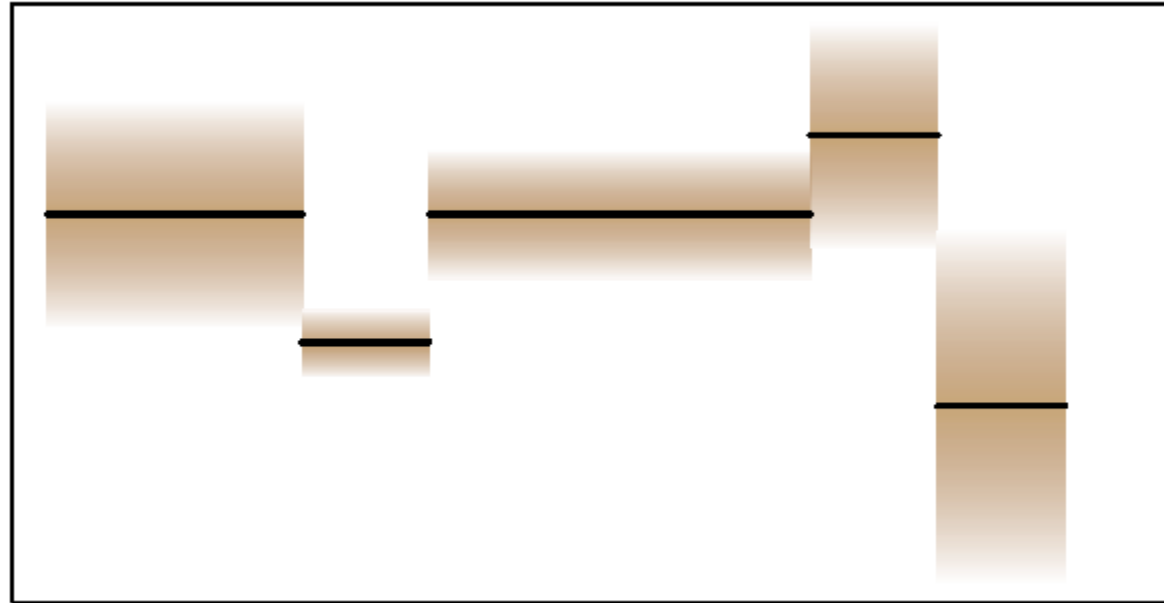
$$\hat{q} = \operatorname{argmax}_{q^{(n)}} p(q^{(n)} | \Lambda)$$



3.1 PARAMETRE ÜRETİMİ VE SENTEZİ

- Çerçeveler arasında bir kısıtlama olmadığından; çözüm bir durumun bütün çerçevelerinin, durumun dağılımının ortalamasına eşit olduğu bir parametre dizisi yaratır.

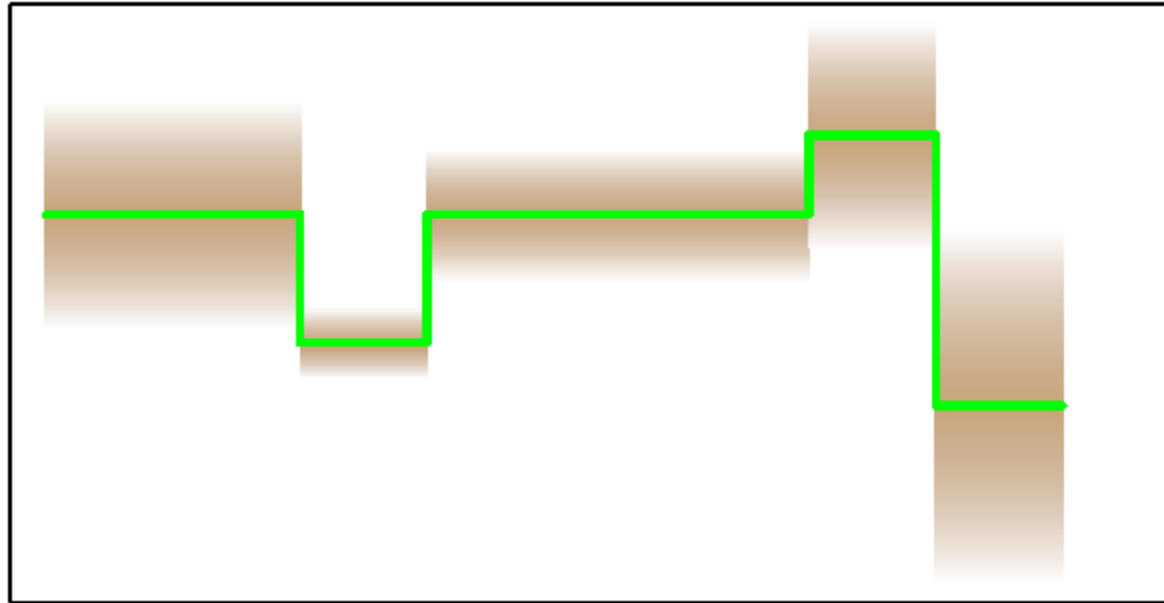
$$\hat{O} = \{\mu_{q_1}, \mu_{q_2}, \dots, \mu_{q_T}\}$$



3.1 PARAMETRE ÜRETİMİ VE SENTEZİ

- Çerçeveler arasında bir kısıtlama olmadığından; çözüm bir durumun bütün çerçevelerinin, durumun dağılımının ortalamasına eşit olduğu bir parametre dizisi yaratır.

$$\hat{O} = \{\mu_{q_1}, \mu_{q_2}, \dots, \mu_{q_T}\}$$



3.1 PARAMETRE ÜRETİMİ VE SENTEZİ

- Üretilen bu parametreler doğallığı ciddi bir şekilde düşüren dinamiklikten yoksundur. Denklemi yeniden düzenleyerek dinamikliği uygulatabiliriz.

$$\mathbf{O} = W\mathbf{C}$$

W türev matrisidir

$$\mathbf{C} = \underset{c}{\operatorname{argmax}} p(W\mathbf{C} | \hat{q}, \Lambda)$$

$$\frac{\partial \log(p(W\mathbf{C} | \hat{q}, \Lambda))}{\partial \mathbf{C}} \Big|_{\mathbf{C}=\hat{\mathbf{C}}} = 0$$

$$\mathbf{M} = [\mu_{q_1}^T, \mu_{q_2}^T, \dots, \mu_{q_T}^T]$$

- çözüm

$$\mathbf{U}^{-1} = \operatorname{diag}[\Sigma_{q_1}^{-1}, \Sigma_{q_2}^{-1}, \dots, \Sigma_{q_T}^{-1}]$$

$$W^T \mathbf{U}^{-1} W \hat{\mathbf{C}} = W^T \mathbf{U}^{-1} \mathbf{M}$$

* Bu çözüm, maksimum olabilirlik parametre üretimi şeklinde ifade edilir. (MLPG)



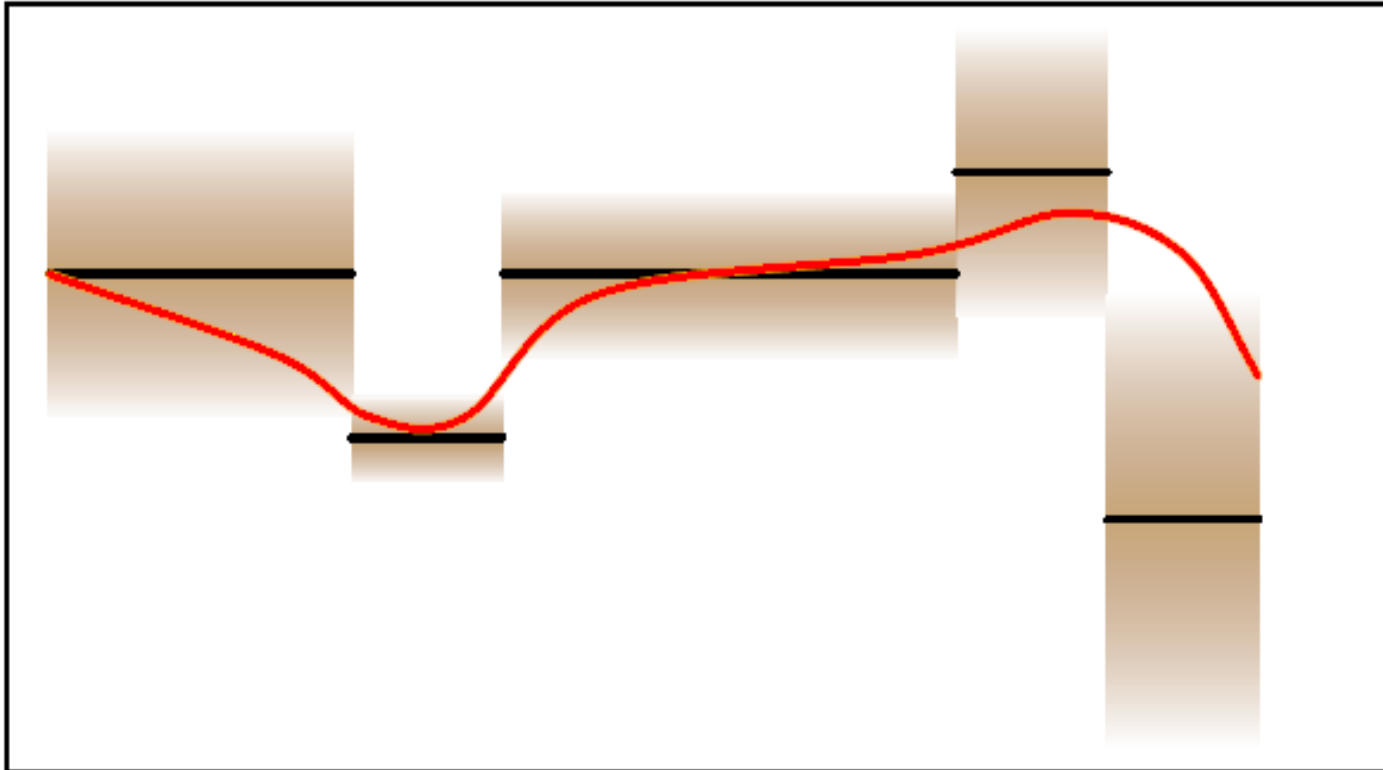
3.1 PARAMETRE ÜRETİMİ VE SENTEZİ

- **MLPG'den yaratılmış** yörünge dinamik ve ortakdeğişke(covariance) kısıtlarını yansıtır.
- Konuşma parametreleri üretildiğinde **final konuşma ses kodlayıcı(vocoder) tarafından sentezlenebilir.**



3.1 PARAMETRE ÜRETİMİ VE SENTEZİ

- **MLPG'den yaratılmış** yörünge dinamik ve ortakdeğişke(covariance) kısıtlarını yansıtır.
- Konuşma parametreleri üretildiğinde **final konuşma ses kodlayıcı(vocoder) tarafından sentezlenebilir.**



İÇERİK

1. Giriş
2. Birim Seçimi tabanlı TTS
3. HMM tabanlı TTS
4. **Melez sistemler**
 1. UTTS ve HTSS karşılaştırması
 2. Melez yöntemler
5. Önerilen melez sistem
 1. Temel HTTS sistemi
 2. Melez tasarım– Genel değerlendirme
 3. Morfem Seçim Algoritmaları
6. Deneyler ve sonuçlar



4.1 UTTS VE HTTPS KARŞILAŞTIRMASI

- **Esneklik**
 - UTTS
 - UTTS sistemleri dönüşüm yapabilme kabiliyetine sahip olmalarına rağmen HTTPS kadar esnek değildirler.
 - HTTPS
 - HTTPS'nin **parametrik yapısı** ses karakteristiklerini, duyguyu ve konuşma tarzını değiştirmek için çok uygun bir altyapı yaratır.



4.1 UTTS VE HTTPS KARŞILAŞTIRMASI

- **Kalite/Doğallık**

- UTTS

- UTTS, HTTPS'ye kıyasla **daha kaliteli** konuşma yaratır.
- Birim sınırlarında **ani bozulmalar** meydana gelebilir.

- HTTPS

- UTTS'de böyle bir endişe olmamasına kıyasla HTTPS'de **ses üretimi henüz optimal değildir**.
- Öte yandan HTTPS, sinir bozucu bozulmaları yok eden, **düzgün parametre yörüngeleri** yaratır.
- Fakat, **fazla düzgünleştirme** üretilen sesin doğallığının bozulmasına neden olabilir.



4.1 UTTS VE HTTPS KARŞILAŞTIRMASI

- **Kaynak Gereksinimleri**

- UTTS

- Verilen bir alandaki birimleri yeterince kapsayabilmek adına kaydedilen konuşmanın ciddi bir bölümü **çalışma süresi(run time)** boyunca erişilebilir olmak zorundadır.

- HTTPS

- Eğitim aşamasında modellerin istatistiği hesaplandıktan sonra sentezleme boyunca veri tabanına ihtiyaç yoktur. Sadece modeller saklanır. Ayrıca, genellikle HTTPS sistemlerinde kullanılan veri tabanının büyüklüğü UTTS'dekilere kıyasla çok daha küçüktür.
 - İstatiksel model eğitimi **daha az veri** gerektirir.



İÇERİK

1. Giriş
2. Birim Seçimi tabanlı TTS
3. HMM tabanlı TTS
4. Melez sistemler
 1. UTTS ve HTSS karşılaştırması
 2. **Melez yöntemler**
5. Önerilen melez sistem
 1. Temel HTTS sistemi
 2. Melez tasarım– Genel değerlendirme
 3. Morfem Seçim Algoritmaları
6. Deneyler ve sonuçlar



4.2 MELEZ YÖNTEMLER

- Literatürde çeşitli melez tasarımlar öne sürülmüştür. Algoritmaya bağlı olarak bunlar birkaç grup altında toplanabilir.
 - HMM Yönlendirmeli Birim Seçimi
 - HMM İle Birimleri Düzgünleştirme
 - HMM-güdümlü sistemler
 - Bölütleri karıştırma
 - Birleştirme



4.2.1 HMM Yönlendirmeli Birim Seçimi

- Rouibia ve Rosec **HMM'leri** (mfcc, ses perdesi, enerji) birim seçimi için **akustik hedeflerini** yaratmak amacıyla kullanmıştır.
- Sentezlemede HTTS sistemlerinin süresi **durum olasılığıyla doğru orantılı** olacak şekilde ayarlanır.
- **Akustik parametreler MLPG** ile yaratılır.
- HMM ile üretilen akustik parametreler **maliyet hesaplamaları** için **ikili ses(diphone) büyüklüğünde birimlere bölünür**.
- İşleri kolaylaştırmak adına her bir birim için **en iyi N aday** belirlenir. (DTW)
- En uygun birimlerin dizisi **ses perdesi sapmasını en aza** indirgemeye çalışan **dinamik aramayla** belirlenir.



4.2.1 HMM Yönlendirmeli Birim Seçimi

- Yang ve arkadaşları **HMM'leri** ayrıca (spektrum ve ses perdesi) **hedef maliyet formüllerinde** kullanmışlardır. Fonemler viterbi ile eğitilmiş HMM'leri kullanarak hizalanır. **Ardından ayrıca süre HMM'ini de yeni hizalandırma ile eğitmişlerdir.**
- Her bir birimin içeriğini, doğal dağılımını, C_{nat} , yakınsamaya yarayan UDB şeklinde tutmuşlardır.
- Sentez aşamasında, en az **log olabilirlik oranlı** (LLR, basit ML değil) K-en iyi birimler seçilir.
- Varsayım sentetik ortalamasının doğal olmadığı, HMM(doğal olmayan) ortalamasından olabildiğince en uzak olmasıdır. $\sum_{all_feature_sets} w_{feat} (p(O_n | C_{nat,n}) - p(O_n | C_{hmm,n}))$ W_s elle ayarlanır
- $LLR(Un) =$
- Final konuşma dalga formunu yaratmak amacıyla **seçilen birimler(viterbi) çapraz-sönme tekniği** kullanarak eklenir.



4.2.1 HMM Yönlendirmeli Birim Seçimi

- Pan ve arkadaşları HMM'leri ayrıca; spektrum, F0, güç, ve süre ile maliyet hesaplamalarında kullanmışlardır.
- Sentezleme aşamasında, 4 parametre seti için **MLPG** ile parametre üretmişlerdir.
- Bir birimin **hedef maliyeti** üretilen ve aday birim arasındaki Öklid mesafesinin ortalamasıdır.
- $TC(U) = d_{\text{spek}} + d_{F0} + d_{\text{güç}} + d_{\text{süre}}$
- **Ekleme maliyeti** uzaklık tabanlı olarak hesaplanır. 4 **F0** istatistiği (başlangıç, bitiş, başlangıç türevi, bitiş türevi) ve bir spektrum kullanmışlardır.
- **En uygun dizi** viterbi ile belirlenir.



4.2.1 HMM Yönlendirmeli Birim Seçimi

- Tarif edilen sistemlerin bir **ortak** noktası hepsinin HMM eğitimlerinde **ML hesaplamaları** kullanmasıdır.
- HMM tabanlı birim seçimi sistemlerini geliştirmek için Ling ve Wang minimum birim seçimi hatasını(**MUSE**) öne sürmüşlerdir.
- Genel tasarımda iki problemi tanımlamışlardır.
 - HMM'lerin ağırlıkları manuel işlemlerle ayarlanır.
 - Birim seçimi için ML kistasının kullanımı ne derece önemlidir
- **Eğitim verisi için, seçilmiş ve doğal birim dizisi arasındaki farklı birimlerin** sayısı olan Birim Seçim Hatası'nı(USE) tanımlamışlardır.
- Birim seçimi hedef fonksiyonunu yeniden yapılandırıp USE'yi formüle yerleştirmişlerdir. **Genel olasılıksal inişi**(GPD) kullanarak ve **USE**'yi en aza indirirerek HMM parametrelerinin model ağırlıklarını yinelemeli şekilde en uygun hale getirmişlerdir.



4.2.2 HMM ile Birimleri Düzgünleştirme

- **HMM tabanlı sistemlerin** önemli bir özelliği de **düzgün yörüngeler yaratabilme** kabiliyetidir.
- Bu yöntem **birimleri düzgünleştirmek** için kullanılabilir.
- Bir dezavantaj
 - HMM iskeletine uymak için **parametrik konuşma** modeli benimsenmiştir. Kalitede kayıp.



4.2.2 HMM ile Birimleri Düzgünleştirme

- Plumpe ve arkadaşları, düzgünleştirme algoritmalarını seçilmiş birimlerin **LSP parametreleri** için **kaynak parametreleri değiştirmeden** uygulamışlardır.
- Verilen bir cümle HMM'i için ML çözümü aşağıdaki denklemin minimize edildiği haline indirgenebilir. (köşegen eşdeğişke(covariance) matrisi ile, i.i.d varsayımı altında)

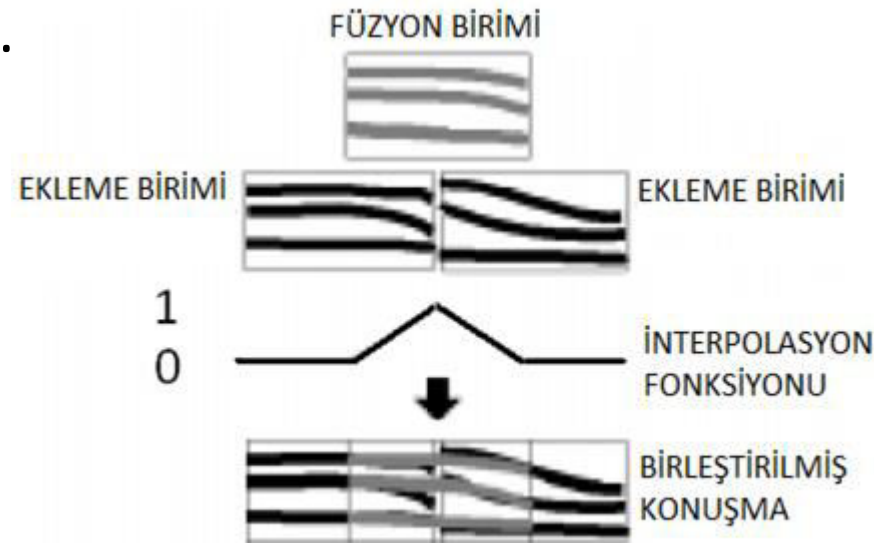
$$E = \sum_{p=1}^P \sum_{t=1}^T \frac{(x_p(t) - \mu_p(t))^2}{\sigma_p(t)^2} + D \frac{(x_p(t) - x_p(t-1) - \Delta\mu_p(t))^2}{\Delta\sigma_p(t)^2} \quad \begin{matrix} D \text{ MLPG'de} \\ 1'dir. \end{matrix}$$

- Fikirleri, eklemeli konuşma sentezinde bulunan doğallığı korumak amacıyla, **HMM ortalamaları yerine μ için gerçek konuşma bölüntülerinin** LSP vektörlerini kullanmaktır.
- Düzgünleştirme seviyesi D parametresinin ayarlanmasıyla belirlenebilir.



4.2.2 HMM ile Birimleri Düzgünleştirme

- Wouters ve Macon benzer bir HMM formülizasyonunu **birim füzyonu** için kullanmıştır(ilave interpolasyon ile). İki birim tipi öne sürülmüştür.
- **Ekleme birimleri** iki ses büyüklüğündedir ve anlamsız kelimelerden toplanmıştır. **Füzyon birimleri** bir ses büyüklüğündedir ve normal sözcelerden(utterance) elde edilmiştir.



4.2.3 HMM güdümlü sistemler

- Gelenegın aksine, bu tip sistemler HTTS sistemlerini geliřtirmek adına konuřmanın doęal bölüntülerini veya doęal parametrelerini kullanırlar.



4.2.3 HMM güdümlü sistemler

- Raitio ve arkadaşları, **sentezlemede** HTTS sistemlerindeki kaliteyi geliştirmek amacıyla sentetik uyarılmaya **en yakın ön depolanmış uyarılma sinyali**ni kullanmışlardır.
- HMM'leri şu altı parametre ile eğitmişlerdir.
 - F0, enerji, Harmoniklerin gürültüye oranı, Harmonik büyüklükleri, kaynak spektrumu, ses yolu spektrumu
- Ayrıca gırtlaksı kaynak vurumlarından veri tabanı yaratmışlardır ve her bir vurumu bu 6 parametre ile ilişkilendirmişlerdir.
- Sentezleme boyunca; cümle HMM'de her bir HMM için en iyi gırtlaksı kaynak vurumu, hedef(LL) ve ekleme masraflarının(RMS) ortak masrafını en aza indirerek seçilir.
- Sentezlenen konuşma, **seçilmiş kaynak vurumlarını** ve **HMM tahminli vokal ses yolu** filtresi katsayılarını kullanan lpc ses kodlayıcı tarafından elde edilir.



4.2.3 HMM güdümlü sistemler

- Gonzalvo ve arkadaşları HTTS eğitmişlerdir. Sentezleme aşamasında, HMM tabanlı birim seçim modülü, en uygun birim dizisini verilen metin için belirler.
- Diğer bir modül **MLPG** kullanarak konuşma parametrelerini yaratır.
- **HMM**'den üretilen parametre ve **en uygun birim dizisi** arasındaki hatayı en aza indirmek amacıyla **Bölgesel Minimum Üretim Hatası'nı** (LMGE) öne sürmüşlerdir. **Bunlar DTW ile uygunluk gösterir.**
- Hizalamadan sonra, HMM modelinin ortalaması ve değişimi öne sürülen **yinelemeli algoritma** ile güncellenir. Sonunda **MLPG** güncellenmiş modeller ile çalıştırılır.



4.2.4 Bölüntüleri karıştırma

- Bu tarz sistemler seyreklik problemlerini çözmeyi önerirler.
- Öne sürülen algoritmalar adreslemeyi amaçlasa bile doğal ve sentetik konuşma arasındaki değişim halen problemdir.



4.2.4 Bölüntüleri karıştırma

- Aylett ve Yamagishi **doğal birim veri tabanına yardım etmek için HMM tabanlı konuşma sentezinden yaratılan yardımcı bir birim veri tabanı** kullanmayı önermişlerdir.
- Veri seyrek ve ekleme maliyetleri yüksek olduğunda doğal olana ek olarak bu sentetik veri tabanını kullanmışlardır.
- Sonra Cereproc's sistemlerinde, **bu karıştırılan birimler hatasız bir şekilde** sistemlerine eklenmiştir. (ticari ürün)



4.2.4 Bölüntüleri karıştırma

- Pollet ve Breen Çok Biçimli Bölüntü (**MFS**) sentezi sistemlerini tanımlamışlardır. Bir **MFS** dizisi, sırasıyla **orijinal** ve **HMM ile üretilmiş birimler** olarak sayılabilecek, **taslak** ve **model** bölüntülerinin karma dizisidir.
- Amaç **model bölüntülerinin insan algısının düşük kaliteye duyarlı olmadığı konuşma parçalarında kullanımıdır**. Seçimi otomatize etmek adına bu tarz yerlerde model bölüntüleri için **düşük maliyet** kullanılır.
- İlk olarak, iki dizi yaratılır: **en uygun taslak** dizi ve **en uygun model** dizi. **En uygun MFS** dizisi **viterbi** ile belirlenir.
- Konuşma **ekleme** ile yaratılır.



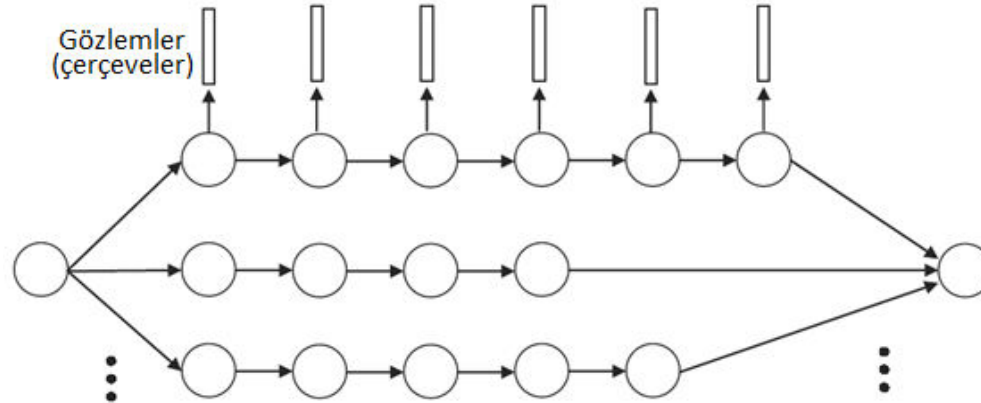
4.2.5 Birleştirme

- Taylor iki yöntemin birleştirilmesini araştırmıştır.
- **HMM tabanlı** ve **birim seçimi tabanlı** sentezleyiciler için ortak taban yaratmak amacıyla çok genel bir HMM iskeleti tanımlamıştır.
- Birleştirme adına
 - Birim seçimi sistemini bir HMM iskeletinde sunabilmek amacıyla, birimler için mükemmel konuşma dalga formları yerine parametrik konuşma modelleri kullanılır.
 - Veri tabanında aynı içeriğe sahip birimler için **birim-HMM ağını** tanımlamıştır. Bu yüzden bütün veri tabanı **her bir içerik** için bir **birim-HMM ağı** olacak şekilde kümelenmiştir.



4.2.5 Birleştirme

- Sonra, $1/N$ olan ilk durumun kolunun olasılıklarının haricinde bütün geçiş olasılıkları 1'e eşitlenir. Her bir birim-HMM ağı yalnızca **veriyi** bilgi kaybı olmadan **hatırlar**. Bu ilk biçimde, saf olarak **birim seçimi** iskeletini temsil eder.



- Çerçeveleri farklı **topolojiler** altında **kümeleyerek**, HMM ağının herhangi bir seviyesi bu saf **birim seçimi** ağından üretilebilir.



İÇERİK

1. Giriş
2. Birim Seçimi tabanlı TTS
3. HMM tabanlı TTS
4. Melez sistemler
 1. UTTS ve HTSS karşılaştırması
 2. Melez yöntemler
5. Önerilen melez sistem
 1. Temel HTTS sistemi
 2. Melez tasarım– Genel değerlendirme
 3. Morfem Seçim Algoritmaları
6. Deneyler ve sonuçlar



5. ÖNERİLEN MELEZ SİSTEM

- Literatürdeki çoğu melez sistem, temel UTTS sistemlerinin kalitesini HTTS tekniklerinin yardımıyla geliştirmeye odaklanmıştır.
- Kendi melez sistemimizde, amaç **morfolojik** dillerin avantajlarından yararlanarak; **HTTS sistemlerinin kalitesini geliştirerekten** ses veri tabanı boyutunu **küçük** tutmaktır.
- Fikir Türkçe'ye uygulanmasına rağmen herhangi morfolojik olarak zengin bir dil için kullanılabilir.



5.1 TEMEL HTTPS SİSTEMİ

- Bir HTTPS sistemi, önerilen melez tasarım için; **HTS** ile, **temel sistem** olarak kullanılmak üzere eğitilmiştir. (HTS: <http://hts.sp.nitech.ac.jp/>)
- Peş peşe birçok metin işleme algoritması **Normalleyici** modülü içinde kaydedilmiş konuşmayı yazıya çevirme için kullanılmıştır. Bu modül numaralar, günler, noktalama işaretleri vs. meselesini adresler.
- Telaffuz Etme Modeli '**CART**'(**Classification and Regression Tree**) tarafından ele alınmıştır. Vurgu modeli içinse harici bir araç kullanılmıştır.
- **İçerikEtiketleyici(ContextLabeler)** her bir cümle için etiket dosyası yaratır. Etiket her bir satırı sözcedeki **bir fonemin içerik bilgisine** karşılık gelir.



5.1.1 ETİKET YAPISI

- Etiketlerde kullanılan bilgiler.

İçerik	Bilgi
Fonem seviyesi	İki önceki, güncel ve iki sonraki fonemler; güncel hecedeki pozisyon(ileri, geri); ekleme bilgisi
Hece seviyesi	Önceki, güncel, sonraki hecelerdeki fonem sayısı. Güncel kelimedeki pozisyon(ileri, geri). Vurgu bayrağı. Vurgulanan heceye(önceki, sonraki) olan mesafe.
Kelime seviyesi	Kelimelerdeki(önceki, güncel, sonraki) hecelerin sayısı. Güncel deyişteki sıra.
Deyiş seviyesi	Hecelerin, kelimelerin deyişteki(önceki, güncel, sonraki) sayısı. Güncel deyişin sözcedeki sırası.
Sözce seviyesi	Hecelerin, kelimelerin, deyişlerin sözcedeki sayısı. Cümle tipi için bir bayrak



5.1.2 TELAFFUZ VE VURGU MODELLEME MODÜLÜ

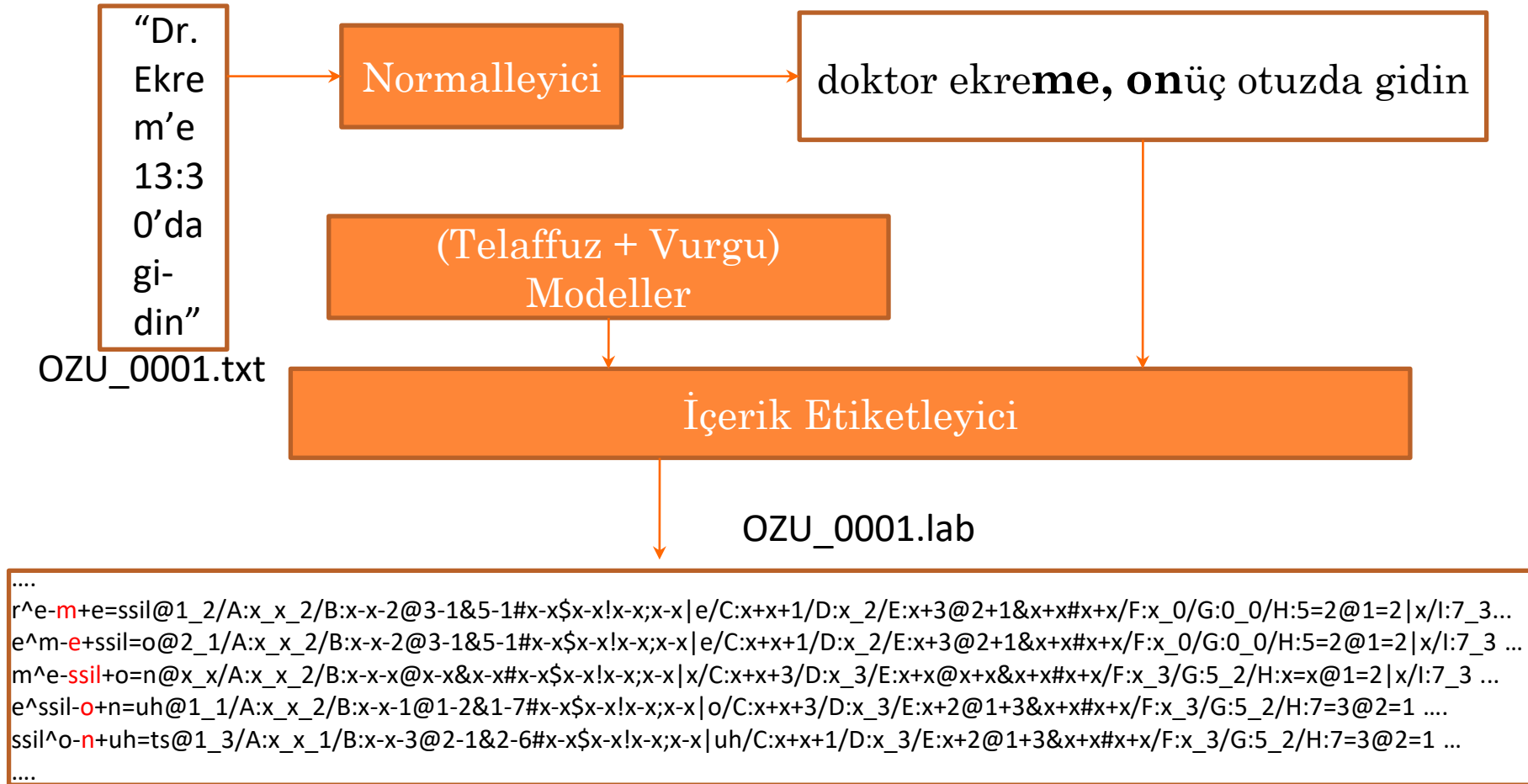
- Türkçe'nin **G2P** (grapheme-to-phoneme) eşleşmelerini modellemek için bir telaffuz sözlükçesi(lexicon) kullanarak CART modeli eğitilmiştir. Sonra, G2P çevirmeleri **CART** kullanarak yapılmıştır.
- **Vurgu** modelleme için, harici bir araç kullanılmıştır. Bu melez sistemlerin **morfem analizi** için kullanılan araç ile aynıdır.

(K. Oflazer and S. Inkelas, "A finite state pronunciation lexicon for Turkish," in Proceedings of the EACL Workshop on Finite State Methods in NLP, Budapest, Hungary, vol. 82, pp. 900–918, 2003)



5.1 TEMEL HTTS SİSTEMİ

- Cümleden etikete



5.1 TEMEL HTTS SİSTEMİ

- **Deneyler**
- Sistem; 24 LSF, 1 log F0 katsayısı ve 5 ses gücü parametreleri ile 30 boyutlu vektörler tarafından eğitilmiştir. Ses güçleri, 0 ile 8000 Hz arasında eşit aralıklı 5 spektral bant için normalleştirilmiş oto-korelasyon kullanarak hesaplanmıştır.
- Test için 37 cümle haber alanından, 66 cümle roman alanından seçilmiştir.
- 17 dinleyici MOS kalite testini yapmıştır
 - Ortalama 3.27, medyan 3
 - Sapma 1



İÇERİK

1. Giriş
2. Birim Seçimi tabanlı TTS
3. HMM tabanlı TTS
4. Melez sistemler
 1. UTTS ve HTSS karşılaştırması
 2. Melez yöntemler
5. Önerilen melez sistem
 1. Temel HTTS sistemi
 2. Melez tasarım– Genel değerlendirme
 3. Morfem Seçim Algoritmaları
6. Deneyler ve sonuçlar



5.2 MELEZ TASARIM – GENEL DEĞERLENDİRME

- Morfolojik olarak zengin dillerde, birçok farklı kelime kısıtlı **morfem** seti kullanarak aynı kökten üretilebilir.
- Önerilen sistemde **en çok tekrar eden morfemlerin veri tabanı** eğitimde yaratılmıştır.
- Sentezde, **en iyi oturan morfemler** önerilen **morfem seçimi** algoritmasını kullanarak seçilmiştir.
- Sonra; **seçilen morfemler, önerilen parametre üretimi** algoritmasıyla **parametre üretimi** aşamasını **geliştirmek** amacıyla HTTS'de kullanılmıştır.



5.2.1 MORFOLOJİK ANALİZ

- “*kazanabileceğini*” sözcüğünün morfolojik analizi

kazanabileceğini (k a z a n)kazan+Verb+

Pos(a b i l)^DB+

Verb+Able(e dZ e)^DB+

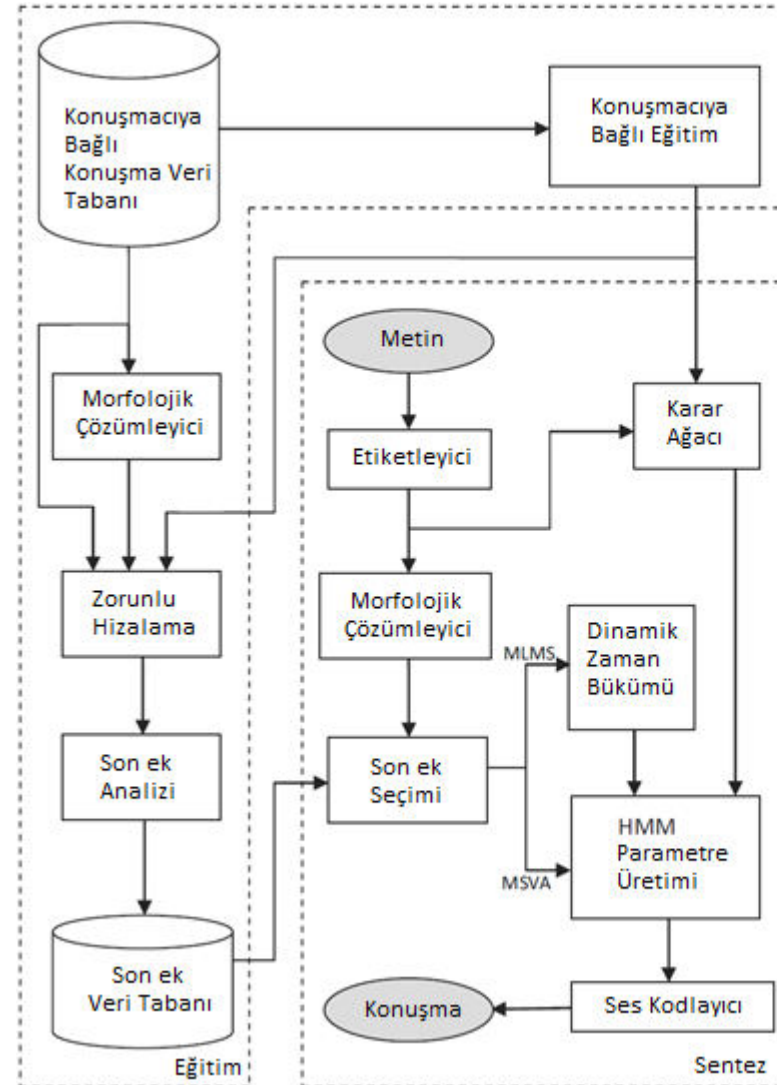
Noun+FutPart+A3sg(G i)+

P3sg(“n i)+Acc

- Sondan ekler/Morfemler: abil, ece, ği, ni



5.2 MELEZ TASARIM – GENEL DEĞERLENDİRME



5.2.1 MORFEM SEÇİM ALGORİTMALARI

- Bu çalışmada; ilk olarak, **maksimum olabilirlik tabanlı hedef maliyeti, seçim için hiçbir ekleme maliyeti** olmadan kullanıldığında sistemin davranışı incelenmiştir. **(MLMS)**.
- MLMS yönteminin test sonuçları analiz edildikten sonra, çıkarılan dersleri kullanarak, diğer bir morfem seçim algoritması geliştirilmiştir **(MSVA)**. Yeni algoritmada ML tabanlı hedef maliyetleri yerine **ses perdesi ekleme maliyetleri** kullanılmıştır. **Hedef maliyetleri, karar ağacı tabanlı morfem filtreleme algoritması ile değiştirilmiştir.** Filtreleme aşamasından sağ kalan morfemler daha sonra Viterbi algoritması ile seçilir.



5.3.1 MAKSIMUM-OLABİLİRLİK TABANLI MORFEM SEÇİMİ (MLMS)

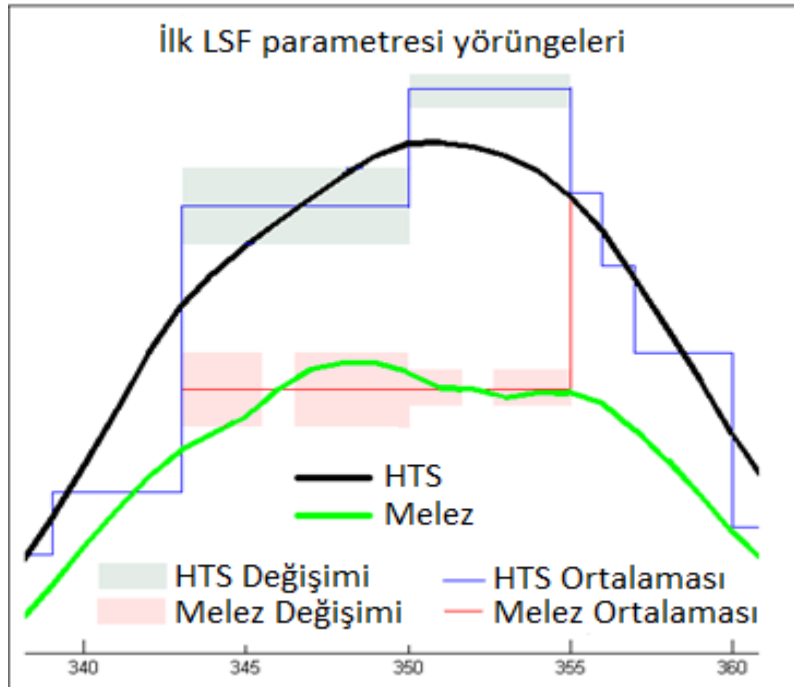
- **w-ML** ölçüsü, **ekleme etrafındaki durumların** ağırlıklarını daha fazla yaparak **ekleme noktalarını düzgünleştirmeye** yardımcı olur.
- LSF ve LF0 için olan hedefler **bağımsız** olarak seçilir.
- **LSF Hedef Seçimi**
 - En fazla **w-ML’li** birim o son ek için hedef olarak seçilir.



5.3.1 MAKSIMUM-OLABİLİRLİK TABANLI MORFEM SEÇİMİ (MLMS)

• LSF Parametre Üretimi

- Bir son ekin **her bir durumunun ortalaması hedefteki en yakın çerçeve** ile değiştirilir. Mesafe log olabilirlik ile ölçülür. (Eğer log-olabilirlik değeri eşik değerden fazla ise)



- Üretilen yörüngenin ortalamadan geçeceğini garanti etmek için, her bir durumun orta çerçevesi **Ö**'ya eşitlenir.

5.3.1 MAKSIMUM-OLABİLİRLİK TABANLI MORFEM SEÇİMİ (MLMS)

- MLMS test sonuçlarını analiz ettikten sonra, şu gözlemler yapılmıştır
 - Olabilirlik tabanlı hedef seçimi, **değişimi** ve dolayısıyla **doğallığı** düşüren düzgün yörüngeleri **önemli derecede tercih etmiştir.**
 - Ses perdesi ve LSF parametreleri için çeşitli morfemler seçilmiştir. Fakat, bu ikisi arasındaki eşleşmezlik kalitede **önemli derecede azalmaya neden olmuştur.**
 - Vurgulanmış heceler boyunca dinleme testlerinde gelişim algılanmıştır. Bu yüzden, **vurgulanmış heceler boyunca hızlı ses perdesi değişimlerini iyi modelleyebilen melez sistemlerin tasarımı önemli bulunmuştur.**



5.3.2 VITERBI ALGORİTNASINI KULLANARAK MORFEM SEÇİMİ (MSVA)

- ML tabanlı hedef maliyetleri düzgün yörüngeleri tercih ettiğinden, ses perdesi değişimleri dinleyici terciğinde en büyük etkiyi yaratmıştır, **MSVA** algoritması **sadece ses perdesi yörüngelerinin ekleme maliyetlerine odaklanmıştır.**
- **Hedef maliyetleri karar ağacı tabanlı morfem filtreleme algoritması ile değiştirilmiştir.** Filtreleme aşamasından sağ kalan morfemler daha sonra Viterbi algoritmasını kullanarak seçilmiştir.
- Ses perdesi konturlarının seçildiği **aynı bölüntülerin** LSF konturlarını kullanmanın iyi performans verdiği bulunmuştur. Böylece, ses perdesi ve LSF konturları MSVA yöntemindeki **aynı morfemden** elde edilmiştir.



5.3.2 VITERBİ ALGORİTMASINI KULLANARAK MORFEM SEÇİMİ (MSVA)

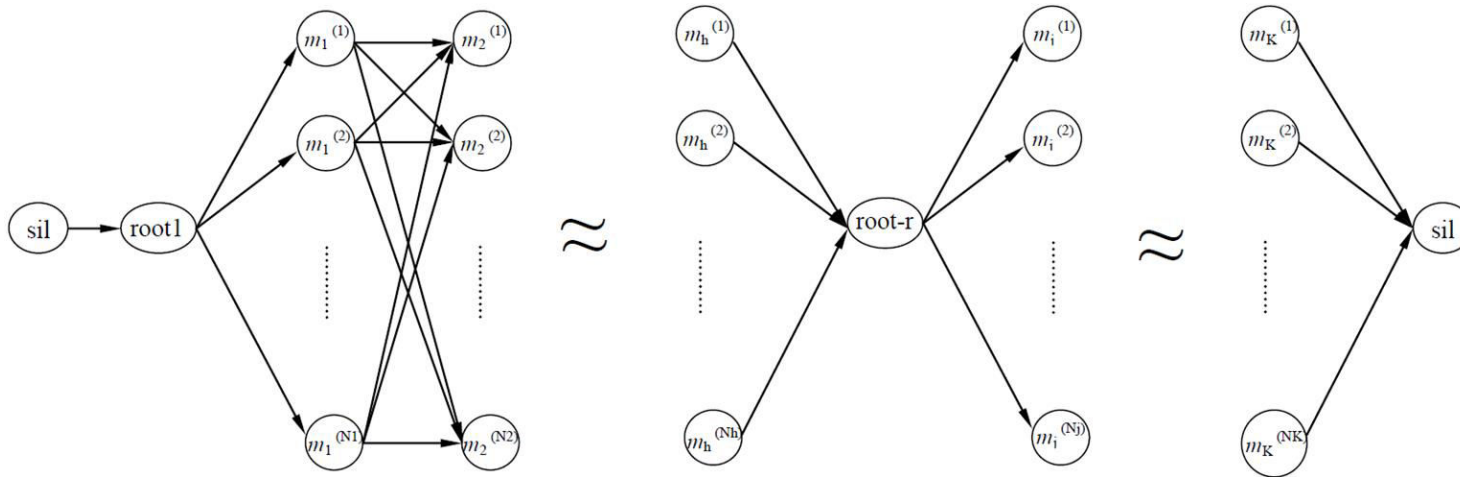
- **Karar ağacı tabanlı ön filtreleme**
 - Viterbi'deki **ağın boyutunu küçültmek** ve **daha içeriksel eşleşme** için her bir morfemin seçiminde küçük karar ağacı kullanılmıştır. Ağaçtan döndürülen morfem kümesi hesaplamalarda kullanılmıştır.
 - Sorular ağacın yaratılmasında kullanılmıştır

Seviye	Soru
Hece Seviyesi	Vurgu: Morfemi içeren hecenin vurgu seviyesi nedir?
Kelime Seviyesi	Kelimedeki sıra: Morfem kelimenin sonunda mı?
Değiş Seviyesi	Değişteki sıra: Morfemi içeren kelime değişin sonunda mı?



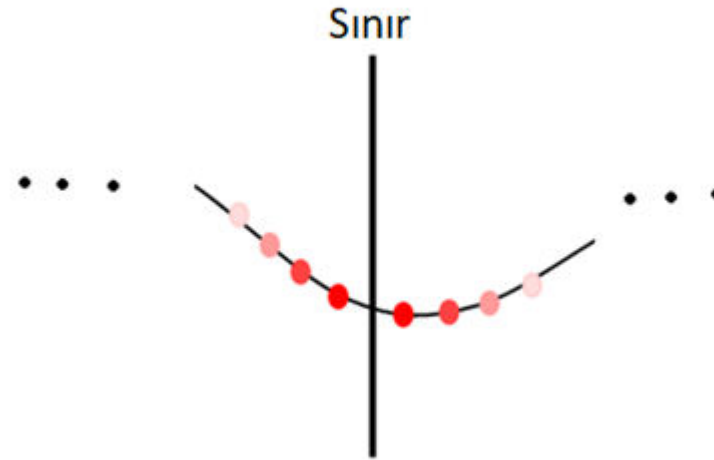
5.3.2 VITERBI ALGORİTMASINI KULLANARAK MORFEM SEÇİMİ (MSVA)

- MLMS yönteminde, ekleme dikkate alınmadığından bütün son ekler bağımsız olarak seçilir.
- MSVA'da, **ekleme** dikkate alınır. Bu yüzden, son ekler **grafik** seçiminden seçilmelidir (viterbi).
- **Kök kelime** için **alternatif yoktur**, fakat **son ekler** için morfem veri tabanında çok sayıda aday vardır.



5.3.2 VITERBI ALGORİTMASINI KULLANARAK MORFEM SEÇİMİ (MSVA)

- En iyi morfem dizisi, Viterbi algoritmasıyla, grafikten **ses perdesinin toplam ekleme maliyetini** en aza indirerek seçilir.
- Ekleme maliyeti **ağırlıklandırılmış Öklid mesafesi** olarak hesaplanır.



- Sınırdaki gerektiği zaman lineer interpolasyon.



İÇERİK

1. Giriş
2. Birim Seçimi tabanlı TTS
3. HMM tabanlı TTS
4. Melez sistemler
 1. UTTS ve HTSS karşılaştırması
 2. Melez yöntemler
5. Önerilen melez sistem
 1. Temel HTTS sistemi
 2. Melez tasarım– Genel değerlendirme
 3. Morfem Seçim Algoritmaları
6. Deneyler ve sonuçlar



6. DENEYLER VE SONUÇLAR

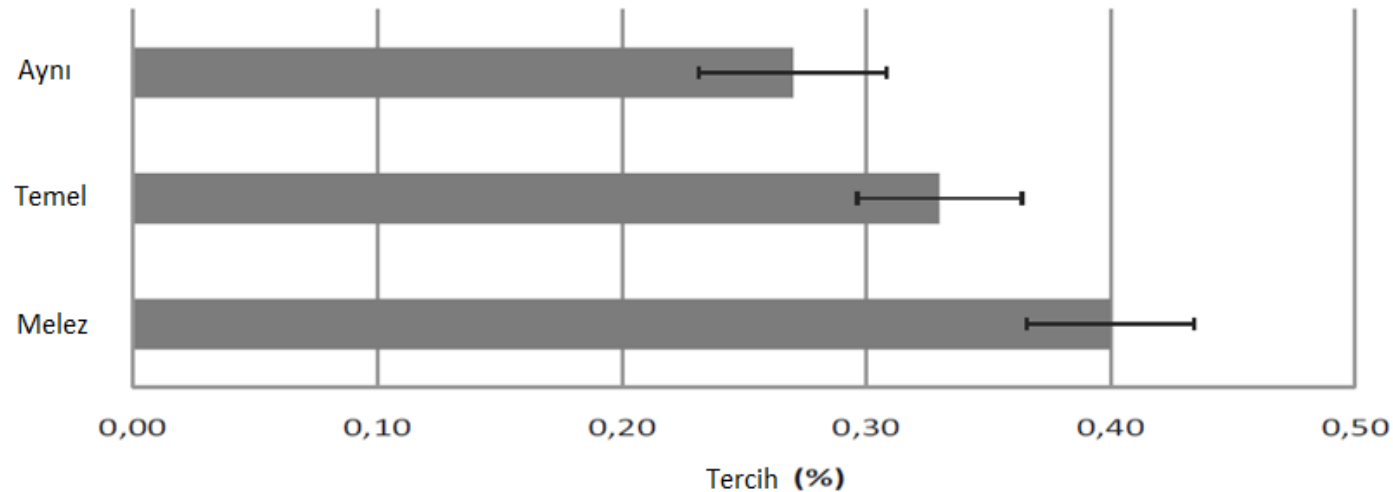
- Bütün sistemler; 24 LSF, 1 log F0 katsayısı ve 5 ses gücü parametrelerini içeren 30 boyutlu vektörler ile eğitilmiştir. Ses güçleri, 0 ile 8000 Hz arasında eşit aralıklı 5 spektral bant için normalleştirilmiş oto-korelasyon kullanarak hesaplanmıştır.



6. DENEYLER VE SONUÇLAR

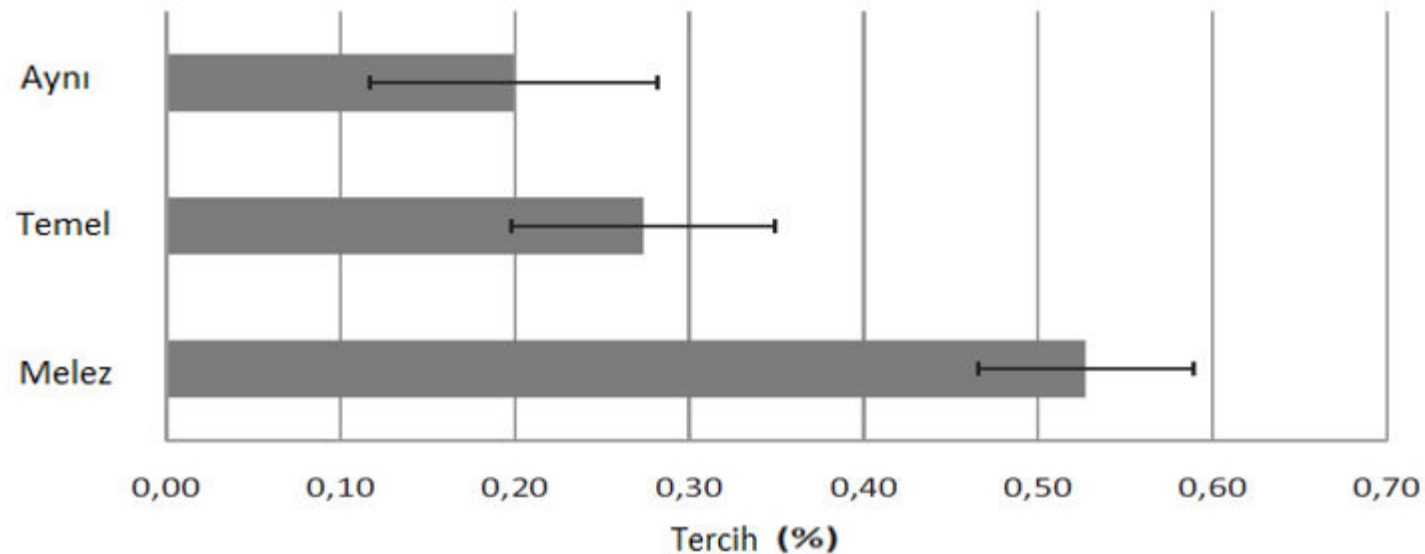
- **MLMS Algoritması performansı**

- AB tercih testi yapılmıştır
- Testte 50 test cümlesi kullanılmıştır
- 17 insan testi yapmıştır
- Tercih çarpıcı değildir (**pearson testi**, önemli alanda az olarak).



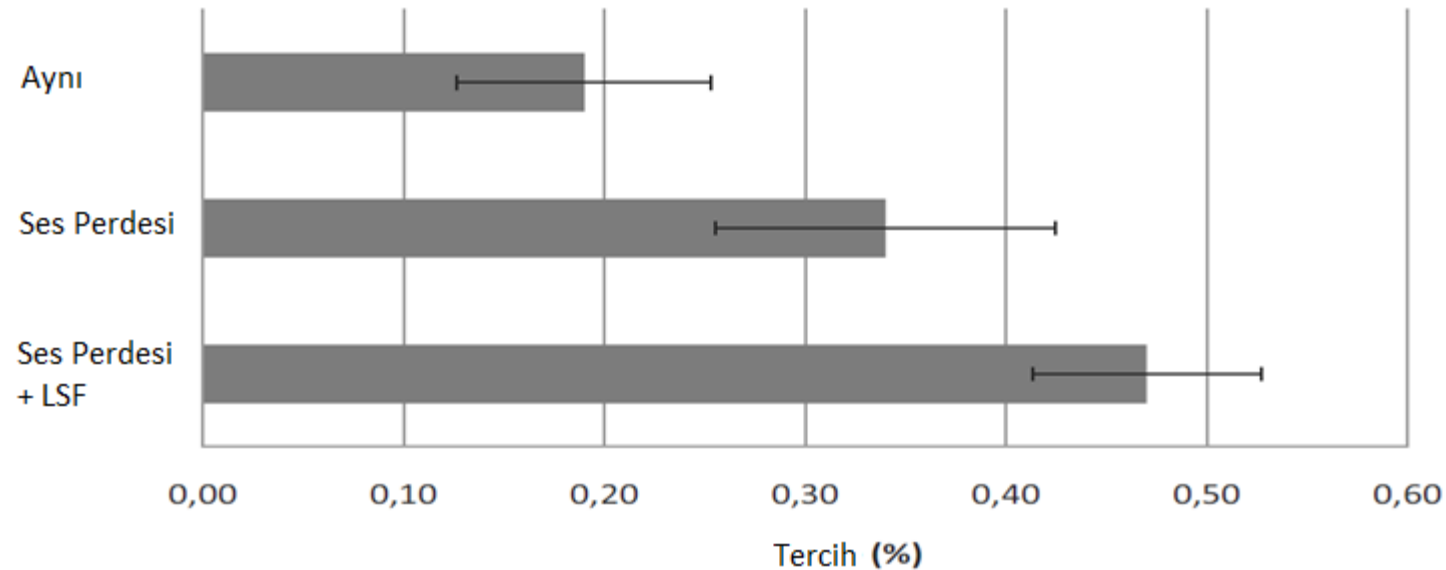
6. DENEYLER VE SONUÇLAR

- **MSVA Algoritmasının Performansı – Sadece Ses Perdesi**
 - AB tercih testi yapılmıştır
 - Testte 30 test cümlesi kullanılmıştır
 - 10 insan testi yapmıştır
 - Önemli bir tercih vardır



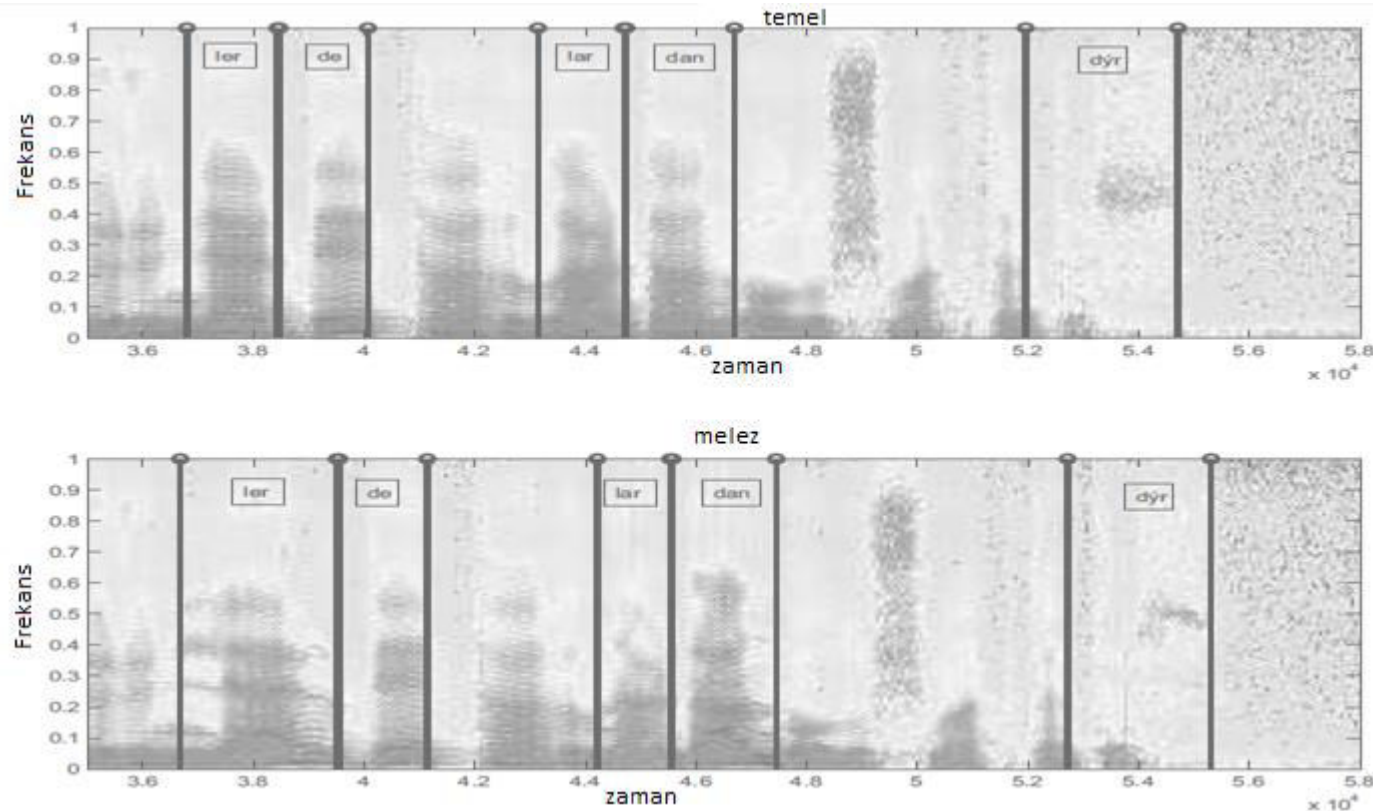
6. DENEYLER VE SONUÇLAR

- **MSVA Algoritmasının Performansı – (Ses Perdesi + LSF)**
 - AB tercih testi yapılmıştır
 - Testte 30 test cümlesi kullanılmıştır
 - 10 insan testi yapmıştır
 - Sadece ses perdesi olandan iyi



6. DENEYLER VE SONUÇLAR

- **MSVA Algoritmasının Performansı**
 - Melez MSVA yöntemi ve temel sistemler için spektrogramların karşılaştırması. Morfemler ve sınırları figürde belirtilmiştir.



6. DENEYLER VE SONUÇLAR










- **Sonuç**

- MSVA algoritması ses perdesi değişimlerini MLMS algoritmasından daha iyi modellemekle kalmayıp MSVA yöntemiyle seçilmiş **morfemlerin LSF yörüngeleri** morfem içeriğine daha iyi uyum sağlamıştır ve kaliteyi geliştirmeye yarayan daha az düzgünleştirmeyi gerektirmiştir.
- Temel HTTS sistemi performansı çoğunlukla **konuşma parametrelerindeki daha fazla değişimlerin** yardımıyla geliştirilmiştir ve bu çoğunlukla **ses perdesi içeriğinde** önemlidir.



6. DENEYLER VE SONUÇLAR

- Örnekler

	Temel	MSVA (sadece ses perdesi)	MSVA (ses perdesi + Isf)
Cümle-1			
Cümle-2			
Cümle-3			



KONUŐMACIYA UYARLANMA



ÖNCEKİ ÇALIŞMALAR

- İki yaygın konuşmacı adaptasyonu yöntemi:
 - Lineer regresyon tabanlı adaptasyon
 - Kısıtlandırılmış maksimum olabilirlik lineer regresyonu (CMLLR)[Gales, 1998]
 - Kısıtlandırılmış yapısal maksimum sonsal lineer regresyonu (CSMAPLR) [Yamagishi, 2009]
 - Özses(Eigenvoice) tabanlı adaptasyon
 - SSS için özsesler [Shichiri, 2002]
 - Küme adaptasyonu eğitimi (CAT) [Gales, 2000]



LİNEER REGRESYON TABANLI ADAPTASYONLAR

- CMLLR
 - Aşırı uyumlamaya(overfitting) eğilimlidir
 - CMLLR durumunda eşlenik önsel yoktur.
 - Düzgünleştirme sayım teknikleri kullanılabilir [Breslin, 2010]
- CSMAPLR
 - Küçük adaptasyon verisi boyutları için daha güvenilirdir.
 - Çünkü sonsal dağılım kullanır.



DÜZGÜNLEŞTİRME SAYIM TEKNİĞİ

- Lineer regresyon dönüşümü hesaplaması için gereken istatistikler düzgünleştirilir.
- Düzgünleştirme, CMLLR ve aşağıdakilerden bir tanesinin interpolasyonunu kullanarak yapılır:
 - Ses yolu uzunluğu normalleştirilmesi (VTLN)
 - Tahminsel CMLLR(PCMLLR)



İNTERPOLASYON

- Farklı konuşmacılar ve stiller arasında interpolasyon şunun için kullanılmıştır:
 - Önceden eğitilmiş sesler/stillerin ağırlıkları interpolate edilmiş sesin hedefe yakın duyulması için ayarlanmıştır.
- Bu özses yöntemine şunun haricinde benzerdir:
 - Özseslerin aksine, adaptasyon konuşmacı-adapte sesleri interpolate ederek yapılır.
 - Ağırlıklar el ile ayarlanır.



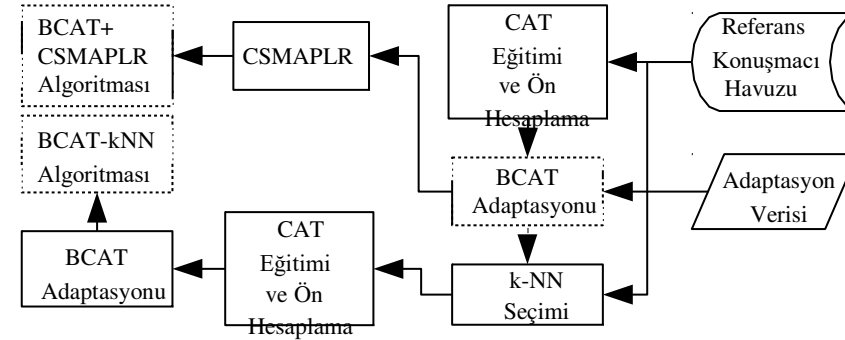
ÖNERİLEN YÖNTEMLER

- Biz şunları yapan 3 yöntem önerdik:
 - Temel özses adaptasyonu algoritması kalite problemlerini düşüren
 - En az veri ile konuşmacı adaptasyonunu daha çok geliştiren
- Bayes CAT (BCAT):
 - Bayes özses yöntemini kullanır
 - Yapaylığı düşürmek ve konuşmacı uzayında daha gerçekçi yönde gitmek için adaptasyon algoritmasını kısıtlandırır.



ÖNERİLEN YÖNTEMLER

- BCAT k-En yakın-Komşular (BCAT k-NN):
 - Hedef konuşmacıya yakın önceden eğitilmiş referans konuşmacıları bulur.
 - Sadece 2. özses adaptasyonu yinlemesindeki referans konuşmacı modellerinden yararlanır.



Figür: Önerilen yöntemlerin genel gözden geçirmesi

ÖNERİLEN YÖNTEMLER

- BCAT algoritması CSMAPLR'ye kıyasla çok daha az sayıda parametreye sahiptir.
- BCAT, sadece önceden belirlenmiş yönlerde kaba taslak adaptasyonlar yapabilir.
- BCAT, daha fazla adaptasyon verisi mevcut olduğunda modeli düzeltemez.
- İkili BCAT / CSMAPLR Yöntemi
 - BCAT ilk önce hedef konuşmacıya hızlı bir şekilde yaklaşmak için kullanılır.
 - BCAT algoritması çıktısı, CSMAPLR algoritması için yeni SI modeli olarak kullanılır.
 - Yeni SI modeli hedef konuşmacıya zaten yakın olduğundan, CSMAPLR ön kısıtları daha az önemli hal alır.
 - CSMAPLR daha fazla parametreler ile final modeli ince ayarlayabilir.



KÜME ADAPTASYONU EĞİTİMİ (CAT)

- Özses yöntemi konuşma tanıma ve SSS'de hızlı adaptasyon için kullanılmıştır.
- Özses yönteminde, R özvektörler seti verilsin $\mathbf{e}_r \in \mathbb{R}^n$, konuşmacı s için orijinal süpervektör

$$\boldsymbol{\mu}^{(s)} = \boldsymbol{\mu}_{si} + \mathbf{E}\mathbf{w}_s + \mathbf{E}_s \quad (1)$$

$\mathbf{E} = [\mathbf{e}_1 \mathbf{e}_2 \dots \mathbf{e}_R]$, \mathbf{w}_s konuşmacı s'nin ağırlık vektörüdür, ve \mathbf{E}_s ise yakınsama hatasıdır.



BCAT K-EN YAKIN-KOMŞU (BCAT K-NN) YÖNTEMİ

- Burada daha iyi konuşmacı adaptasyonu için Σ_w^{-1} ve E hedef özelinde eğitilmiştir.
- Önceki yinelemelerde öğrenilen bilgilerde yinelemeli BCAT
- Hedef özelinde parametreler aşağıdaki gibi eğitilir.
 - İlk BCAT adaptasyonu adımından sonra
 - Sonuç model, k en yakın-komşuları(k-NN) , referans konuşmacıların geniş havuzundan bulmak için kullanılır.
 - Bu komşular daha sonra E matrisi ve Σ ön dağılımının eşdeğişkesini yaratmak için kullanılır.



İKİLİ BCAT/CSMAPLR YÖNTEMİ

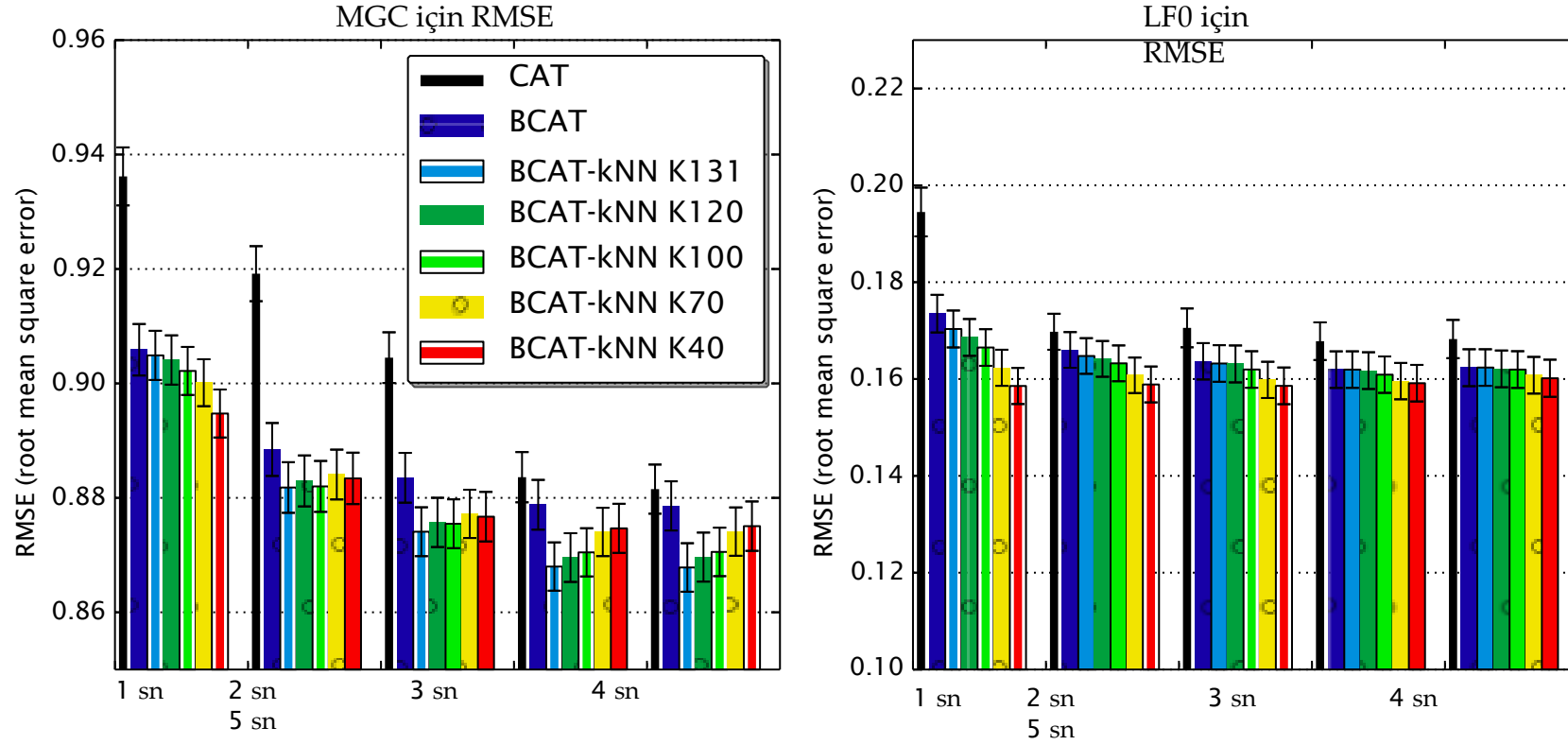
- Sadece verinin saniyeleri mevcut olduğunda:
 - CSMAPLR algoritmasının adaptasyon esnekliği, transformasyon matrislerindeki çok sayıdaki serbest parametrelerinden dolayı düşüktür.
 - Sadece birkaç transformasyon matrisi eğitilebilir.
 - Lakin, BCAT algoritması çok daha az sayıda parametre içerir.
 - BCAT önceden belirlenmiş yönlerde sadece kabataslak adaptasyonlar yapabilir.
 - BCAT daha fazla veri mevcut olduğunda modelleri düzeltemez.



İKİLİ BCAT/CSMAPLR YÖNTEMİ

- İkili yöntemde,
 - İlk önce, hedef konuşmacıya hızlı yaklaşmak için BCAT algoritması kullanılır.
 - BCAT algoritması çıktısı, CSMAPLR algoritması için yeni SI modeli olarak kullanılır. SI modeli hedefe zaten yakın olduğundan, CSMAPLR önseli tarafından uygulanan kısıtlamalar daha az önemli hale gelir.
 - Dahası, CSMAPLR algoritması modeli BCAT modeliyle mümkün olmayacak yönler götürebilir.
 - CSMAPLR , modeli daha fazla veri olduğunda düzeltebilir.





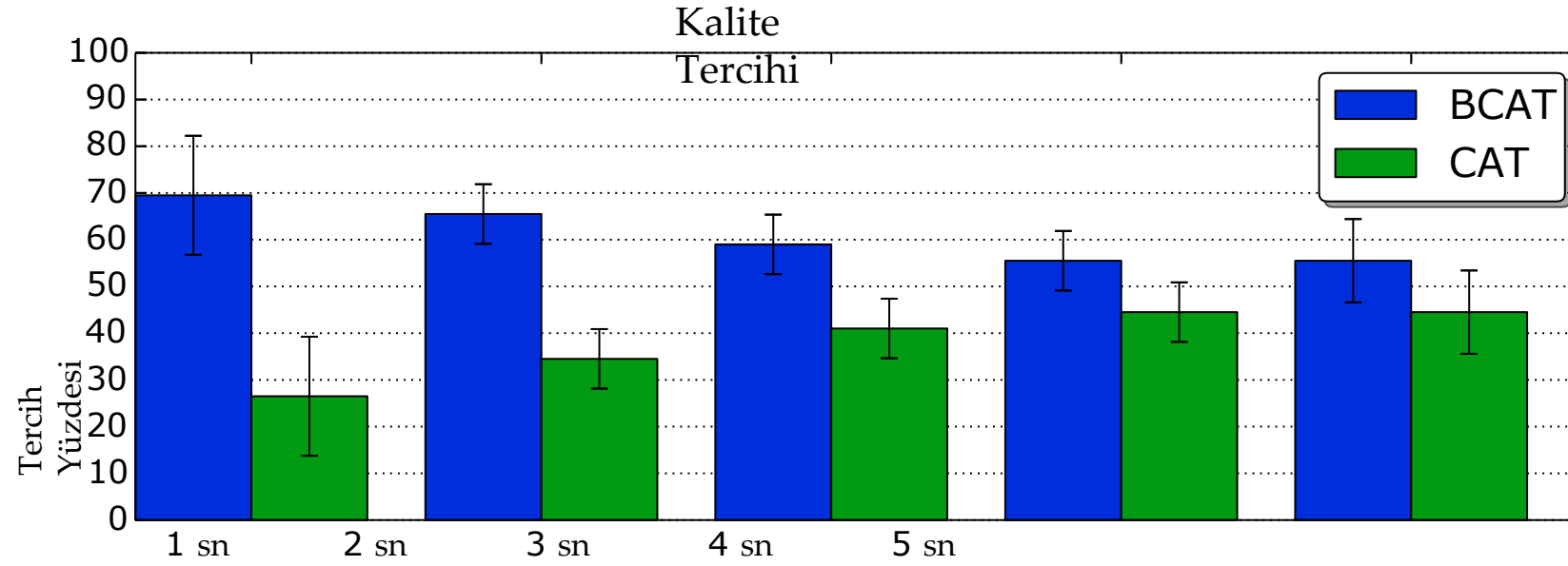
Figür: Farklı k değerleri için BCAT-kNN algoritması performansı. %95 güven aralıklı LF0 ve MGC özellikleri için RMSE sonuçları. Bu sonuçlara dayanarak k MGC için 131'e LF0 için 40'a eşitlenir.



ABX VE AB DİNLEME TESTLERİ

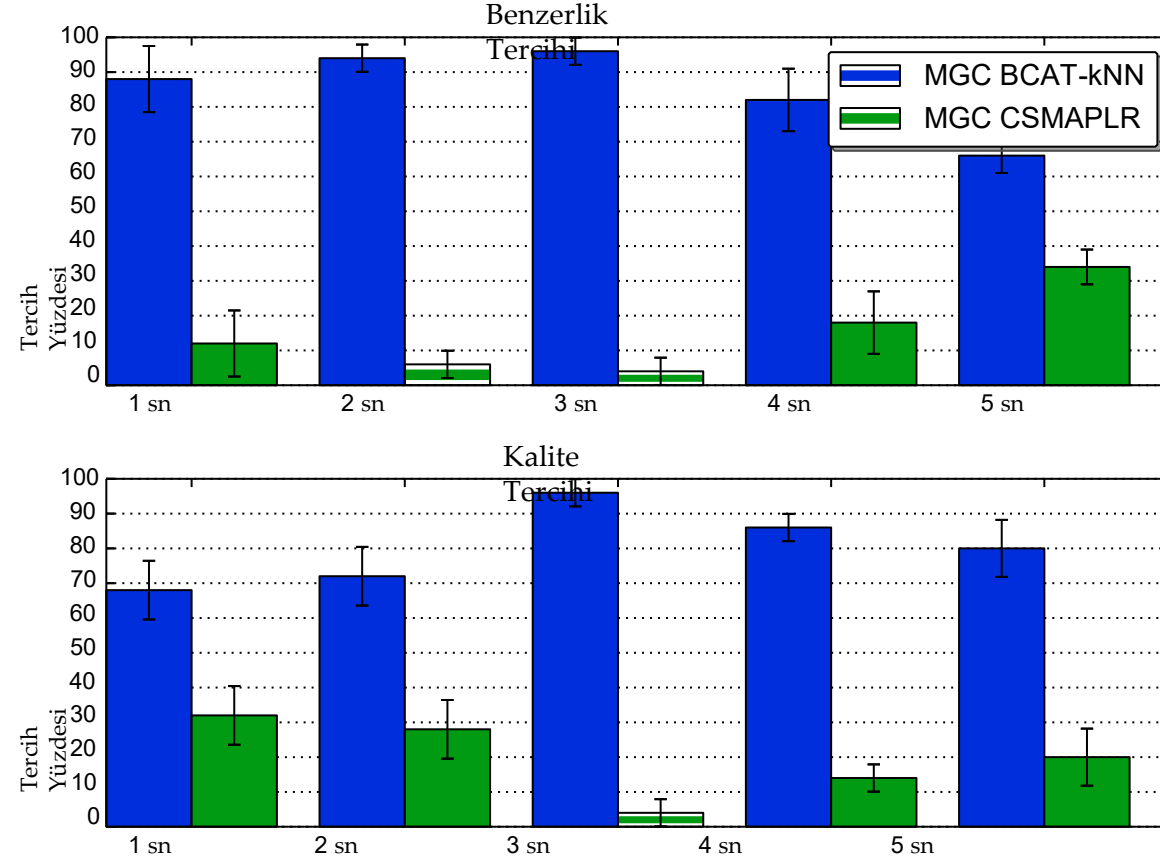
- ABX testi, benzerliği öznel olarak test etmek için kullanılır.
- Dinleyiciler, referans örnek X'e olan algılanmış benzerliğe bağlı olarak A örneğini veya B örneğini tercih etmişlerdir.
- A ve B örnekleri farklı adaptasyon yöntemlerinden rastgele sentezlenmiştir.
- X örnekleri orijinal kaydın "sentetik kopyalarıdır".
- AB testi kalite farklarını ölçmek için yapılmıştır.
- Algılanan konuşma kalitesine bağlı olarak dinleyiciler A veya B örneğini tercih etmişlerdir.
- Her bir adaptasyon verisi büyüklüğü için, bir hedef konuşmacıya bir sözce sentezlenmiştir. 10 konuşmacı kullanılmıştır.
- Sekiz dinleyici testi yapmıştır.





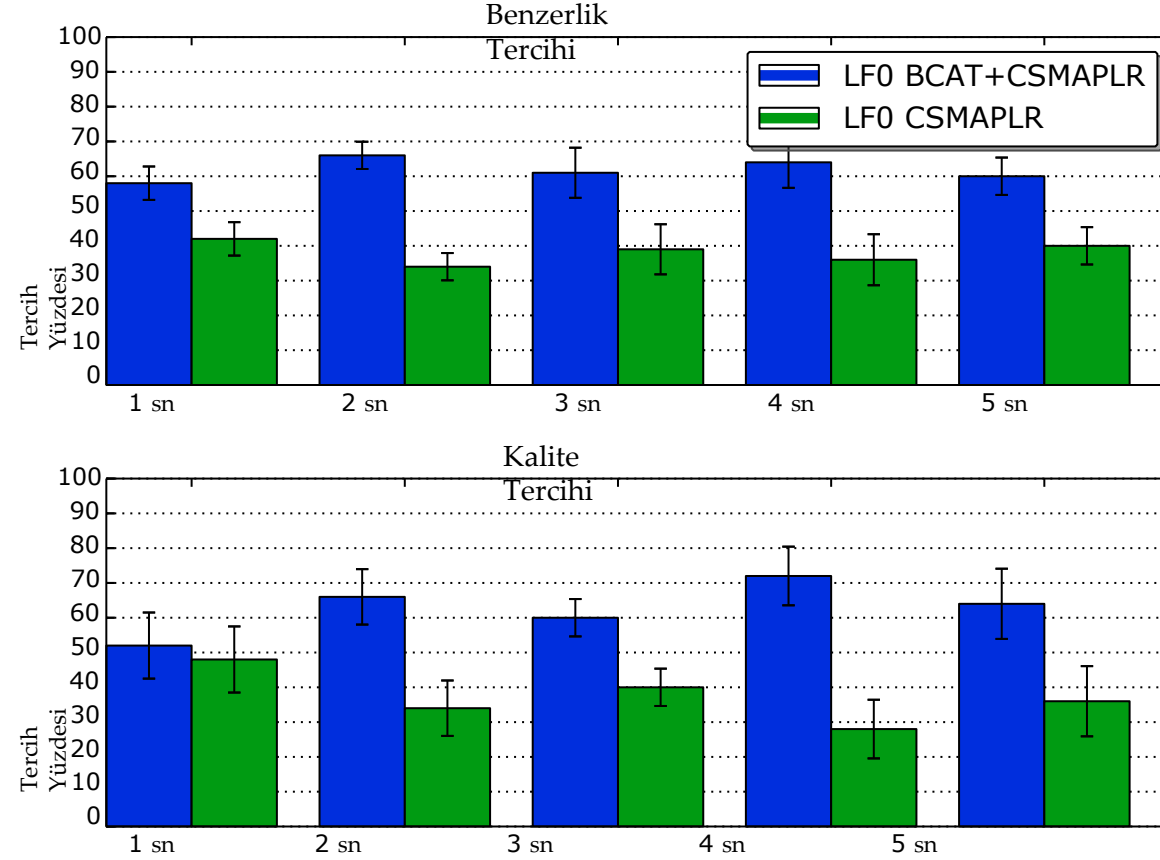
Figür: %95 güven aralıklı MGC ve LF0 özelliklerinin kalitesi açısından öznel AB tercihi testi sonuçları.





Figür: %95 güven aralıklı MGC özellikleri için benzerlik (ABX) ve kalite (AB) açısından öznel tercih testi sonuçları. LF0, BCAT+CSMAPLR'e sabitlenmiştir.





Figür: %95 güven aralıklı LFO özellikleri için benzerlik (ABX) ve kalite (AB) açısından öznel tercih testi sonuçları. MGC, BCAT-kNN'e sabitlenmiştir.



DERİN ÖĞRENEN AĞLARLA AKUSTİK MODELLEME



MERLIN TTS Tool

- DNN-tabanlı konuşma sentezi için açık kaynak kodlu bir yazılım aracı
- Önyüzde bir metin işleme aracına ve bir ses kodlayıcısına ihtiyaç duyar
- Python ile yazılmıştır
- Theano ve Tensorflow destekler
- DNN model repertuarı geniştir
 - Feedforward Networks
 - RNNs(LSTM, GRU)
 - Mixture Density Neural Networks

ÖNE ÇIKAN DERİN SINIR AĞ TABANLI SENTEZ TEKNİKLERİ

- WAVENET: A GENERATIVE MODEL FOR RAW AUDIO (van den Oord et al, 2016)
- Tacotron (Wang et al, 2017)
 - Tacotron 2 (Shen et al, 2017)
- Deep Voice (Arik et al, 2017)
 - Deep Voice 2 (Arik et al, 2017)
 - Deep Voice 3 (Ping et al, 2017)
- Char2Wav (Sotelo et al, 2017)



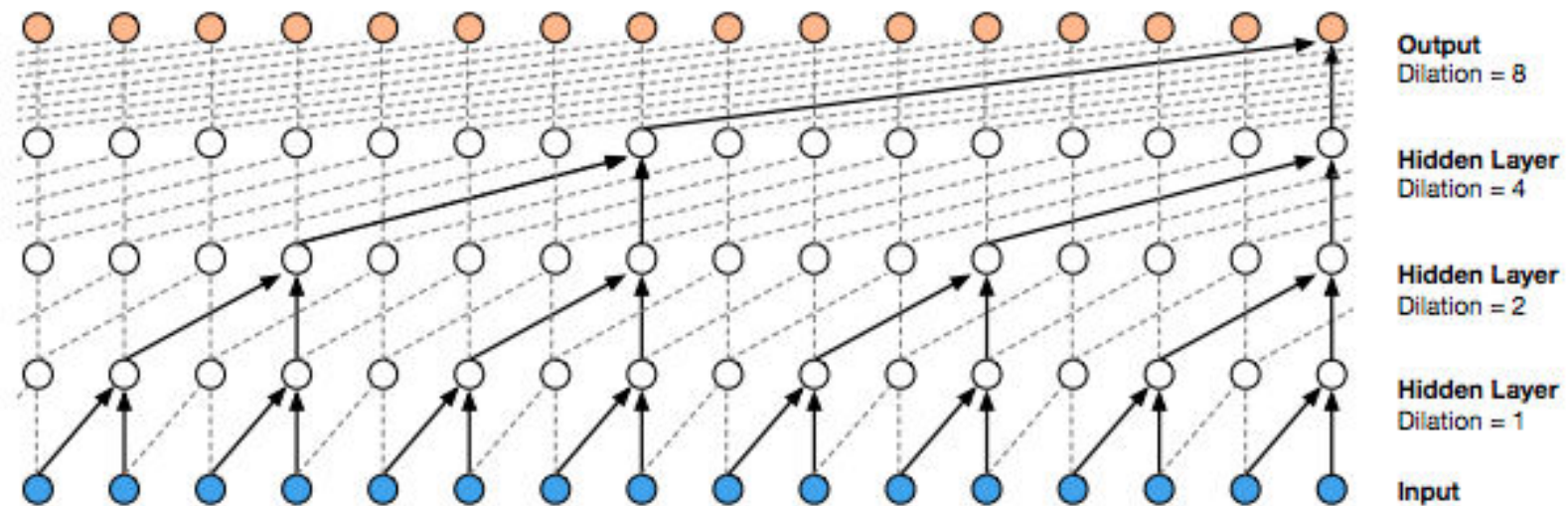
WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

- Wavenet'ten önce, sentez paradigması, önce akustik parametreleri üretip, ardından vocoder(Voice Encoder) yardımıyla ses dalgasını üretmek üzerineydi
- Wavenet doğrudan ses dalgası üretebiliyor
- Doğrudan ses dalgasını modelleyebilmesi müzik gibi diğer sesleri de üretebilmesini mümkün kılıyor
- İmaj üretiminde kullanılan(PixelCNN) oto-regresif yapının sese uyarlanmasından oluşuyor
- Ağ yapısı birkaç residual bağlantı ile birlikte tamamen convolutional ağlar kullanılıyor



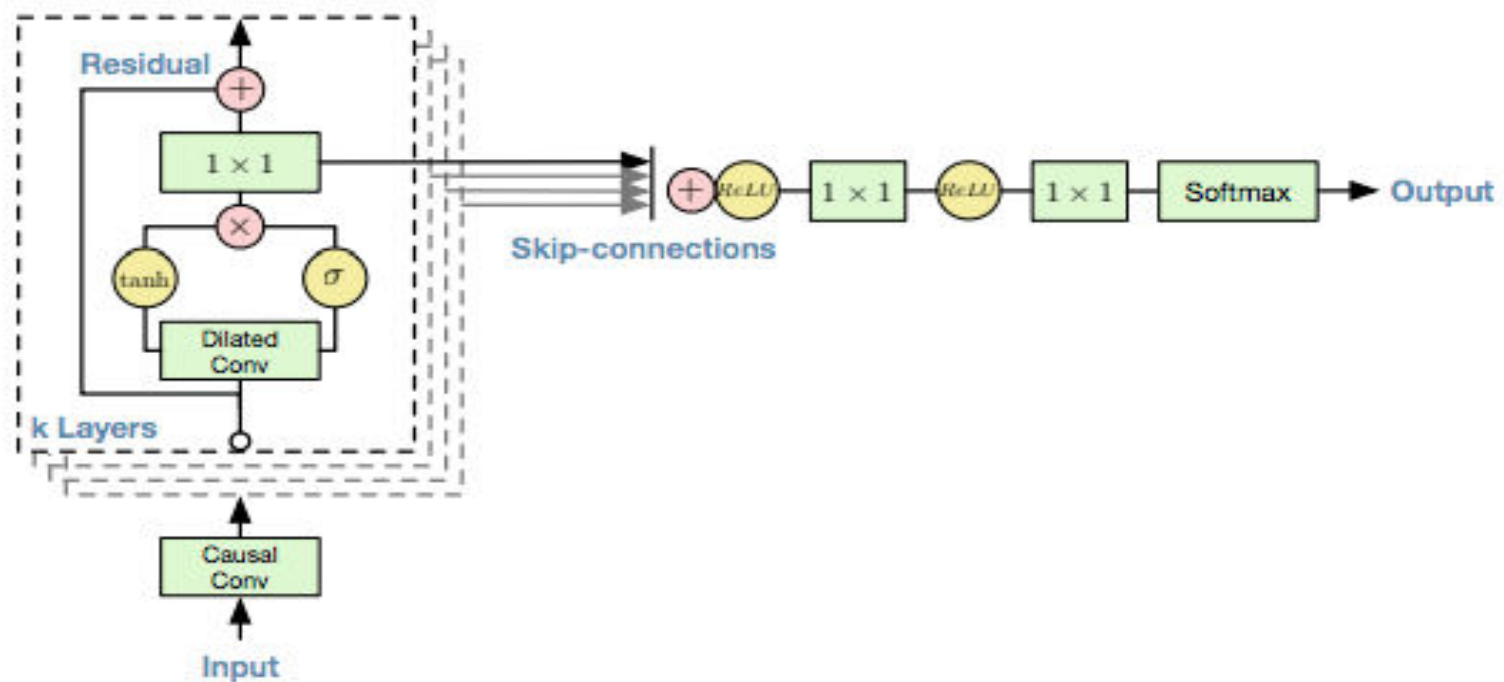
WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

- Dilated Causal Convolutions



WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

- Ağ yapısına genel bakış



WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

- Wavenet yapısının bir andaki ses dalgası değerinin sadece ondan önceki değerlere bağlı olduğu (causal) varsayımı üzerine kuruludur

$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1})$$

- Dilated \quad yüksek hesaplama maliyetinin etkisini azaltmaya yardımcı olur
- Dilated convolutions sayesinde bir andaki ses dalgasını hesaplarken bakılan girdi penceresinin büyüklüğü ağların derinliği ile üssel oranda artar



WAVENET: A GENERATIVE MODEL FOR RAW AUDIO

- Artılar
 - Doğallık açısından en gelişkin seviyede
 - Doğrudan ses dalgasını modelleyebiliyor
 - Convolutional yapısı sayesinde paralel olarak çalışabilir
- Eksiler
 - Eğitimi büyük veritabanları gerektiriyor
 - Hesaplama maliyeti yüksek
 - Endüstriyel kullanım için bu haliyle yeterince hızlı değil
 - Ses üretimi paralel değil ardışık
 - İleriye dönük zamansal bilgileri değerlendirmiyor



TACOTRON: A FULLY END-TO-END TEXT-TO-SPEECH SYNTHESIS MODEL

- Uçtan uca ses sentezi modeli
- Karmaşık bir ağ yapısı
 - Dikkat-tabanlı encoder-decoder
 - Convolutional
 - Highway Networks
 - Bi-directional RNN
 - Residual Bağlantılar
- Öznitelik mühendisliği ve alan bilgisi gerektirmiyor
- 1984'ten bir vocoder kullanıyor
- Wavenet'ten daha hızlı
- Yazım yanlışlarına karşı gürbüz ve noktalamaya karşı duyarlı



NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS (TACOTRON 2)

- 2 fonksiyondan oluşuyor
 - Karakter embedding'den mel-spectrogram'a
 - Mel-spectrogram'dan ses dalgasına (Wavenet'in değiştirilmiş bir hali)
- Vocoder'un yerine değiştirilmiş bir Wavenet yapısı
- Ağ yapısının içinde:
 - Dikkat-tabanlı katman
 - LSTM
 - Bi-LSTM
 - Convolution



DEEP VOICE: REAL-TIME NEURAL TEXT-TO-SPEECH

- Baidu'nun TTS modeli
- Öznitelik işleme, ön-yüz işlemleri DNN ile modelleniyor
- Çıktıları bir diğerinin girdisi olacak şekilde 5 farklı ağ
- Performans odaklı, gerçek zamanlı ses sentezleyebiliyor
- Ağlar genel olarak RNN ve CNN'lerden oluşuyor



DEEP VOICE 2: MULTI-SPEAKER NEURAL TEXT-TO-SPEECH

- Benzer bir uçtan uca sistem ancak öncekine göre daha hızlı sentezleme ve üretilen ses kalitesinde artış
- Önceki Deep Voice'un yanında Tacotron'dan daha iyi performansa sahip
- Tacotron'un kalitesinin bir post-işlem sinirsel vocoder geliştirildiği gösteriliyor
- Çok konuşmacılı bir TTS sistemi olarak yüksek performans
 - Binlerce benzersiz ses
 - Her bir konuşmacı için 30 dakikadan az veri



DEEP VOICE 3: SCALING TEXT-TO-SPEECH WITH CONVOLUTIONAL SEQUENCE LEARNING

- RNN bulundurmeyan convolutional and dikkat-tabanlı bir sistem
- 3 farklı vocoder ile uyumlu (World, Griffin-Lim, Wavenet)
- Büyük veritabanlarında eğitilmek üzere ölçeklenebilir bir model (800 saatten fazla ses)
- Konuşmacı embeddingleriyle çalışan çok konuşmacılı bir TTS modeli
- Muhtemel endüstriyel kullanımı tarif ediliyor
- Tek GPU ile günlük 10 milyon sorguyu işleyebilir düzeyde



CHAR2WAV: END-TO-END SPEECH SYNTHESIS

- Facebook'un uçtan uca modeli
- 2 bileşenden oluşuyor
 - Okuyucu
 - Sinirsel Vocoder
- Okuyucu dikkat mekanizması içeren bir encoder-decoder yapısı
 - Encoder: girdi olarak yazı ve ya fonem alan bir bi-RNN
 - Decoder: çıktı olarak akustik parametreler üreten dikkat tabanlı bir RNN
- Model sadece işlenmemiş yazı ve ses ile eğitilebilir



POSTFİLTRE-TABANLI İYİLEŞTİRME



POSTFILTER

Neden Postfilter kullanılıyor?

- Sentezlenen sesin kalitesi 3 ana nedenden dolayı doğal sesten farklı olabilir [1]:
 - Akustik modelde kullanılan özelliklerin doğal sesi temsil edecek kadar ayrıntılı olmaması
 - Doğal sesin kalitesinin akustik modelin genellemesinden dolayı kaybolması
 - Akustik modelin performansının yeterince iyi olmaması



POSTFILTER

Postfilter Türleri

1. Analitik Metodlar

- ✓ Sentezlenen ses kalitesinde artış
- Akustik modelde yeni özellik kullanmak

2. Veriden Öğrenilen Metodlar

- ✓ Sentezlenen ses kalitesinde artış
- ✓ Akustik modelde yeni özellik kullanmak



POSTFILTER

1. Analitik Metodlar

- I. Maximum Likelihood Parameter Generation [2]
 - Sentezlenmiş ses özelliklerini kullanılarak en yüksek olasılığa sahip doğal sesi oluşturmak amaçlanır.
- II. Global Variance Metodu [3][10]
 - Sentezlenmiş ses özelliklerinin varyansını doğal ses özelliklerinin varyansına yaklaştırarak akustik modelin genellemesinden dolayı oluşan kaybı azaltmaya çalışır.



POSTFILTER

1. Analitik Metodlar

III. Modulation Spectrum Metodu [4]

- Sentezlenmiş ve doğal ses parametrelerinin güç spektrumu arasındaki farkın azaltılması hedeflenir.

III. Spectral Peak Enhancement [5, 6]

- Ses spektrumunda bulunan tepe noktaları, sentezlenmiş sesin ve doğal sesin formantları arasındaki farkı azaltmak için kullanılır.



POSTFILTER

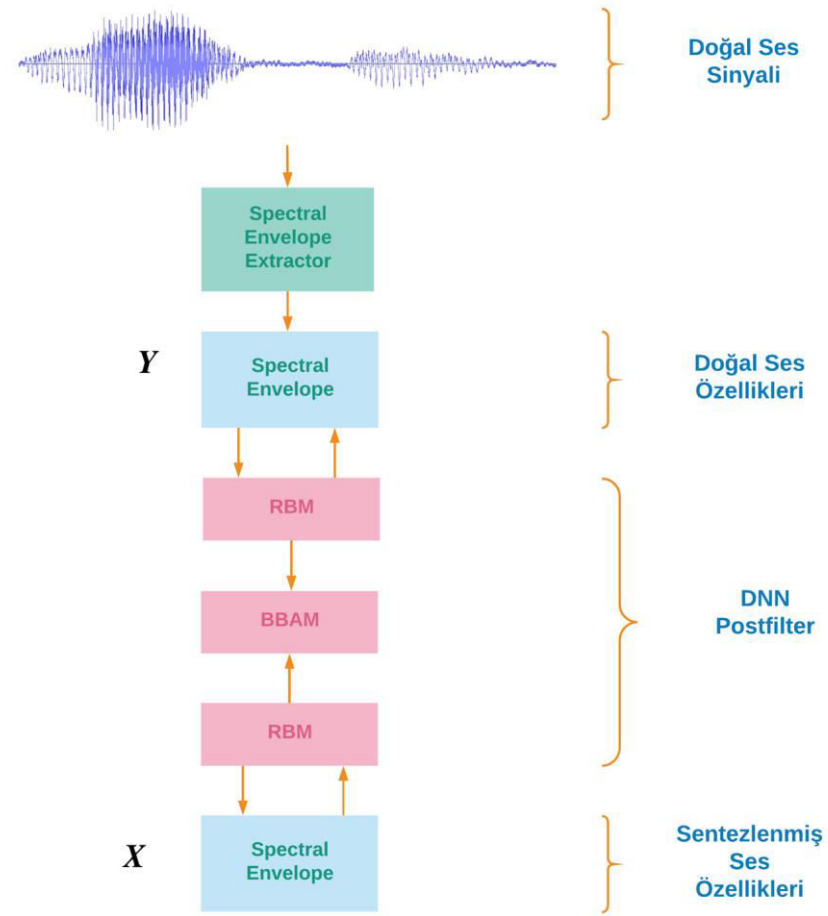
2. Veriden Öğrenilen Metodlar

- I. Deep Neural Network-based Postfilter [7]
- II. Recurrent Neural Network-based Postfilter [8]
- III. Generative Adversarial Network-based Postfilter [9]



I. Deep Neural Network-based Postfilter [7]

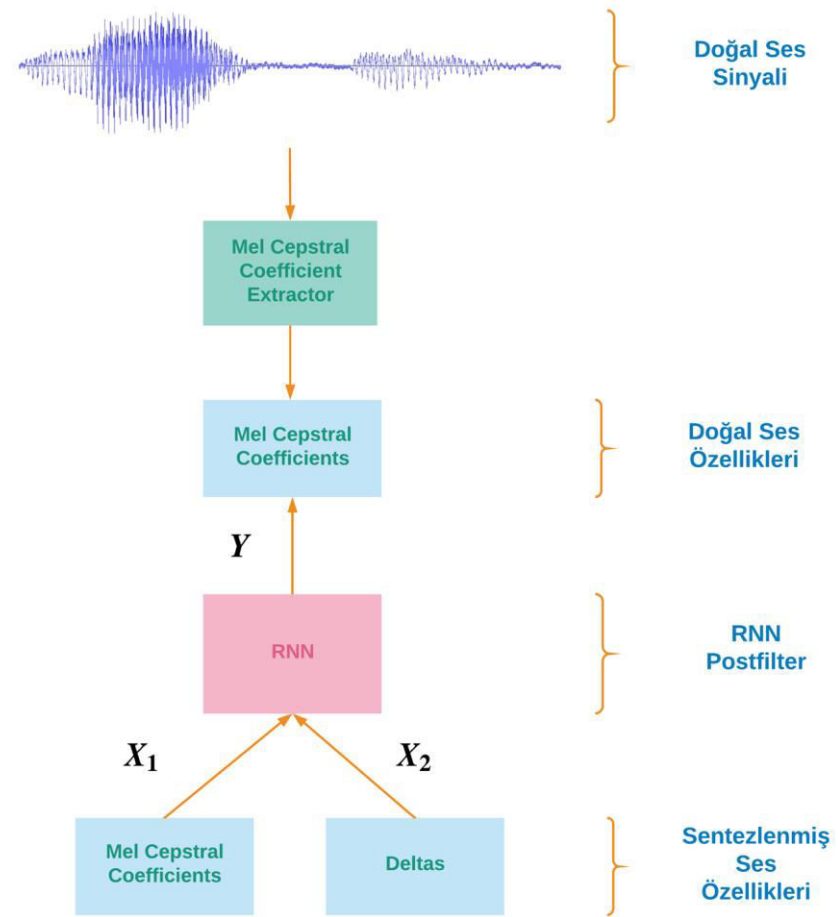
- RBM Gauss dağılımını varsayarak doğal ve sentezlenmiş sesi modellemiştir.



DNN-based Postfilter model yapısı



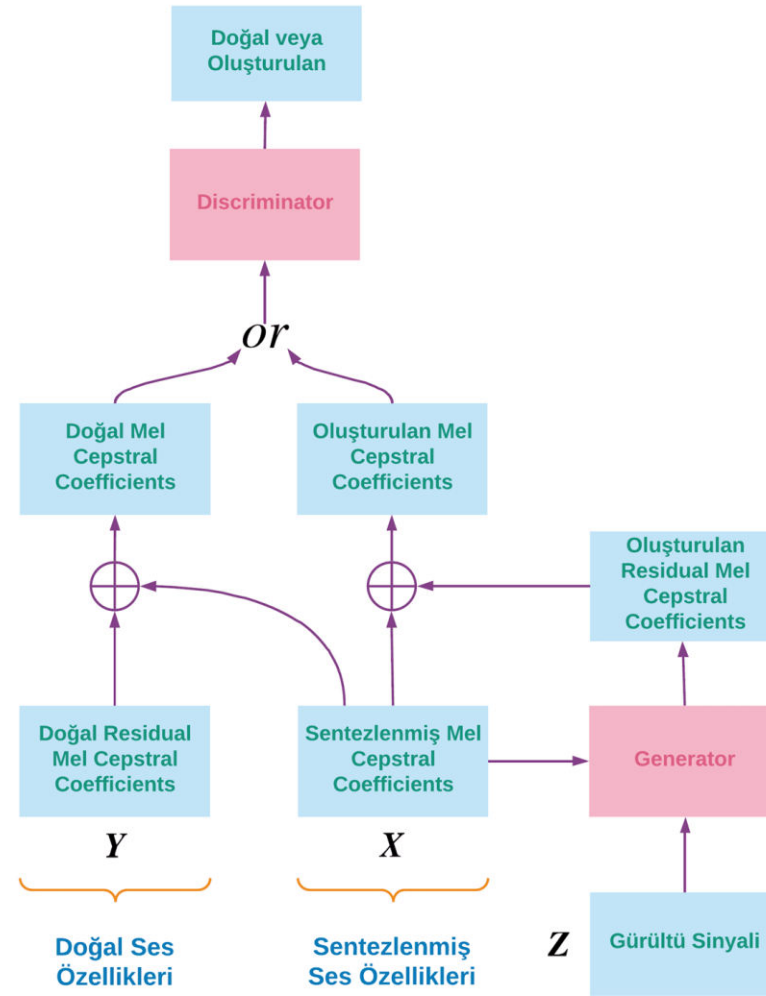
II. Recurrent Neural Network-based Postfilter [8]



RNN-based Postfilter model yapısı



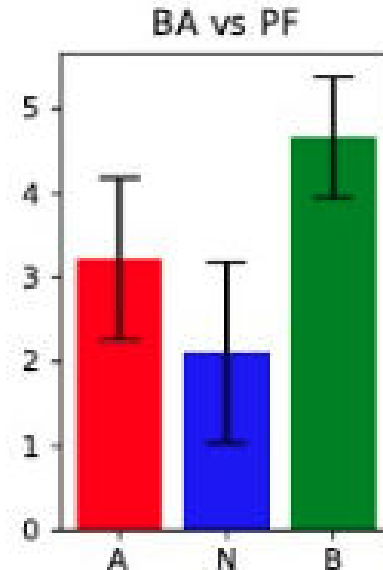
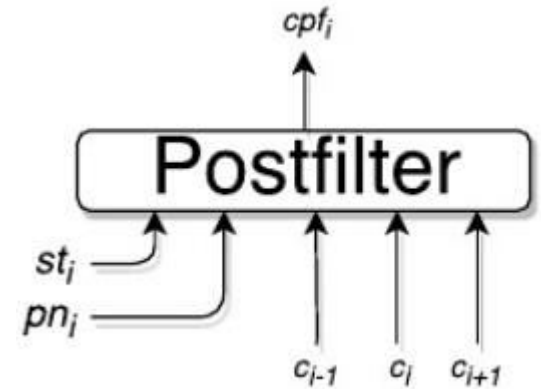
III. Generative Adversarial Network-based Postfilter [9]



GAN-based Postfilter model yapısı

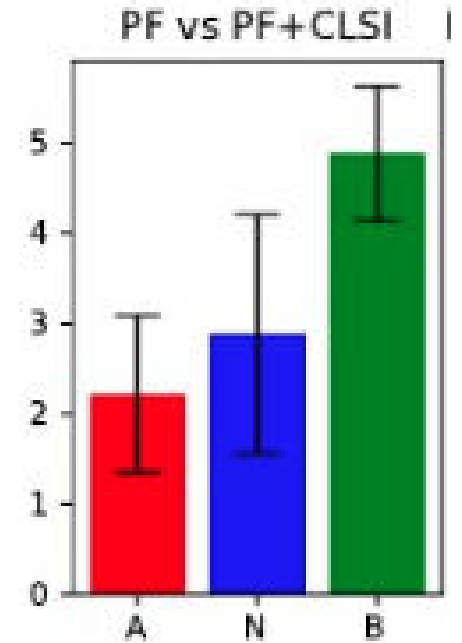
ÖNERİLEN POSTFILTER YÖNTEMLERİ

- Spektral öznitelikler üzerinde sığ sinirsel ağlar kullanımı
 - Post-filter, TTS modelinden bağımsız olarak eğitilen bir model
 - State ve fonem bilgisi de veriliyor
 - Doğallığı ve hedeflenen kullanıcıya olan yakınlığı arttırıyor
 - Ses kalitesinde hafif bir düşüş
 - Baseline ve Post-filter için ABX testi



ÖNERİLEN POSTFILTER YÖNTEMLERİ

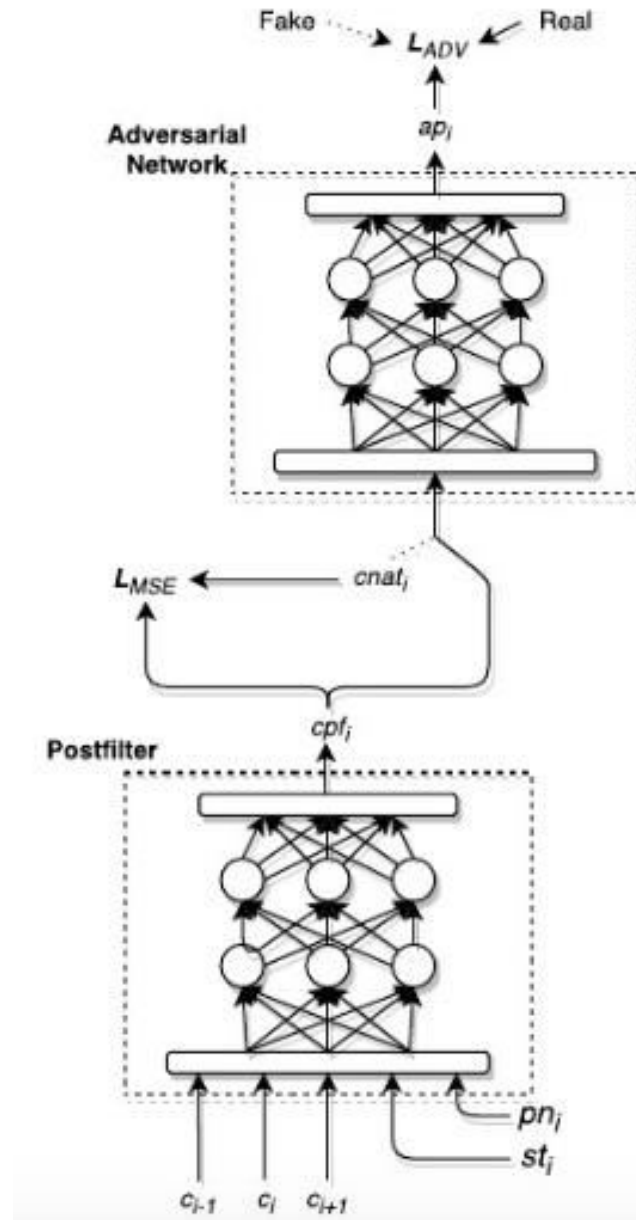
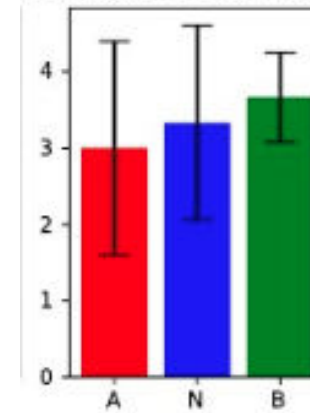
- Post-filter için küme-tabanlı ağırlık başlatma
 - Eldeki konuşmacıların vektörleri k gruba kümelenir
 - Her bir grup için ayrı bir post-filter eğitilir
 - Hedeflenen konuşmacı için en uygun grup bulunur
 - Hedeflenen konuşmacı için uyarlanacak post-filter bu grubun post-filter'ının ağırlıklarıyla başlatılır
 - Normal post-filter ve küme-tabanlı ağırlık başlatmalı post-filter için ABX testi(sağda)



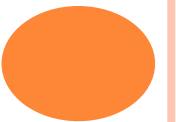
ÖNERİLEN POSTFILTER YÖNTEMLERİ

- Adversarial eğitim eklentisi
- Post-filter eğitiminde ek olarak, doğallığı arttırmak için GAN'larda olduğu gibi bir adversarial hata fonksiyonu ekleniyor
- Önerilen modelin ürettiği parametreler global varyans açısından kesin referansa daha yakın sonuçlar veriyor
- Küme-tabanlı ağırlık başlatmalı post-filter ve ona ek olarak GAN eklentisinin ABX karşılaştırması(yukarıda)

PF+CLSI vs PF+CLSI+GAN



GAN AĞLARI



TARİHİ

- GAN Goodfellow ve ark. tarafından 2014 yılında bulundu.
 - 3000 fazla atıf aldı
 - Birçok çalışmaya ilham oldu
- İlk olarak bilgisayarla görme alanında ortaya çıktı
 - Üretilen ilk örnekler düşük çözünürlüklü yüzler, hayvanlar ve objelerdi.
 - Sonradan konuşma ve doğal dil işleme alanlarında kullanılmaya başlandı
- Lecun tarafından “Son 10 yılın en sıradışı fikri” olarak tanımlandı
 - Dağılımları taklit edebilme özelliği
 - Beyin gibi gözetimsiz olarak çalışması



GİRİŞ

- Kapsamlı ek açıklamalar olmaksızın belirli bir dağıtımdan örnekler üretir
- Belirli bir dağılımı modelleyen denetimsiz öğrenim ile çalışan üretken bir model
- Sıfır-toplam oyunda birbiriyle yarışan bir çift sinir ağı
 - Ayırt edilemez örnekler üretmeye çalışan bir üretken ağı
 - Gerçek ve üretilen örnekler arasında ayırım yapmaya çalışan bir ayırmacı ağı
- Konveks durumdaki Nash dengesine yakınlaşır
 - Üretilen örneklerin verilen örnekler kadar "gerçek" olduğu ideal bir nokta
 - Gerçek dünya problemlerinin çoğu konveks değildir
 - Eğitim kırılıgandır ve ideal yakınsama zordur



GİRİŞ(DEVAMI)

- GAN'lar ayırmacı olmaktan ziyade üretken modellerdir
 - X özellikler y ise etiketler olsun
 - Ayırmacı bir model koşullu olasılığı öğrenmek için çalışır. $p(y | x)$
 - Örnek, “görüntü pikselleri” (özellikler) verilen “bir at resmi olma olasılığı” (etiket) olasılığı
 - Bir generatif model bileşik olasılığı öğrenmeye çalışır. $p(x, y)$
 - Bu fotoğrafın elma(y) olduğunu varsayarsak, bu parametreler(x) ne kadar olasıdır?

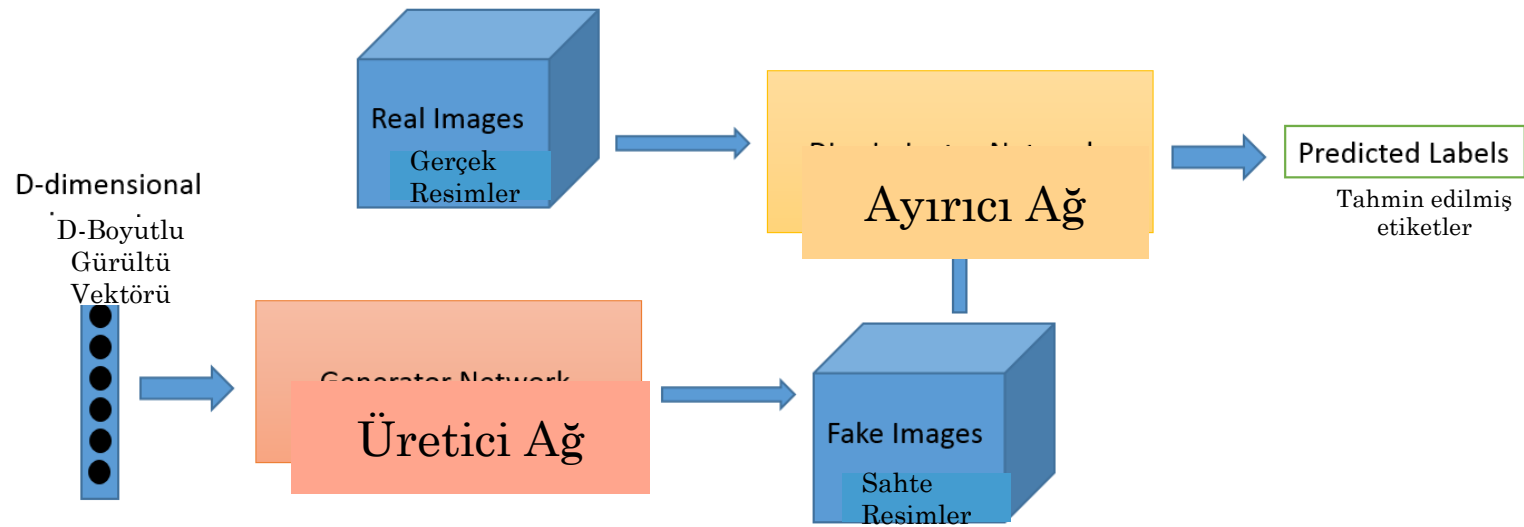


BASITÇE GAN'LARIN EĞİTİMİ

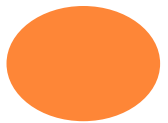
- Üretken Ağ (G) tamamen veya kısmen rastgele giriş alır
- G bir örnek oluşturur (örn., resim)
- Üretilen örnek ve gerçek örneklerle birlikte ayırıcı (D) beslenir.
- D, hem gerçek hem de oluşturulan örnekler için orijinalliği tahmin eden olasılıkları döndürür
- Ayırıcı, etiketler vasıtasıyla bir geri bildirim döngüsünde çalışır
- Üretici ise ayırıcı ile bir geri bildirim döngüsünde çalışır



EĞİTİMİN BASİT BİR DİYAGRAMI

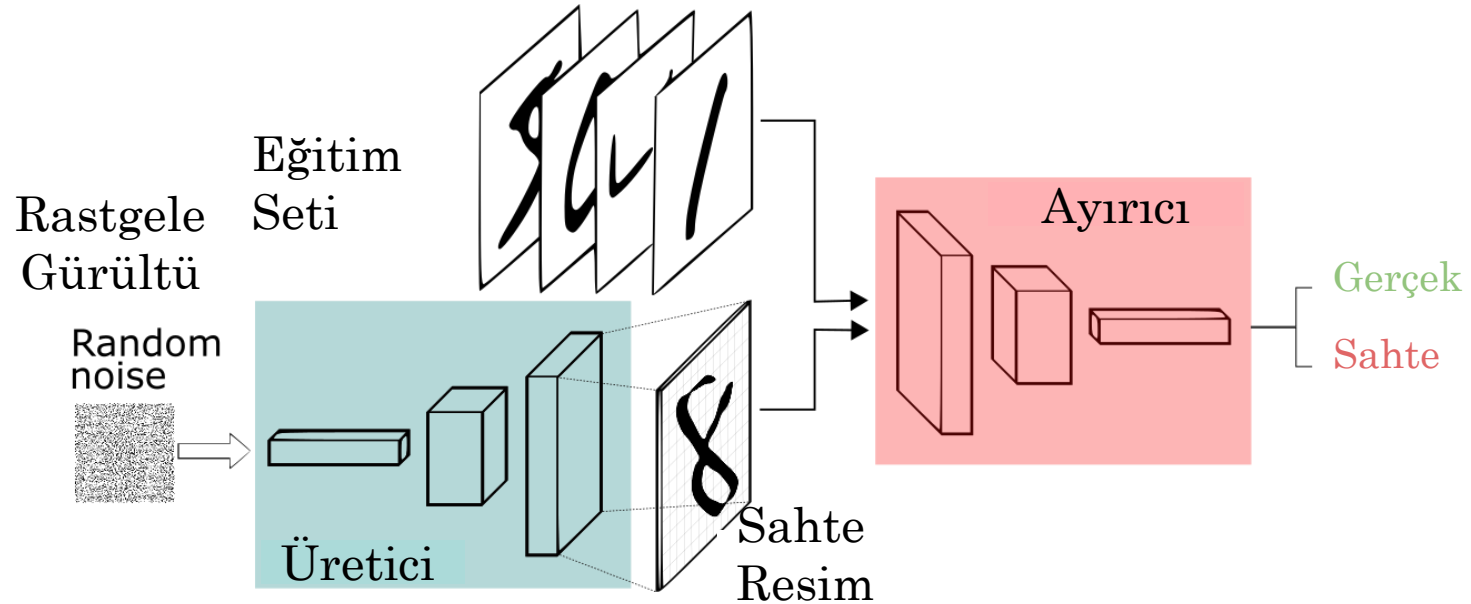


(courtesy of O'Reilly media)



KALPAZAN - POLİS ANALOJİSİ

- Basit bir analogi ile,
 - Üretici(kalpazan) sahte para basar
 - Ayırıcı(polis) sahte parayı saptamaya çalışır
 - İkisi de daha iyi olabilmek için kendilerini geliştirir



EĞİTİME TEKNİK YAKLAŞIM

- Üretici (G) ve Ayırıcı (D) türevlenebilir fonksiyonlardır
- G ve D parametreleri gradient descent yöntemi ile öğrenilir
- G bir gürültü girdisi olan z vektörünü alır, x 'e benzer bir örnek olan $G(z)$ 'yi döndürür
- D, x ve $G(z)$ değerini girdi olarak alır ve $D(x)$ ve $D(G(z))$ için tahmin yapar
- D and G için parametreler $\theta(D)$ ve $\theta(G)$ olsun
- Ayırıcı $J(D)(\theta(D), \theta(G))$ fonksiyonunu minimize etmeye çalışır
 - Ama $\theta(G)$ üzerinde bir kontrolü yoktur
- Üretici $J(G)(\theta(G), \theta(D))$ fonksiyonunu minimize etmeye çalışır
 - Ama $\theta(D)$ üzerinde bir kontrolü yoktur



EĞİTİME TEKNİK YAKLAŞIM

- İdeal eğitim süreci, Nash dengesinde sonuçlanır
 - $J(D)(\theta(D), \theta(G))$ fonksiyonu $\theta(D)$ 'ye göre minimum
 - $J(G)(\theta(D), \theta(G))$ fonksiyonu $\theta(G)$ 'e göre minimum

$$J^{(D)} = -\frac{1}{2}\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \log D(\mathbf{x}) - \frac{1}{2}\mathbb{E}_{\mathbf{z}} \log (1 - D(G(\mathbf{z})))$$

$$J^{(G)} = -J^{(D)}$$

- $J(D)$ 'nin ilk teriminde D , verilen örnek (x) gerçek olduğunda 1'in log olasılığını maksimize etmeye çalışır
- $J(D)$ 'nin ikinci teriminde D , verilen örnek ($G(z)$) üretilmiş olduğunda 0'ın log olasılığını maksimize etmeye çalışır
- G , verilen örnek $G(z)$ olduğunda D 'nin tahmininden 1'in log olasılığını en üst düzeye çıkarmaya çalışır



ÜRETİLEN BAZI ÖRNEKLER (2014)

Toronto Face Database kullanarak
oluşturulmuş yüzler

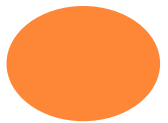
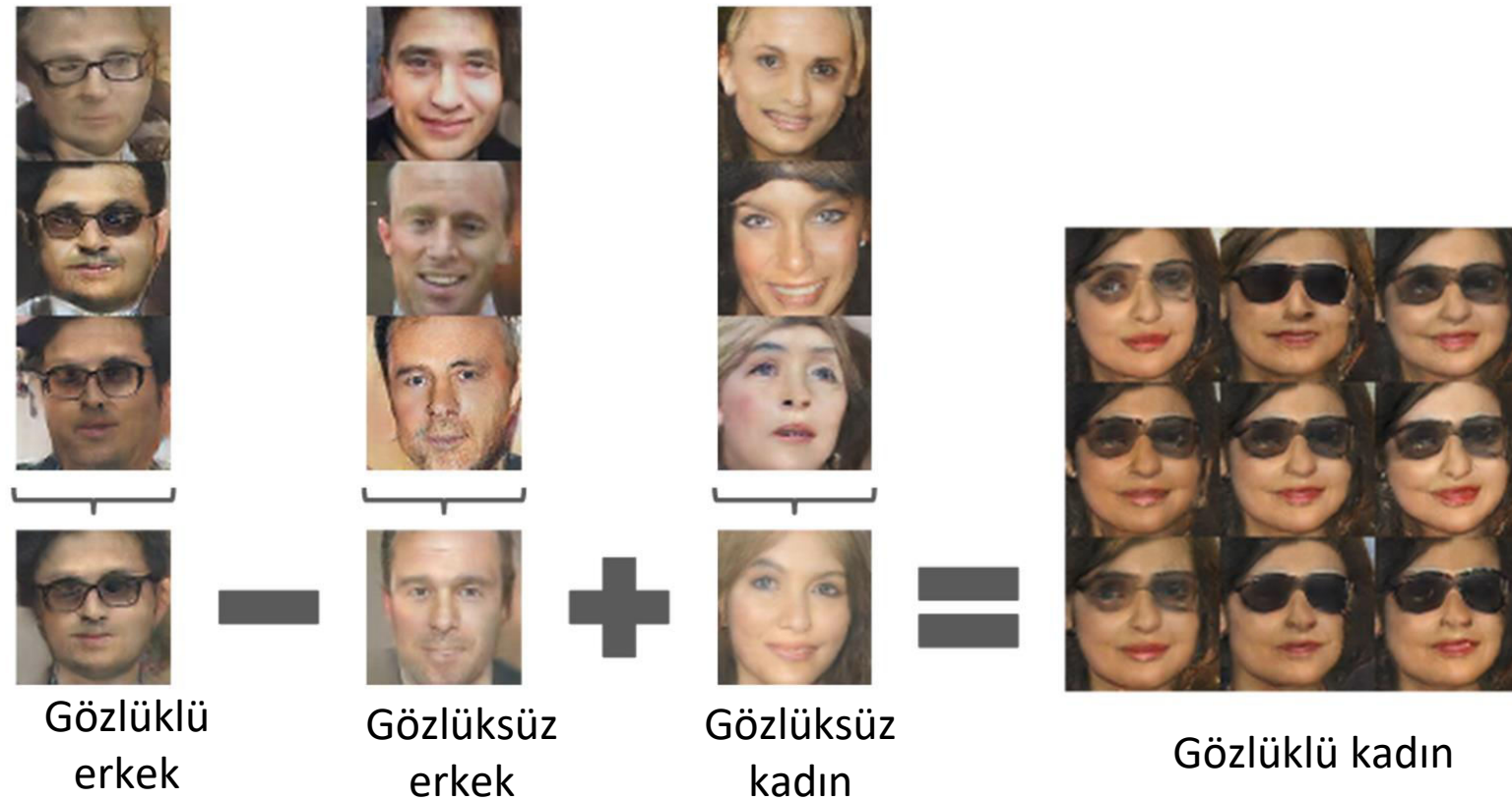


ÜRETİLEN BAZI ÖRNEKLER (2018)

CelebA-HQ veri kümesini kullanarak oluşturulmuş yüzler



BAZI OLASI ARİTMETİK İŞLEMLER



UYGULAMA ALANI

- Görüntü sentezi
- Anlamsal görüntü düzenleme
- Style transfer
- Image super-resolution
- Yazıdan resim üretimi
- Bir sonraki video karesi tahmini



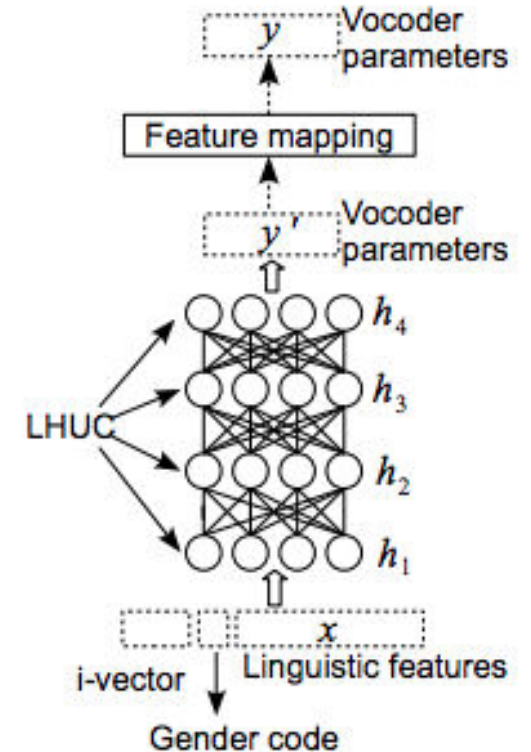
DNN TABANLI TTS SİSTEMLERİNDE KONUŞMACIYA UYARLAMA

- A study of speaker adaptation for DNN-based speech synthesis (Wu et al, 2015)
- An Investigation of DNN-Based Speech Synthesis Using Speaker Codes (Hojo et al, 2016)
- Adapting and controlling DNN-based speech synthesis using input codes (Luong et al, 2017)
- Neural Voice Cloning with a Few Samples (Arik et al, 2018)
- Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis (Jia et al, 2018)



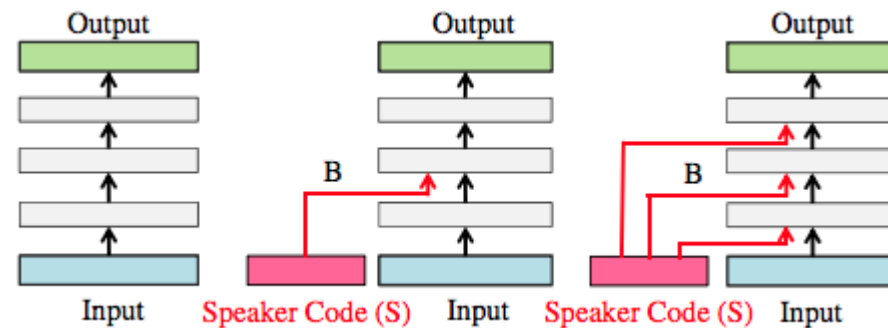
A STUDY OF SPEAKER ADAPTATION FOR DNN-BASED SPEECH SYNTHESIS

- 3 farklı seviyede uyarlama
 - Girdi katmanında konuşmacıya özel vektör
 - Konuşmacının ses özelliklerini temsil eden düşük boyutlu bir vektör
 - Ağ içerisinde Learning Hidden Unit Contributions(LHUC)
 - Saklı ünitelerin çıktılarını kişiye göre ölçekleyen bir sistem
 - Çıktı seviyesinde bir lineer dönüşüm
 - Joint Density Gaussian Mixture Modeli (JD-GMM)



AN INVESTIGATION OF DNN-BASED SPEECH SYNTHESIS USING SPEAKER CODES

- Konuşmacı vektörü olarak one-hot vektörler
- One-hot daha önce görülmemiş konuşmacılar için işe yaramıyor (bu haliyle sadece çok konuşmacılı sistemler için)
- Konuşmacı vektörü birden fazla katmana girdi olarak verilebiliyor
- Uyarlama için konuşmacı vektörleri donduruluyor sadece B ağırlıkları güncelleniyor



ADAPTING AND CONTROLLING DNN-BASED SPEECH SYNTHESIS USING INPUT CODES

- Uyarlamada konuşmacıya özel vektörlere ek olarak cinsiyet ve yaş bilgisi de veriliyor
- Konuşmacı, cinsiyet ve yaş vektörleriyle konuşmacıdan bağımsız bir model oluşturuluyor
- Yeni konuşmacılar için bu vektörler backpropagation ile elde ediliyor
- Uyarlanmış sentezler bu elde edilen vektörler ile yapılıyor



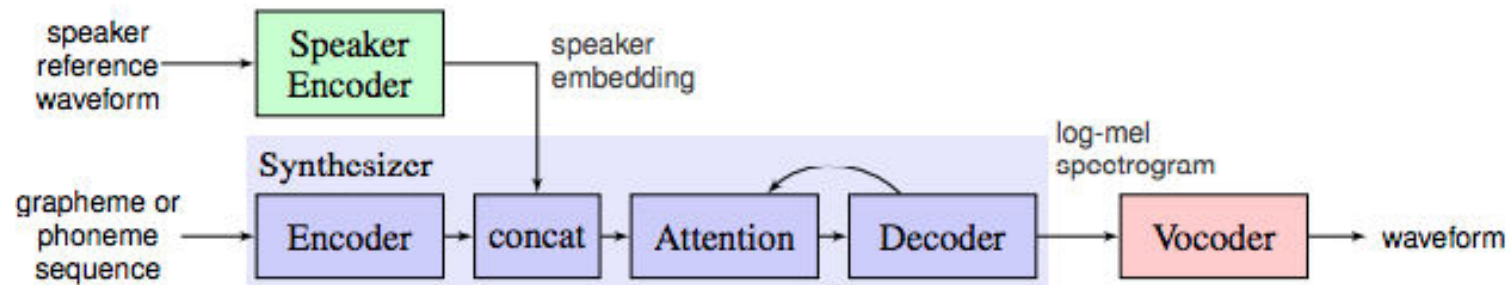
NEURAL VOICE CLONING WITH A FEW SAMPLES

- Konuşmacı embeddingleri kullanılarak konuşmacıdan bağımsız bir model oluşturuluyor
- Uyarlama sırasında model ağırlıkları yerine bu embeddingler öğreniliyor
- Ayrıca konuşmacı encode eden bir DNN eğitip bu DNN'in çıktısını konuşmacı vektörü olarak kullanarak uyarlama sunuluyor
- Sistem çok az sayıda örnek ile çalışabiliyor



TRANSFER LEARNING FROM SPEAKER VERIFICATION TO MULTISPEAKER TEXT-TO-SPEECH SYNTHESIS

- Sistem 3 Sinir ağından oluşuyor
 - Konuşmacı encode eden ağ, önceden konuşmacı doğrulama işi için eğitilmiş(yeşil)
 - Tacotron 2'den uyarlanan bir diziden-diziye ağ modeli(mavi)
 - Oto-regresif Wavenet-tabanlı bir vocoder ağı(kırmızı)
- 3 ağ da birbirinden bağımsız bir şekilde eğitiliyor
- Konuşmacı encode eden ağ tüm sistemin darboğazı olabilir ve eğitimi büyük bir veritabanı gerektiriyor



DNN TABANLI ÖNERİLEN UYARLAMA YÖNTEMLERİ

- Konuşmacıdan bağımsız model üzerinde transfer öğrenmesi
 - Hedef konuşmacının vektörünün kullanılması(baseline)
 - Hedef konuşmacıya en yakın konuşmacının vektörünün kullanılması
 - Küme bilgisinin konuşmacı vektörü olarak kullanılması
 - DNN modeline ait bazı ağırlıkların tekrar eğitilmesi



DNN TABANLI ÖNERİLEN UYARLAMA YÖNTEMLERİ

- Hedef konuşmacıya en yakın konuşmacının vektörünün kullanılması
 - a. Konuşmacıdan bağımsız modelin eğitimi sırasında kullanılan konuşmacılar arasında hedef kullanıcıya en yakın olan bulunur
 - b. Uyarlama bu en yakın komşunun vektörü ile yapılır
- Küme bilgisinin konuşmacı vektörü olarak kullanılması
 - a. Eldeki tüm konuşmacılar k tane gruba kümelenir
 - b. Konuşmacıdan bağımsız bir model sadece küme bilgisiyle eğitilir
 - c. Hedef konuşmacı için en uygun grup seçilir
 - d. Uyarlama işlemi bu grup bilgisiyle yapılır



DNN TABANLI ÖNERİLEN UYARLAMA YÖNTEMLERİ

- DNN modeline ait bazı ağırlıkların tekrar eğitilmesi
 - a. Önceden belirlenen bazı ağırlıkları hedef konuşmacının ses verisine göre güncelle
 - b. Bütün ağırlıklar değil sadece bir alt kümesi güncelleniyor(overfit'ten kaçınmak için)
 - c. Genel olarak çıktı üzerindeki etkisi daha yüksek(gradyanı yüksek) ağırlıklar seçilir(üst katmanlara ait ağırlıklar)



DNN TABANLI ÖNERİLEN UYARLAMA YÖNTEMLERİ

- Ağırlıklar tekrar eğitilmeden elde edilen hata değerleri
 - DNN modelin ağırlıkları değişmeden
 - Sadece konuşmacı vektörlerin değişimi var

SI-NN	SI-tar	Cluster
5.46	5.53	5.22

- Bazı ağırlıklar güncellendiğinde

	5utt	15utt	30utt
NN i-vector	5.06	4.93	4.87
Target i-vector	5.26	5.17	5.09
Cluster ID	5.24	5.16	5.12



FUTURE

- GAN'ların eğitilmesi hala zor, yakınsama sorunu henüz çözülmedi
- Oluşturulan örnekler çoğunlukla insan algısında görülebilen kusurlara sahip
- GAN'lar sahte örnekler üretirken daha iyi hale geldikçe, doğal olarak güvenlik sorunu ortaya çıkıyor
- GAN'lar, çıktıları ayırt edilebilir olduğunda daha iyi çalışırlar, bu da doğal dil işleme gibi alanlarda kullanılmasını zorlaştırır



REFERANSLAR

1. Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11), 1039-1064.
2. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on* (Vol. 3, pp. 1315-1318). IEEE.
3. Toda, T., & Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE TRANSACTIONS on Information and Systems*, 90(5), 816-824.
4. Takamichi, S., Toda, T., Neubig, G., Sakti, S., & Nakamura, S. (2014, May). A postfilter to modify the modulation spectrum in HMM-based speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 290-294). IEEE.
5. Ling, Z. H., Wu, Y. J., Wang, Y. P., Qin, L., & Wang, R. H. (2006, September). USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method. In *Blizzard Challenge Workshop*.
6. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (2005). Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis. *Systems and Computers in Japan*, 36(12), 43-50.



REFERANSLAR

7. Chen, L. H., Raitio, T., Valentini-Botinhao, C., Yamagishi, J., & Ling, Z. H. (2014, May). DNN-based stochastic postfilter for HMM-based speech synthesis. In INTERSPEECH (pp. 1954-1958).
8. Muthukumar, P. K., & Black, A. W. (2016). Recurrent neural network postfilters for statistical parametric speech synthesis. arXiv preprint arXiv:1601.07215.
9. Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K., & Kashino, K. (2017, March). Generative adversarial network-based postfilter for statistical parametric speech synthesis. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on (pp. 4910-4914). IEEE.
10. Toda, T., Black, A. W., & Tokuda, K. (2005, March). Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on (Vol. 1, pp. I-9). IEEE.



REFERANSLAR

11. Generative Adversarial Networks(Goodfellow et al, 2014)
12. Learning Factorial Codes by Predictability Minimization(Schmidhuber, 1992)
13. <https://deeplearning4j.org/generative-adversarial-network>
14. <https://towardsdatascience.com/generative-adversarial-networks-history-and-overview-7effbb713545>
15. Generative Adversarial Networks: An Overview(Creswell et al, 2017)
16. <https://www.quora.com/What-research-directions-is-Ian-Goodfellow-pursuing-to-improve-Generative-Adversarial-Networks>
17. <https://www.oreilly.com/learning/generative-adversarial-networks-for-beginners>
18. https://en.wikipedia.org/wiki/Generative_adversarial_network
19. Progressive Growing of GANs for Improved Quality, Stability, and Variation(Karras et al, 2018)
20. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks(Radford et al, 2015)



REFERANSLAR

21. Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
22. <https://deepmind.com/blog/wavenet-generative-model-raw-audio/>
23. Black, A. W., Zen, H., & Tokuda, K. (2007, April). Statistical parametric speech synthesis. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on (Vol. 4, pp. IV-1229). IEEE.
24. Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. Speech Communication, 51(11), 1039-1064.

