Türkçe Metinden Konuşma Sentezlemede Doğallığın Artırılması İçin Öneriler

Recommendations for Increasing the Naturalness in Turkish Text-to-Speech Synthesis

İ. Baran Uslu[1], H. Gökhan İlk[2], A. Egemen Yılmaz[2]

[1] Elektrik-Elektronik Mühendisliği Bölümü
Başkent Üniversitesi
ibuslu@baskent.edu.tr

[2] Elektronik Mühendisliği Bölümü
Ankara Üniversitesi
h.gokhan.ilk@eng.ankara.edu.tr, aeyilmaz@eng.ankara.edu.tr

Özet

Metinden konuşma sentezleme; yazılı bir metnin geliştirilen sistem tarafından otomatik olarak okunmasıdır. Bu çalışmada, difon tabanlı, eklemeli bir konuşma sentezleyici tasarlanmış ve gerçekleştirilmiştir. Birleştirmede PSOLA yöntemi kullanılmaktadır. Genellikle konuşma sentezleyicilerin ezgi modeli yoktur veya eksiktir. Bu durum sentezlenen konuşmanın doğallığını olumsuz yönde etkiler. Çalışmamızda bu eksikliğin giderilmesi için yeni bir model önerilmiştir. Sentezlenen konuşmanın doğallığının artırılması için, konuşmanın ezgisi üzerinde süre ve vurgu temelli kurallar tanımlanmıştır. Bu kurallar, hazırlanan ara yüzde yapılan pek çok denemenin sonucunda bulunmuştur. Uygulanan kuralların sentezlerin doğallığındaki başarısı öznel dinleme testleriyle ölçülmüştür. Sonuç olarak, tanımlanan kuralların geliştirilen konuşma sentezleyicide uygulanması ile CMOS testi sonucunda 1,86/5,00 puanlık bir artış elde edilmiştir. Bu sonuç, ezgi modelimizin başarılı olduğunu göstermektedir.

Anahtar kelimeler: Metinden konuşma sentezleme, difon, PSOLA, ezgi modeli, doğallık, CMOS

Abstract

Text to speech synthesis (TTS) is the automatic reading of a text by a system. In this work, a TTS system which concatenates diphones has been designed and implemented. For concatenations, PSOLA method was used. Usually speech synthesizers lack an intonation model. This degrades the naturalness of the synthesized speech. For increasing the naturalness of the synthesized speech, duration and accent based rules were defined in this study for a proper intonation. These rules were determined after an extensive set of experiments performed in the designed testbed. In the end, an improvement of 1.86/5.00 in the CMOS score was obtained by applying the defined rules in the developed synthesis platform. This result shows the success of our intonation model.

Keywords: Text to speech synthesis (TTS), diphone, PSOLA, intonation model, naturalness, CMOS

1. Giriş

Metinden konuşma sentezleme (MKS); pek çok dil için ortak ve önemli bir araştırma konusudur. Görme ve konuşma engelli insanlar için iletişim imkânı sağlaması [1], sesli yanıt, uyarı ve okuma sistemleri [2-4], dil ediniminde ve yabancı dil öğretiminde kolaylıklar sağlaması [5], MKS'nin uygulama alanlarından bazılarıdır. Bu alandaki temel kaynaklar [6-9]'da verilmiştir.

Bu konuda yapılan araştırmaların hedefi; insan sesi doğallığında konuşma sentezidir. Sentezlenen konuşmanın kalitesi ve doğallığı arttıkça, MKS sistemleri gündelik hayata daha çok girecektir. 1993'ten bugüne kadar Türkçe MKS sistemleri üzerinde pek çok lisansüstü çalışma yapılmıştır. Bu çalışmalarda sinyal işleme yönteminin ve kullanılan ses parçalarının senteze olan katkısının yanı sıra, frekans değişiminin ve sürenin modellenmesi de incelenmiştir [10]. Bu makalede Türkçe metinden konuşma sentezlemede ezgi modelleri üzerinde durulmuş, önerdiğimiz yeni ezgi modeli anlatılmıştır.

Doğal bir konuşmanın sentezlenmesinin önündeki engeller arasında büyük bir konuşma parçası veritabanının (farklı uzunluk ve temel frekanslı) oluşturulması, konuşma parçalarının sürelerinin modellenmesi ve uygun ezgi kurallarının tanımlanması sayılabilir.

Türkçede vurgu ve ezgi yapıları, bazı sinyal işleme ve bilgisayar bilimleri araştırmacıları tarafından daha önce incelenmiştir [11-13]. Şayli [11], Türkçe MKS sistemlerinde süre modelleri üzerinde çalışmış, fonem ve trifon tabanlı incelemelerin sonucu olarak ortalama süreleri rapor etmiştir. Şayli'nin çalışmasındaki önemli sonuçlardan birisi de; cümle içinde kullanıldıklarında fonem ve trifon ortalama sürelerinin belirli oranlarda düşmesidir. Bunun sebebi, daha uzun bir konuşmanın tek nefeste söylenebilmesi için, tüm birimlerin belirli oranlarda sıkıştırılmasıdır. Öztürk [12], fonemler için süre ve F_0: temel frekans eğrilerinin modellenmesini ele almıştır. İstatistiksel olarak metinsel özellikler (fonem türü, hece sayısı, hecenin konumu, hecenin vurgu alıp almaması vb.) incelenmiş ve regresyon analizi yapılmıştır. Sonuç olarak, ortalama süre için en etkili parametreler: fonemin türü, ön ve arkadaki fonemlerin türleri ve fonemin hece içindeki yeri olarak rapor edilmiştir. Temel frekans eğrileri ise hece frekansları baz alınarak incelenmiştir. Öztürk, çalışmasının sonunda bu modellerin duyumsal olarak değerlendirilmesini önermektedir. F_0 üzerinde yapılan bir diğer çalışmada, Oskay vd. [13], cümle bazında temel frekans eğrilerinin genelleştirilmesi üzerinde durmuşlardır. F_0 eğrileri, olumlu, olumsuz ve soru cümleleri için doğrusal ve ikinci derece fonksiyonlar ile modellenmeye çalışılmıştır. Külekçi ve Oflazer [14], metin içerisindeki söz gruplarını belirlemeye çalışmışlar, bunlara 3 kademeli (0: yok, 1: az ve 2: fazla) ezgi seviyesi atamışlardır. %85 başarıyla söz gruplarını ayırmayı ve doğru vurgu seviyesini belirlemeyi başarmışlarsa da nesnel değerlendirmenin bir Türkçe MKS sistemiyle birleştirilmesi sonucunda elde edilebileceğini belirtmişlerdir. Uslu ve İlk [15]'de, Fujisaki ezgi modelini, birkaç Türkçe cümleye ilk defa uygulamışlardır. Bu yöntemde cümlenin perde frekansı değişimi bir toplamsal modelle ele alınmakta, tamlama (phrase) ve vurgu (accent) olmak üzere iki bileşene ayrılmaktadır. Modelin matematiksel ifadesi Eşitlik (1)'de verilmiştir.

$$\ln(F_0) = \ln(F_{\min}) + P + A$$
$$P = \sum_{k=1}^{N_p} A_{p,k} . g_p(t - T_{p,k})$$
$$A = \sum_{k=1}^{N_a} A_{a,k} \left[g_a(t - T_{a1,k}) - g_a(t - T_{a2,k}) \right] \quad (1)$$

Burada $g_p(t) = \alpha^2 t . e^{-(\alpha t)} u(t)$ tamlama dürtü tepkisi,
$g_a(t) = \min\left(1 - (1 + \beta t) . e^{(-\beta t)}, \gamma\right)$ aksan basamak tepkisidir.

$A_{p,k}$, $A_{a,k}$, $T_{p,k}$, $T_{a1,k}$, $T_{a2,k}$, α, β, γ; model parametreleri ve F_{min} ; taban frekansıdır.

Model, beklendiği gibi konuşmanın doğallığını arttırmış, PESQ testi sonucunda 0,15/4,00 puanlık bir iyileşme elde edilmiştir. PESQ: Perceptual Evaluation of Speech Quality, telefon hatlarının kalitesini ölçmek için önerilmiş bir yöntem olup öznel dinleme testlerindeki dinleyici bulma ve dinletme zahmetlerinden kurtulmak için tercih edilen bir nesnel değerlendirme testidir (ITU-T P.862). İki konuşma parçasının birbirine olan yakınlığını, bunları hizalayıp aralarındaki özilintiden bulmaya çalışır.

Dilbilimciler bir sözcükteki vurgunun yerini tespit etmek için seslemleri (heceleri) sırayla baskın bir şekilde okurlar. Hangi okuyuş kulağı tırmalamıyorsa vurgunun o şekilde doğru olduğuna karar verirler. Türkçede fiil çekimleri, istisnalar haricinde, kurallara bağlıdır [16]. Aydemir ve Yılmaz [16], çalışmalarında fiillerin otomatik çekimlenmesi ve vurgularının belirlenmesi üzerinde durmuşlar, Türkçedeki yaklaşık 4600 adet fiilden 1100 adedi için 5400 farklı çekimin yapıldığını ve vurgu pozisyonlarının doğru bulunduğunu belirtmişlerdir. Bu sonuçlar bizim de çalışmamızın temelini oluşturmaktadır. Uslu vd. [17], tasarladıkları MKS ara yüzünde fiil çekimleri için akustik özellikler (süre, perde frekansı ve enerji) ile ilgili ezgi kuralları önermişlerdir. Pek çok fiil çekimi için yapılan denemeler, web üzerinden dinleme testleri ile değerlendirilmiş ve en çok beğenilen yöntem sonuç olarak önerilmiştir.

Bu çalışmaların yanı sıra, duygusal sentez konusunda yapılan çalışmalar bulunmaktadır [18, 19]. Bu alanda, Bulut vd. [18] yaptıkları çalışmada; sesbirim düzeyinde süre, perde frekansı, enerji ve izge değişikliklerinin duygusal senteze olan etkisini incelemişlerdir. Sonuç olarak sesbirim düzeyinde duygu dönüştürmede, izgesel zarf değişikliklerinin yerel prozodi değişikliklerine göre daha etkili, yerel prozodi değişikliklerinde ise; süre değişiminin perde frekansı değişiminden daha başarılı olduğunu belirtmişlerdir. Burkhardt vd. [19], farklı dillerde (Türkçe, Yunanca, Almanca ve Fransızca) duygusal sentezler yapıp birbirleriyle karşılaştırmışlardır. Temel frekans, süre ve "jitter" parametreleriyle senteze duygu katmaya çalışmışlar; sonuçta, hem o dile özgü, hem de tüm dillerde ortak noktalar bulunduğunu belirlemişlerdir.

Bu çalışmada geliştirilen ezgi modelinde; seçilen cümlelerde sözcüklere, difon sayısına göre süre değiştirme işlemi uygulanmaktadır. Daha sonra, vurgulu hece dikkate alınarak, cümlenin fiiline ezgi verilmeye çalışılmakta ve ayrıca cümle içinde yer alan öbek vurguları için ezgi kuralları araştırılmaktadır. Tüm bu işlemler temel frekans, süre ve enerji parametrelerinin sistematik bir şekilde değiştirilmesi temeline dayanmaktadır. Farklı cümle yapıları ve öbek vurguları için adı geçen akustik parametrelerin optimum değerlerine duyumsal değerlendirme testleri sonucunda ulaşılmaya çalışılmıştır.

Bu Giriş bölümünün ardından, çalışmanın 2. Bölümünde izlenen yöntem ve önerilen ezgi modeli ayrıntılı olarak anlatılmıştır. 3. Bölümde elde edilen bulgular verilmekte, 4. Bölümde ise sonuçlar tartışılmaktadır.

2. Önerilen Ezgi Modeli

2.1. Yöntem

Konuşma sentezlemede en çok kullanılan tekniklerden biri eklemeli sentezlemedir [6]. Önceden kaydedilen konuşma parçaları bu yöntemde, uygun süre, perde frekansı ve enerji düzenlemelerinin ardından uç uca eklenir. Çalışmamızda konuşma parçası olarak ikili sesbirim de diyebileceğimiz difonlar kullanılmaktadır. Difon; bir fonemin ortasından takip eden fonemin ortasına kadar olan ses parçası [6] olduğu için, ortalama difon süreleri, fonem sürelerinin ortalaması ile hesaplanmakta [11] ve bunlar yaygın olarak kullanılan

PSOLA (Pitch Synchronous OverLap and Add) yöntemiyle [20] birleştirilmektedir. Bu yöntemde perde işaretleri adı verilen yerler referans alınarak ses parçaları birleştirilir.

Türkçe'de yer alan 29 harf ve 44 fonem [21] Tablo 1'de verilmiştir. Sentezin doğru ve doğal olması için ilk şart, difonların doğru belirlenmesidir. Şekil 1'de, "b a1" ile "a1k2" difonlarının perde işaretleri, Şekil 2'de ise bu difonların PSOLA yöntemiyle 6 perde örtüştürülerek birleştirilmesi gösterilmiştir.

Tablo 1: Türkçedeki harf ve fonem(ses birim)ler

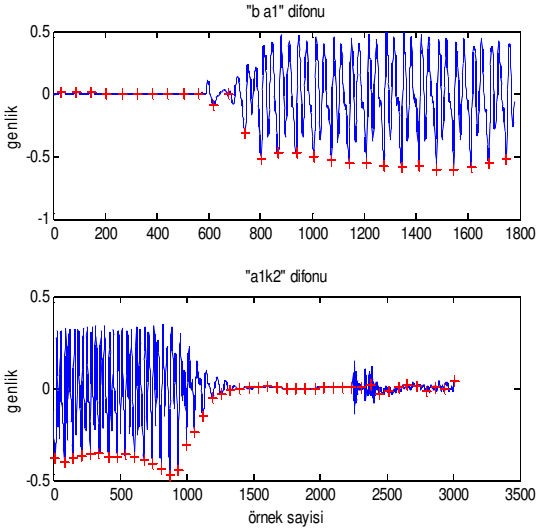
Harf	IPA*	Fonem	Örnek
a	ɑ	a1	**a**1nı
	a	a2	l**a**2f
e	ɛ	e1	m**e**1ç
	e	e2	d**e**2vam
ı	Ï	ı	**ı**slak
i	i	i1	**i**1çecek
	ɪ	i2	**i**2tibar
o	ɔ	o1	s**o**1ru
	o	o2	**o**2ymak
ö	œ	ö1	**ö**1rtü
	ø	ö2	**ö**2ğren
u	ʊ	u1	k**u**1lak
	u	u2	**u**2ğrak
ü	ʏ	ü1	**ü**1mit
	y	ü2	d**ü**2ğme
b	b	b	**b**al
c	ʤ	c	**c**am
ç	ʧ	ç	se**ç**im
d	d	d	**d**emet
f	f	f	**f**asıl
g	ɟ	g1	**g**1ümüş
	g	g2	kar**g**2a
h	h	h	**h**ava
j	ʒ	j	**j**eodezi
k	c	k1	**k**1edi
	k	k2	a**k**2ıl
l	l	l1	**l**1eman
	ɬ	l2	ku**l**2
m	m	m	**m**akarna
n	n	n1	a**n**1ı
	ɳ	n2	sü**n**2gü
p	p	p	**p**ırasa
r	r	r1	**r**1af
	ɾ	r2	ka**r**2şı
	ɣ	r3	da**r**3
s	s	s	**s**ert
ş	ʃ	ş	a**ş**ı
t	t	t	**t**ebeşir
v	v	v1	**v**1ar
	ʋ	v2	ta**v**2uk
y	j	y1	**y**1atak
	ːɪ	y2	du**y**2
z	z	z1	ya**z**1lık
	ʐ	z2	ka**z**2

* IPA: International Phonetic Alphabet

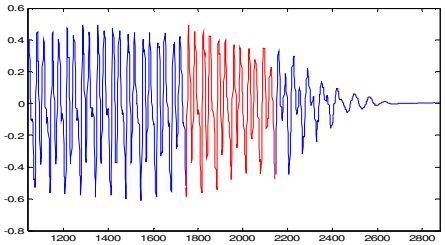
Burada yapılan işlem; birinci difonun sonundan 6 perde, ikinci difonun başından 6 perde almak, bu konuşma parçalarını Hanning pencerenin azalan (birinci difon) ve artan (ikinci difon) bölümleri ile çarpmak, örtüştürmek ve toplamaktır. Hanning pencere Eşitlik (2) ile verilmiştir (*N*; pencerenin boyudur).

$$w(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) \tag{2}$$

Şekil 3'te bu işlem boyunca kullanılan örnek dalga şekilleri görülebilir.



Şekil 1: "b a1" ve "a1k2" difonlarının perde işaretleri



Şekil 2: "b a1" ve "a1k2" difonlarının PSOLA yöntemiyle birleştirilmesi

Eğer birleştirilen ses parçalarının enerjileri arasında seviye farkı varsa, bu da kalitenin düşmesine sebep olacaktır. İzgesel zarf uyumsuzluğu bu çalışmanın kapsamı dışındadır. Ancak temel frekans ve enerji uyumsuzlukları çalışma kapsamında giderilmektedir.

Difonların temel frekansı; perde işaretleri arasındaki farkın (perde periyotlarının) ortalamasının tersi alınarak

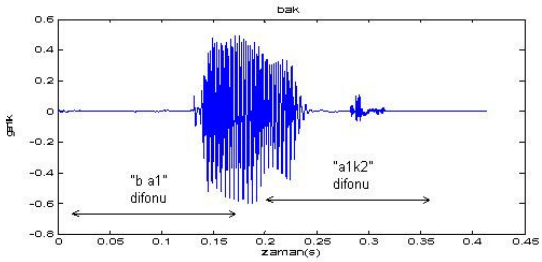
hesaplanmakta ve komşu difonların temel frekansları arada bir değerde eşitlenmeye çalışılmaktadır. Enerji uyumsuzluğu ise komşu difonların enerji oranları kullanılarak giderilmektedir. Eşitlik (3) ve (4)'te bu işlem anlatılmaktadır.

$$E_1=\frac{1}{K}\sum_{n=1}^{K} d_1^2(n) \qquad E_2=\frac{1}{L}\sum_{n=1}^{L} d_2^2(n) \tag{3}$$

E_1 ve E_2; sırasıyla 1. difonun ve 2. difonun ortalama enerjileridir. K ve L; difonların uzunluklarıdır. 2. difon (d_2), enerji oranının karekökü olan katsayı (α) ile çarpılır ve enerjisi eşitlenen yeni difon (s_2) elde edilir (Eşitlik (4)).

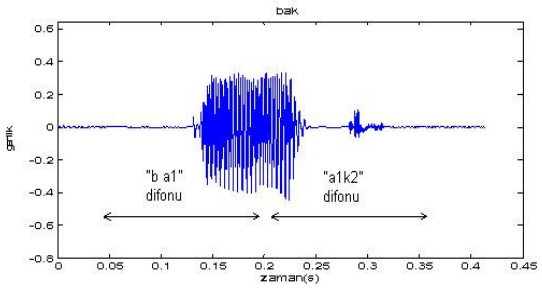
$$\alpha=\sqrt{\frac{E_2}{E_1}} \qquad s_2=\alpha . d_2 \tag{4}$$

Şekil 4'te enerji uyumsuzluğu olan difonlarla yapılan sentez gösterilmiştir.

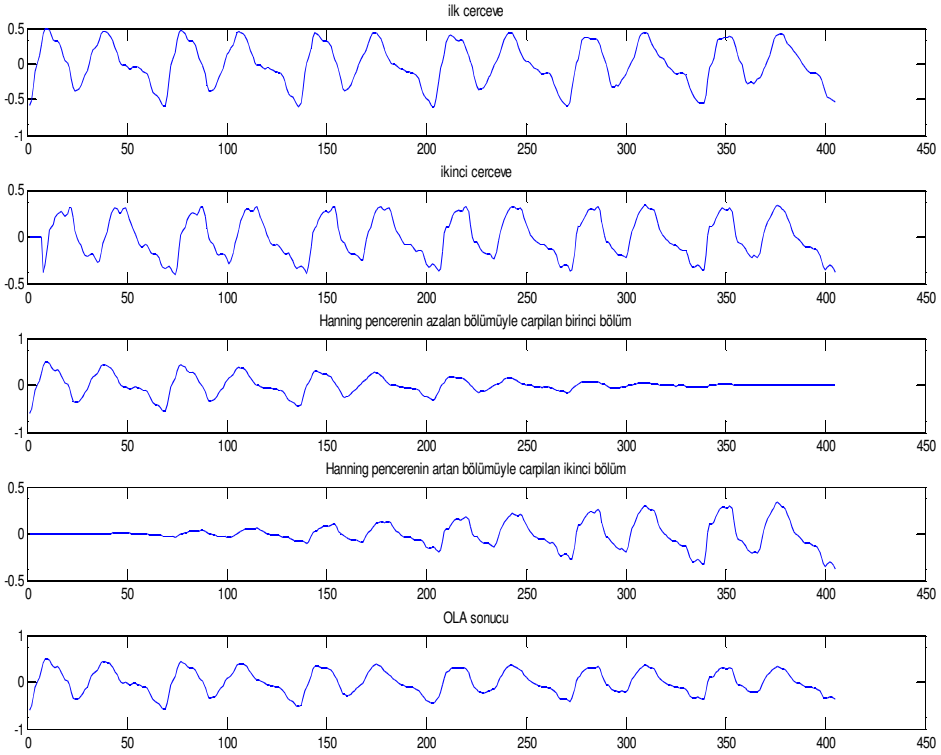


Şekil 4: Enerji eşitlenmemiş durumda sentez

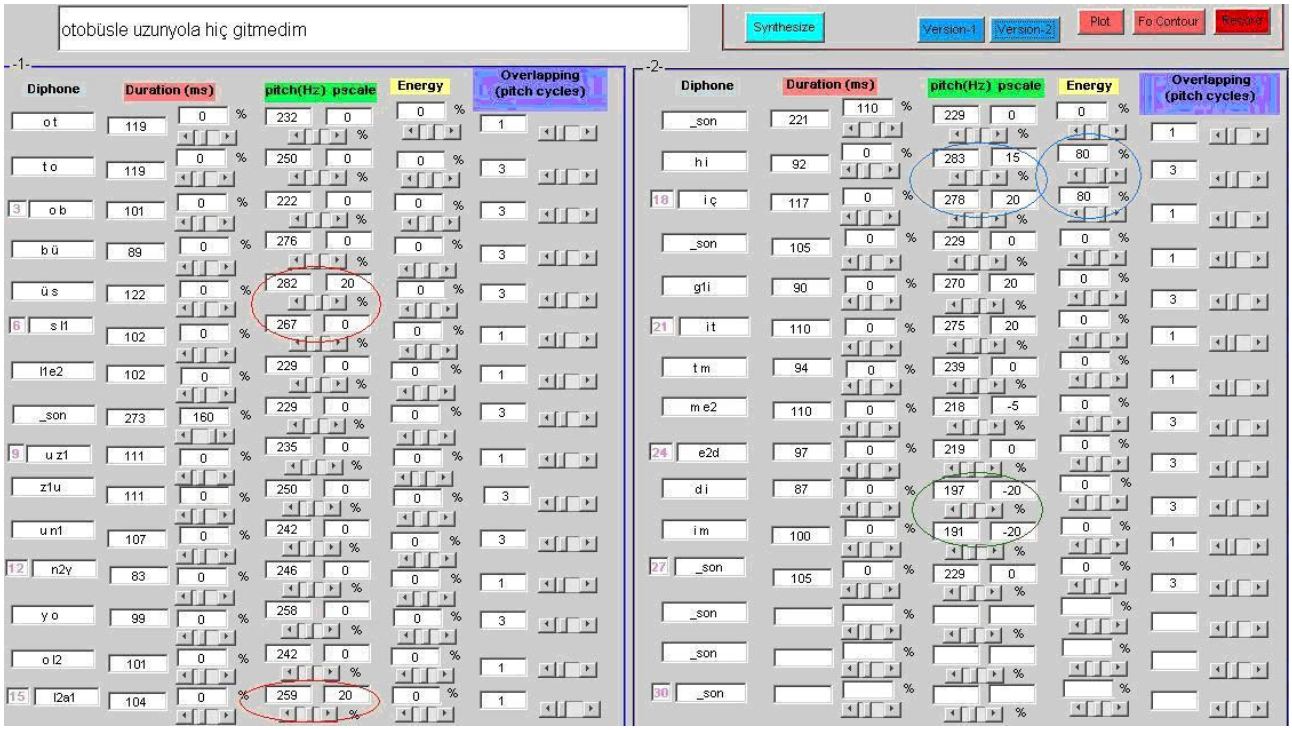
Şekil 5'te ise enerjileri eşitlenmiş difonlarla yapılan sentez gösterilmiştir.



Şekil 5: Enerji eşitleme sonrası sentez



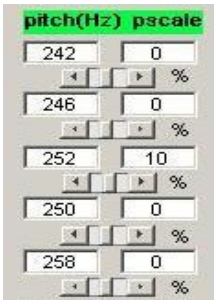
Şekil 3: PSOLA ile yapılan örtüştürüp ekleme işleminin ayrıntıları



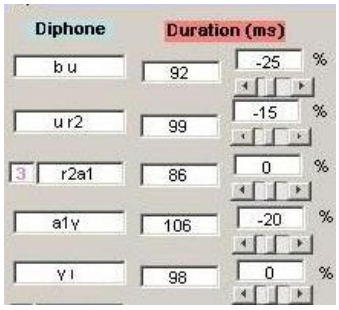
Şekil 6: Tasarlanan ve gerçekleştirilen MKS test platformu

Çalışmada, Şekil 6'da gösterilen ara yüz tasarlanmış ve ezgi denemeleri için bir platform oluşturulmuştur. Matlab GUI© ile hazırlanan bu platformda süre, perde frekansı, enerji ve örtüştürme süreleri ayarlanabilmektedir. Süre ve perde frekansı değiştirme işlemleri yine PSOLA yöntemiyle yapılmaktadır [6].

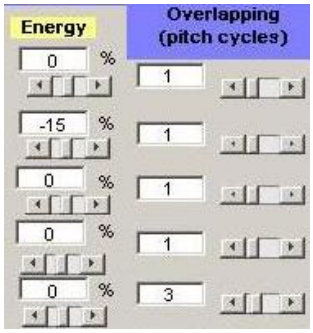
Sol üstte bulunan metin kutusuna girilen yazı, otomatik olarak sözcüklerine ve difonlarına ayrılır. Sözcükler boşluk karakterinden yararlanılarak, difonlar ise Türkçenin telaffuz kurallarına [21] göre belirlenir. Daha sonra difonlar veri tabanından çağrılır. Geliştirilen ara yüzde bulunan süre ayarı Şekil 7'de, perde frekansı ayarı Şekil 8'de, enerji ve örtüştürme süresi ayarı da Şekil 9'da gösterilmiştir.



Şekil 8: Perde frekansı ayarı



Şekil 7: Difonların süre ayarı



Şekil 9: Enerji ve örtüştürme süresi ayarı

2.2. Terminoloji ve Gösterimler

Fonem, difon, ve sözcük için süre, difonlar için perde frekansı ve enerji tanımları aşağıda Tablo 2'de verilmiştir.

Tablo 2: Matematiksel gösterim

$t^{(d)}_{i,j}$	*i*. sözcükteki *j*. difonun süresi (*d*: difon)
$t^{(f)}_{i,j,k}$	*i*. sözcükte *j*. difonun *k*. foneminin süresi (*f*: fonem)
$t^{(s)}_{i}$	*i*. sözcüğün toplam süresi (*s*: sözcük)
$p^{(d)}_{i,j}$	*i*. sözcükteki *j*. difonun orijinal perde frekansı
$(p^{(d)}_{i,j})'$	*i*. sözcükteki *j*. difonun değiştirilen perde frekansı
$e^{(d)}_{i,j}$	*i*. sözcükteki *j*. difonun orijinal enerjisi
$(e^{(d)}_{i,j})'$	*i*. sözcükteki *j*. difonun değiştirilen enerjisi

Sentezleme için yapılan ilk işlem; metinden belirlenen difonların ortalama difon sürelerine otomatik olarak getirilmeleridir. Bunun için, fonem ortalama sürelerinin [11] ortalaması alınır.

Difon ortalama süre hesabı Eşitlik (5)'te görülmektedir.

$$t^{(d)}_{i,j} = \left(t^{(f)}_{i,j,1} + t^{(f)}_{i,j,2}\right)/2 \tag{5}$$

Burada; $t^{(d)}_{i,j}$; *i*. sözcükteki *j*. difonun süresidir. $t^{(f)}_{i,j,1}$ *ve* $t^{(f)}_{i,j,2}$; *i*. sözcükteki j. difonun *k*. fonem süresidir (k = 1: baştaki, k = 2: sondaki fonem).

Sözcüklerdeki difon sayısına göre; difon süresi değiştirme parametresi: α_d, Eşitlik (6)'daki gibi uygulanır.

$$t^{(s)}_{i} = \alpha_d \sum_{k=1}^{K} t^{(d)}_{i,j} \tag{6}$$

Burada, $t^{(s)}_{i}$; *i*. sözcüğün toplam süresidir (bk. Şekil 7) ve α_d parametresinin değeri Tablo 3'te verilmiştir. Eğer sözcükteki difon sayısı 4'ten az ise bu sözcüğün difonları %20 uzatılır (α_d = 1,20); burada amaç sentezlenen sözcüğün yutulmaması ve anlaşılırlığının artırılmasıdır. Eğer sözcükteki difon sayısı 6'dan fazla ise, difonların süresi %5 azaltılır (α_d = 0,95); burada da amaç yavaş okumanın önüne geçmektir. Eğer difon sayısı 4,5 veya 6 ise difonlar ortalama sürelerinde bırakılır (α_d = 1,00). Bu değerlere yapılan denemelerin sonucunda karar verilmiştir.

Tablo 3: Difon süresi değiştirme parametresi (α_d)

α_d = 1,20	Difon sayısı ≤ 3
α_d = 1,00	4 ≤ Difon sayısı ≤ 6
α_d = 0,95	Difon sayısı ≥ 7

Perde frekansı değişikliği için Eşitlik (7)'deki işlem yapılır.

$$\left(p^{(d)}_{i,j}\right)' = \beta_k \left(p^{(d)}_{i,j}\right) \tag{7}$$

β_k; perde frekansı değiştirme oranı olup, σ_k; yüzde değiştirme miktarından $\beta_k = 1 + \sigma_k$ ile elde edilir (bk. Şekil 8).

Örnek olarak; σ_k = –0,15 için, β_k = 0,85 olacak ve *k*. difonun perde frekansı %15 azaltılacaktır.
Enerji ayarı için Eşitlik (8)'deki işlem yapılır.

$$\left(e^{(d)}_{i,j,k}\right)' = \gamma_k \left(e^{(d)}_{i,j,k}\right) \tag{8}$$

γ_k; enerji değiştirme oranı olup, λ_k; yüzde değiştirme miktarından $\gamma_k = 1 + \lambda_k$ ile elde edilir (bk. Şekil 9).

Örnek olarak; λ_k = +0,20 için, γ_k = 1,20 olacak ve *k*. difonun enerjisi %20 artırılacaktır. Bunun için ilgili difon $\alpha = \sqrt{\gamma_k}$ katsayısı ile çarpılır.

2.3. Süre ve Vurgunun Ayarlanması

Türkçede vurgu; zaman ve şahıs eklerinin özelliklerine göre yer değiştirebilmektedir [16]. Bazı ekler vurguyu kendinden önceki ek veya hecelere kaydırırken, bazıları da vurguyu üzerine çekmektedir. Örnek olarak: "***sev**miyorum*" fiilinde vurgu "*–me*" olumsuzluk ekinden önce yer alıyorken, "*gitti**ler** mi?*" fiilinde vurgu –mi soru ekinden öncedir.

Vurgu incelememizde, çekimli fiillerin vurgu alan hecesi üzerinde denemeler yapılmıştır. Olumlu, olumsuz, olumlu soru ve olumsuz soru yapısındaki cümleler için vurgunun yeri, hem dilbilimsel açıdan, hem de bilgisayar destekli yazılımlarla incelenmiştir. Yapılan gözlemler ışığında, vurgu en doğru şekilde senteze kazandırılmaya çalışılmıştır. Süre ve vurgu kurallarına göre sentezlenen konuşma, olduğu gibi birleştirilen konuşmayla dinleme testine tâbi tutulmuş ve kuralların sentezin doğallığına olan katkısı değerlendirilmiştir.

İncelenen cümleler Tablo 4'te verilmektedir. 1 ve 2: olumlu, 3 ve 4: olumsuz, 5 ve 6: olumlu soru, 7 ve 8: olumsuz soru formundadır ve 9 numaralı cümle kontrol amaçlıdır. Kontrol cümlesinin ham ve kurallı halleri tıpa tıp aynıdır ve amaç dinleyicilerin dikkatini ölçmektir. Bu cümleye +2,+1,0,-1 veya -2 dışında puan vermiş olanların test sonuçları değerlendirmeye alınmamıştır.

Tablo 4: Süre ve vurgunun birlikte incelendiği cümleler

No	Cümle
1	*Her şeye rağmen zamanında geldi.*
2	*Çok çalıştığı için başarılı oldu.*
3	*Otobüsle uzun yola hiç gitmedim.*
4	*Yıllardır güneş yüzü görmedi.*
5	*Son sınava yeterince çalıştın mı?*
6	*Biz yokken kendine iyi baktın mı?*
7	*Görevini en iyi şekilde yapmadın mı?*
8	*Saçımı sana süpürge etmedim mi?*
9	*Peki, yeterince çalışmıyor musun?*

Öncelikle, ortalama difon süreleriyle sentez yapılır. Birleştirmede bütün difonların perde frekansları ve enerjileri tasarlanan ara yüzün hassasiyeti ölçüsünde eşitlenir. Bu senteze *ham sentez* adı verilir. Daha sonra ham sentez üzerinde aşağıda maddeler halinde verilen kurallar uygulanır. Burada vurgunun konumunun sisteme girilmiş olduğu varsayılmaktadır.

Bir cümle okunurken, anlamda etkili olduğu için, cümlede yer alan tamlamalara özel bir vurgu katarız. Seçilen cümlelerde böyle tamlamalara ("*uzun yola*", "*güneş yüzü*", "*en iyi şekilde*" vb.) yer verilmiş ve doğru ezgilendirme için yöntem araştırılmıştır. Cümlede öbek vurgusu olarak bilinen bu bölümler belirgin şekilde vurgulanmalıdır. Çalışmamızın en çok zaman alan ve en önemli bölümlerinden birisi de bu bölümdür. Öbek vurgusu için hem temel frekansın hem de enerjinin diğer vurgulara göre daha fazla artırılması önerilmektedir.

Seçilen cümlelere doğal ezgi kazandırmak için izlenen yol şu şekilde sıralanabilir:

- Difon süreleri, *Tablo 3*'teki gibi ayarlanır.
- Söz gruplarının belirlediği duraklar boşluk süresi ile ayarlanır. Buralarda boşluk difonunun süresi iki katına çıkartılır.
- Ham sentezi oluşturan tüm difonların temel frekansları ve enerjileri eşitlenmeye çalışılır.
- Orijinal kayıtlarda ortak olan vurgular senteze verilmeye çalışılır. Bunun için, öbek vurgusunun yer aldığı difonların hem perde frekansları (%30), hem de enerjileri (%70) artırılır (Şekil 6'da mavi işaretli bölüm).
- Sözcük vurgularında ise sadece temel frekans %20 artırılır (Şekil 6'da kırmızı işaretli bölüm).
- Cümlenin fiiline vurgu katılır. Bu amaçla, fiilin vurgulu hecesindeki difonların temel frekansları %20, enerjileri %40 artırılır.
- Ayrıca cümle biten ezgi ile sonlandırılır. Bunun için, son sözcükteki difonların temel frekansları ve enerjileri kademeli olarak (sırasıyla, %10, %15, %20) azaltılır (Şekil 6'da yeşil işaretli bölüm).

Bu önerileri oluşturan perde frekansı ve enerji oranlarına, gerçekleştirilen ara yüzde yapılan pek çok denemeden sonra karar verilmiştir. Farklı değerlerle yapılan sentezler dinlenmiş ve dinleme testleri sonucunda en çok beğenilen orana karar verilmiştir. Tüm bu ayarlamalar elle yapılsa da otomatik hale getirilebileceği düşünülmektedir.

Bu çalışmada, [22]'de oluşturulan 16 kHz ile örneklenmiş difonlar kullanılmıştır. Ayrıca yapılan tüm sentezler uzunluğu 3 örnek olan yumuşatma (smoothing) süzgecinden geçirilmiştir. Bu süzgecin amacı; tıslama ve çatırtı seslerinin etkisini azaltmaktır.

3. Bulgular

Bu bölümde, yukarıdaki bilgiler ışığında yapılan sentezler, web üzerinden CMOS (Comparative Mean Opinion Score testi, ITU-T P.800 standardı olan MOS testinin karşılaştırmalı bir türüdür) testine tâbi tutularak, belirlediğimiz ezgi kurallarının doğallığa olan etkileri ölçülmüştür. Bu amaçla [23] ile adresi verilen ağ sayfası tasarlanmıştır. Dinleyiciler ekrana rastgele sırada gelen ham sentez ile ezgi eklenmiş sentezi dinlemiş ve birbiriyle karşılaştırmışlardır. -5 ile +5 arasında puan vererek hangisinin ezgili sentez olduğunu bilmeden dereceli bir kıyaslama yapmışlardır. Bu öznel teste 40 dinleyici katılmıştır. Elde edilen bulgular Tablo 5'te görülmektedir.

Bu sonuçlardan tüm ezgili sentezlerin ham sentezlerden daha doğal ve başarılı bulunduğu görülmektedir. En yüksek puan; olumsuz cümlelerde (3 ve 4) elde edilmiştir (ort. 2,55/5,00). Daha sonra en yüksek puan; olumlu soru cümlelerinde (5 ve 6) elde edilmiştir (ort. 2,39/5,00). Vurgu yerleri bariz belli olmayan olumlu cümlelerde (1 ve 2) ise ortalama 1,70/5,00'lik artış sağlanmıştır. En düşük artış ise ortalama 0,80/5,00 ile olumsuz soru cümlelerinde (7 ve 8) elde edilmiştir. Tablo 4'teki 7 numaralı cümlenin en düşük puanı almasına süre ve temel frekans değişiklikleri sonucunda, "*görevini en iyi …*" bölümünde anlaşılırlığın azalmasının neden olduğu düşünülmektedir.

Tablo 5: CMOS testi sonuçları

No	Cümle	Puan / 5,0
1	*Her şeye rağmen zamanında geldi.*	1,95
2	*Çok çalıştığı için başarılı oldu.*	1,45
3	*Otobüsle uzun yola hiç gitmedim.*	2,00
4	*Yıllardır güneş yüzü görmedi.*	3,09
5	*Son sınava yeterince çalıştın mı?*	2,32
6	*Biz yokken kendine iyi baktın mı?*	2,45
7	*Görevini en iyi şekilde yapmadın mı?*	0,32
8	*Saçımı sana süpürge etmedim mi?*	1,27
9	*Peki, yeterince çalışmıyor musun?*	0,14

4. Sonuçlar ve Tartışma

Bu çalışmada Türkçe metinden konuşma sentezleyiciler için vurgu kuralları araştırılmış, geliştirilen ara yüzde, bir sistematik dâhilinde, ham senteze ezgi verilmeye çalışılmıştır. Dört farklı formdaki (olumlu, olumsuz, olumlu soru ve olumsuz soru) toplam sekiz adet cümle üzerinde uygulanan vurgu kuralları, sentezin doğallığını ve başarısını arttırmıştır. Tüm bu çalışmaların sonucunda elde edilen bulgular CMOS testi ile değerlendirilmiştir. Sonuçta ezgi kuralları uygulanan sentezler, ham sentezlere göre ortalama 1,86/5,00 puan daha başarılı (doğal) bulunmuştur. Belirlenen kurallar, vurgu yerleri belirli olan cümlelerin daha doğal sentezlenmesini sağlamaktadır.

Çalışmalarımız test kümesini genişletmek için devam etmektedir. Gelecek çalışma olarak; sentezleyicide difondan daha büyük konuşma parçalarının kullanılması ve örnekleme frekansının artırılarak sentezlere olan etkisinin incelenmesi önerilebilir. Bu çalışmada önerilen yöntemin otomatikleştirilmesi de bir diğer çalışma konusudur.

Teşekkür

Çalışmalarımıza sesini veren Dr. Özgül Salor'a ve dinleme testlerimize katılan herkese teşekkür ederiz.

KAYNAKLAR

[1] Braille Teknik Ltd. Şti.
http://www.brailleteknik.com/jaws.html
son erişim: 08/02/2012

[2] Loquendo S.p.A., a Telecom Italia Group Company
http://www.loquendo.com/en/demo-center/tts-demo/

[3] GVZ Ses tanıma ve sentezleme teknolojileri şirketi
http://www.gvz.com.tr/index.html
son erişim: 08/02/2012

[4] DİKTE Yöndata Bilgisayar Ltd. Şti.
http://www.dikte.com.tr/konusmatanima.php
son erişim: 08/02/2012

[5] Google translate
http://translate.google.com
son erişim: 08/02/2012

[6] Dutoit, T., *An Introduction to Text-to-Speech Synthesis,* Kluwer Academic Publishers, 1997.

[7] Tatham, M. and Morton K., *Developments in Speech Synthesis*, Wiley, 2005.

[8] Narayanan, S. and Alwan, A., *Text to Speech Synthesis, New Paradigms and Advances*, Prentice Hall, 2005.

[9] Taylor, P., *Text-to-Speech Synthesis*, Cambridge University Press, 2009.

[10] Uslu, İ. B., "Türkçe metinden konuşma sentezlemede bugünkü durum - 2. Bölüm", *EMO Ankara Şubesi Haber bülteni*, 2010.3.

[11] Şayli, Ö., "Duration analysis and modelling for Turkish text-to-speech synthesis", yüksek lisans tezi, Boğaziçi Üniversitesi Fen Bilimleri Enstitüsü, 2002.

[12] Öztürk, Ö., "Modelling phoneme durations and fundamental frequency contours in Turkish speech", doktora tezi, ODTÜ Fen Bilimleri Enstitüsü, 2005.

[13] Oskay, B., Salor, Ö., Özkan, Ö., Demirekler, M. ve Çiloğlu T., "Türkçe metinden konuşma sentezlemede ezgi belirlenmesi ve uygulanması", IEEE 9. Sinyal İşleme ve Uygulamaları Kurultayı SİU-2001, 2001, s. 238–243.

[14] Külekçi, M. O. ve Oflazer K., "An infrastructure for Turkish prosody generation in text-to-speech synthesis", *TAINN 2006, 15th Turkish Symposium on Artificial Intelligence and Neural Networks*, Muğla, Haziran 2006, s. 49–57.

[15] Uslu, İ.B. ve İlk, H.G., "Türkçe metinden konuşma sentezlemede Fujisaki ezgi modeli, *IEEE 17. Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, SİU-2009, Antalya, Nisan 2009, s. 844–847.

[16] Aydemir T. ve Yılmaz, A. E., "Türkçe fiil çekimlerinde vurgu konumunu belirlemek için bir yazılım kütüphanesi, *IEEE 18. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU 2010)*, 22–24 Nisan 2010, Diyarbakır, Türkiye, s. 696–699.

[17] Uslu, İ.B., Yılmaz A.E. ve İlk, H.G., "Türkçe metinden konuşma sentezlemede fiil çekimleri için yeni bir ezgi modeli", *IEEE 19. Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, SİU-2011, Antalya, Nisan 2011, s. 638–641.

[18] Bulut, M., Busso C., Yıldırım, S., Kazemzadeh, A., Lee, C. M., Lee S. and Narayanan S., "Investigating the role of phoneme-level modifications in emotional speech resynthesis", *Proceedings of Interspeech,* 2005, s. 801–804.

[19] Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L. and Auberge, V., "Emotional prosody – does culture make a difference?", *Speech Prosody*, Dresden Germany, paper no. 207, 2006.

[20] Moulines, E., and Charpentier, F., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, volume: 9, 1990, s. 453–467.

[21] Ergenç, İ., *Konuşma Dili ve Türkçenin Söyleyiş Sözlüğü,* Multilingual, 2002

[22] Salor, Ö., Pellom B, Çiloğlu T. and Demirekler M., "On developing new text and audio corpora and speech recognition tools for the Turkish language", ICSLP-2002: Inter. Conf. On Spoken Language Processing, Denver, Colorado USA, 16–20 Eylül 2002, s. 349–352..

[23] http://demo.reformo.net/baran3/index.php
son erişim: 08/02/2012