Metinden Konuşma Sentezleme

I. Bölüm

Elektronik Mühendisi İbrahim Baran Uslu
ibuslu@baskent.edu.tr

Özet

Yazımızda, metinden konuşma sentezleme (text-to-speech synthesis); yâni yazının bir sistem tarafından otomatik olarak okunması incelenmektedir. Bu konuda, hayata geçirilmeyi bekleyen pek çok uygulama bulunmaktadır. Bunlar arasında; sesli yanıt, bilgi ve uyarı sistemleri, görme engelli kişiler için ekran ve kitap okuma programları; konuşma engelli kişiler için sözlü iletişim imkânı, dil öğrenme araçları, tatildeyken birikmiş olan e-postaların size okunması vb. Sayılabilir.

Giriş

Metinden konuşma sentezleme (MKS); yazılı bir metnin konuşma şekline dönüştürülmesidir. Sistemin blok şeması, en basit haliyle, Şekil-1'de verilmiştir (1). MKS sistemleri iki ana bölümden oluşur: Dil Çözümleme Modülü ve Sinyal İşleme Modülü. Dil çözümleme modülünde, konuşmaya çevrilecek metin ön işlemlerden geçirilir. Burada amaç; metinden sentezlenecek konuşmayı oluşturan ses bilgilerinin elde edilmesidir. Eğer metinde kısaltmalar varsa ('Dr.', 'Sn.', 'Doç.' gibi); bunlar açık yazılışlarındaki hallerine çevrilmelidir. Ayrıca sayılar (tarihler, 1. gibi kısaltmalar, vd.) da metne dönüştürülmelidir. Sinyal işleme modülü ise; sesbilgisel ve bürünsel bilgileri kullanarak konuşmayı oluşturur.

Şekil-1 MKS sistemi blok şeması

Dünya literatürüne bakıldığında, metinden konuşma sentezlemede üç yaklaşım geliştirildiği görülür. Bu yaklaşımlar; 1- Kural-tabanlı formant sentezleyiciler, 2- Söyleyiş (articulatory) sentezleyicileri, 3- Eklemeli (concatenative) sentezleyicilerdir. Birinci grup sentezleyiciler konuşma sinyalinin doğrusal öngörümlü kodlaması (LPC, Linear Predictive Coding) temeline dayanmaktadır. Bu ise; Eşitlik-1'de görüldüğü gibi, konuşma sinyalinin n. örneğinin (ŝ(n)), önceki p adet örneğin doğrusal kombinasyonu şeklinde ifade edilmesidir.

$$\hat{s}(n) = -\sum_{i=1}^{p} a_i s(n-i) \qquad (1)$$

Konuşmayı; durağan kabul edildiği 20-30 ms'lik kısa sürelerde, oldukça başarılı bir şekilde kodlayabilen bu yöntemde, ai' ler; LPC katsayıları olarak adlandırılır ve bu süre zarfında (çerçeve süresi) sabittir. Konuşmanın gerçek değeri ile öngörülen değeri arasındaki fark; hata sinyalidir ve sentezlemede uyartım sinyali olarak adlandırılmaktadır (bakınız Eşitlik-2).

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{i=1}^{p} a_i s(n-i) \qquad (2)$$

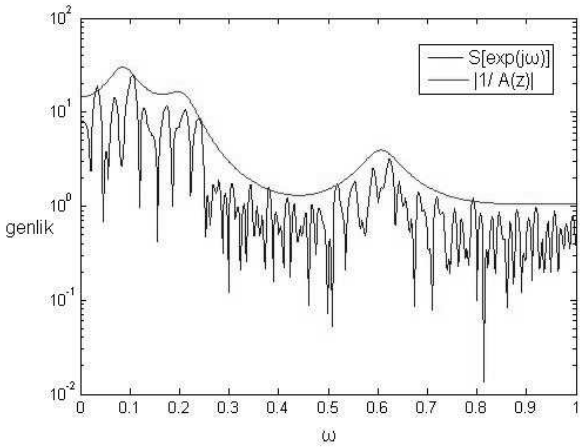
Hata sinyali ile birlikte LPC eşitliğimiz Eşitlik-3'teki gibi olacaktır.

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + e(n) \tag{3}$$

(3)'ten z-dönüşümü ile aktarım işlevi (transfer fonksiyonu) hesaplanırsa; sonlu uzunlukta dürtü tepkili bir süzgeç elde edilir: A(z)-LPC analiz filtresi (4). Süzgeç katsayıları (ai); hatanın karesini minimum yapacak şekilde hesaplanır. A(z) süzgeci; ses yolunu (vocal tract) modeller. Böylece konuşma sinyali, incelendiği zaman aralığında, uyartım sinyaline ve süzgeç katsayılarına (ai) ayrıştırılmış olur.

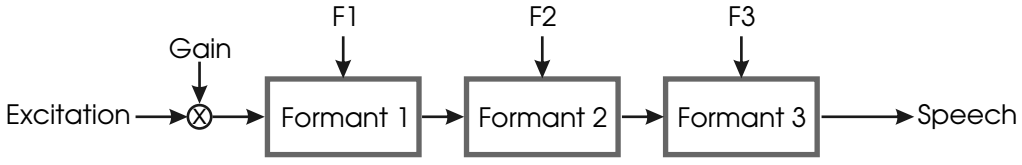
$$\frac{E(z)}{S(z)} = A(z) = 1 + \sum_{k=1}^{p} a^k z^{-i} \tag{4}$$

Analiz süzgecinin genlik tepkisinde yer alan tepeler; ses yolunun rezonans frekanslarına karşılık gelir ve formant frekansları olarak adlandırılmaktadır (bakınız Şekil-2). İşte formant sentezleyiciler; konuşmayı bu ana bileşenleri (formant frekanslarını ve bant genişliklerini) ayarlayarak sentezlemeye çalışır. Tarihsel olarak en eski geliştirilen sentezleyiciler formant sentezleyicilerdir.

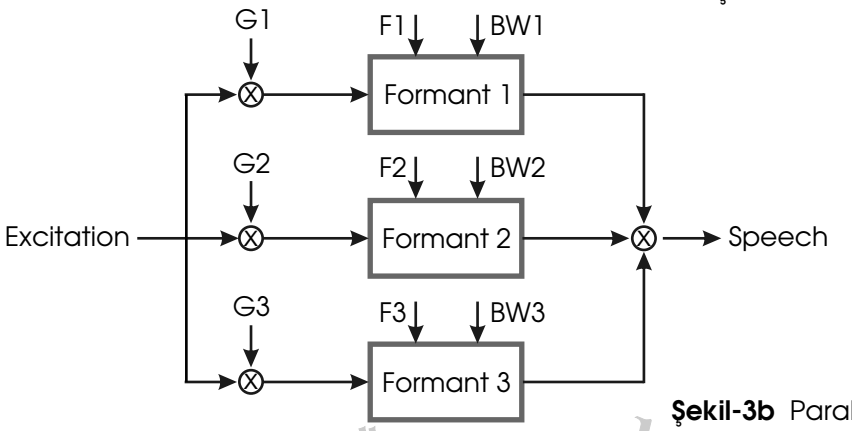


Şekil-2 Formant frekanslarının LPC analiz filtresinin genlik spektrumundan elde edilmesi

Seri formant sentezleyiciler, denetim bilgisi olarak sadece formant frekanslarına ihtiyaç duyar (bakınız Şekil-3a). Ünlülerin sentezinde, formant frekanslarındaki genlik değerlerini ayrı ayrı kontrol etmek gerekmez. Seri formant sentezleyicilerin sürtünücü ve patlamalı ünsüzleri sentezlemede çok başarılı olmadığı belirtilmektedir (2). Paralel yapıya göre daha az denetim bilgisine sahip olduğu için, daha kolay gerçekleştirilebilir.



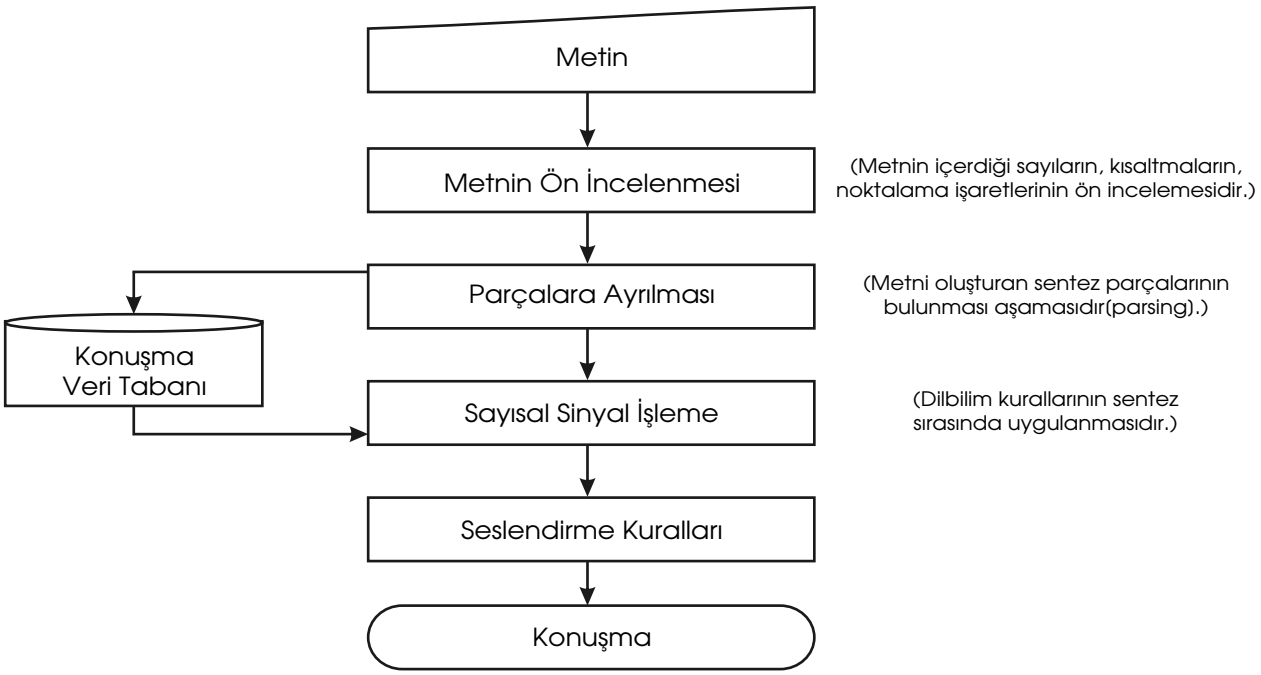
Şekil-3a Seri formant sentezleyici



Şekil-3b Paralel formant sentezleyici

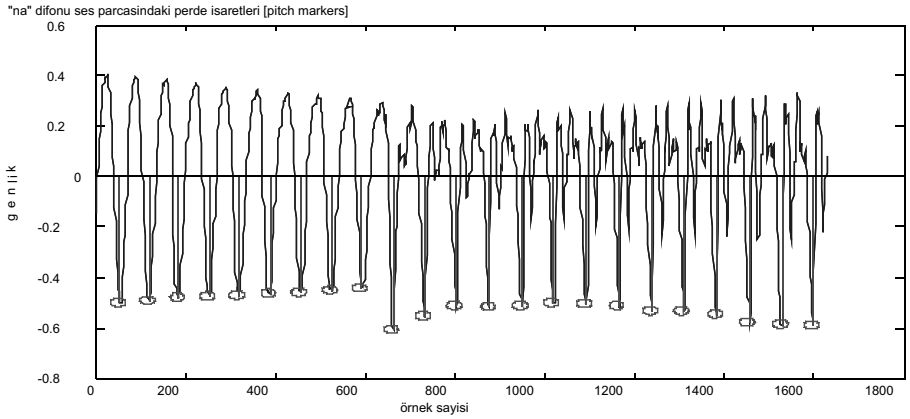
Paralel formant sentezleyiciler, her bir formant frekansı için ayrı frekans, genlik ve bandgenişliği parametrelerine ihtiyaç duyar (bakınız Şekil-3b). Dolayısıyla daha karmaşık yapıdadır. Geniz (nasal) ünsüzlerinin ve sürtünücü ünsüzlerin sentezinde daha iyi sonuçlar verdiği, ancak ünlülerin sentezini gerçekleştiremediği belirtilmektedir (2).

2. grupta yer alan söyleyiş sentezleyicilerinde amaç; tüm ses birimlerin (phoneme), insanın ses üretim mekanizmasında nasıl oluşturulduğunun en hassas şekilde modellenmesidir. Teorik olarak en doğru yaklaşım söyleyiş sentezleyicileridir, fakat; ses yolunda bulunan organların davranışlarını modellemek pek de kolay değildir. Çünkü bu organların (ses telleri, ağız ve geniz boşlukları, dil, dudaklar, vd.) görüntülerini elde etmek zordur. Tahmin edileceği üzere çok sayıda parametre içeren bu modelin elde edebildiği konuşma kalitesi formant sentezleyicilerinkinden daha iyidir; özellikle konuşmanın geçiş bölümlerinde, fakat işlem ödünleşimi (tradeoff) de bir o kadar fazladır.
3. grup sentezleyiciler ise eklemeli (concatenative) sentezleyiciler olarak bilinir. Bu gruptaki sentezleyiciler; bellek kapasitelerinin ve sayısal sinyal işlemci hızlarının artmasına paralel olarak tercih edilmektedir. Bunlar, konuşmayı; sesbirim, difon, trifon, seslem vb. gibi önceden kaydedilen ses parçalarını belirli sinyal işleme teknikleriyle biraraya getirerek oluşturmaya çalışırlar. Bu amaçla, ilk önce birleştirilecek parçaların veritabanı hazırlanmalıdır. Kayıt aşaması zahmetlidir, ne var ki; hedef ses parçalarının kaydedilen konuşma parçalarından ayrılması daha vakit alıcıdır. Bu işlemi belirli bir başarıyla otomatik olarak yapabilen programlar vardır. Eklemeli sentezlemede yer alan temel adımlar Şekil-4 ile gösterilmektedir.

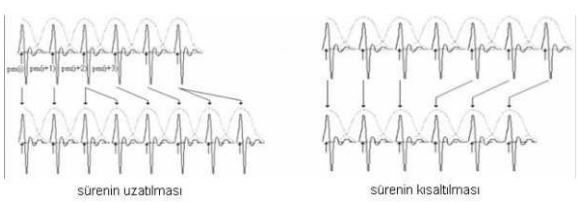


Şekil-4 Eklemeli sentezlemede işlemlerin akış şeması

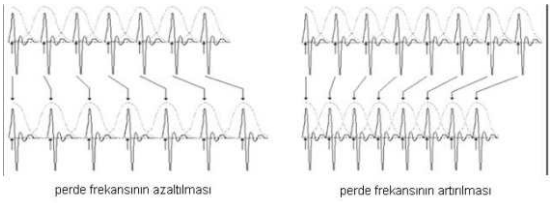
Konuşmayı oluşturacak ses parçalarının belirlenmesinin ardından, ikinci adım olarak sayısal sinyal işleme bloğuna geçilir. Ses parçaları birbirlerine fazları, frekansları ve enerjileri uyumlu olacak şekilde birleştirilmelidir. Bunu sağlayan yöntemler; örtüştürmeli ekleme (OLA: OverLap Add) olarak bilinir. Yapılan işlem; çerçevelerin bir ağırlıklandırma penceresi (genellikle Hanning pencere) ile çarpılması ve yumuşatılan bu bölümlerin örtüştürülüp (genellikle % 50 örtüşme) toplanarak birleştirilmesidir. Bu yöntemin bilinen ve en çok kullanılan türleri: PSOLA (Pitch Synhronous OverLap Add) ve WSOLA (Waveform Similarity OverLap Add) yöntemleridir (3, 4). PSOLA yönteminde perde işaretleri referans alınır (bakınız Şekil-5) ve süre değiştirme, perde frekansı modifikasyonu gibi işlemler perde eşzamanlı olarak yapılır. Böylece ses parçaları birbirine faz uyumlu bir şekilde eklenmiş olur. WSOLA yönteminde ise, dalgaşekli benzerliği (maksimum çapraz ilinti) referans alınır. WSOLA'da perde işaretlerine gerek yoktur. Bu yönden PSOLA'ya göre avantajlıdır. PSOLA yöntemiyle süre (bakınız Şekil-6a) ve perde frekansı (bakınız Şekil-6b) ayarlamaları; WSOLA yöntemiyle ise yalnızca süre ayarlaması (bakınız Şekil 7) yapılabilir.



Şekil-5 PSOLA yönteminde kullanılan perde işaretlerinin gösterilmesi

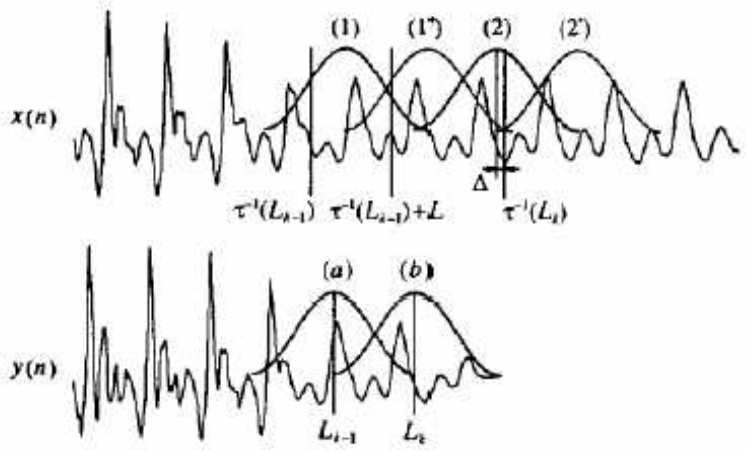


Şekil-6a PSOLA (Pitch Synchronous OverLap and Add) yöntemi ile süre değiştirme (time scaling)



Şekil-6b PSOLA yöntemi ile perde frekansı değiştirme (pitch scaling)

Tüm bu sistemlerin hedefi; insan sesi doğallığında, anlaşılabilir ve kaliteli bir konuşma sentezidir. Bu ise kolay bir problem değildir. Hedeflenen konuşma parametrelerinin elde edilmesi için uygulanan işlemler beraberinde bir miktar bozucu etki getirmektedir. Ayrıca, sentezlenen konuşmanın kalitesinin ölçülmesi de zordur; çünkü konuşma duyumsal açıdan çok boyutlu bir sinyaldir. Yapılan son çalışmalarda; yukarıda özetlenen yaklaşımların birlikte kullanımı, duyumsal maliyet fonksiyonlarının geliştirilmesi, izgel geçişlerin yumuşatılması, konuşmaya duygu katılması gibi konularda araştırmalar dikkat çekmektedir (5, 6). Konuya giriş niteliği taşıyan bu yazının devamında, Türkçe metinden konuşma sentezleme için yapılan çalışmaların özetlenmesi hedeflenmiştir.



Şekil-7 WSOLA yöntemiyle süre değiştirme

Referanslar.....
(1) Thierry Dutoit, An Introduction To Text-To-Speech Synthesis, Kluwer Academic Publishers, 1997.
(1) Lemmetty, S., (1999), "Review of speech synthesis technology", Master Thesis, Helsinki University of Technology, March 1999.
(3) Moulines, E. and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication 9, pp. 453-467, 1990.
(4) Moulines, E. and W. Verhelst "Time-domain and frequency-domain techniques for prosodic modification of speech" in Kleijn and Paliwal (eds.), Speech Coding and Synthesis, pp. 519-555, Elsevier Science B.V., Netherlands, 1995.
(5) Shrikanth Narayanan and Abeer Alwan, Text-To-Speech Synthesis, New Paradigms and Advances, Pearson, 2005.
(6) Mark Tatham and Katherine Morton, Developments in Speech Synthesis, Wiley, 2005.