Disfonik Konuşmanın Yeniden Yapılandırılarak Normal Konuşma Elde Edilmesi

Reconstruction of Dysphonic Speech for Synthesizing Normally Phonated Speech

H. İrem Türkmen, M. Elif Karslıgil

Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
{irem, elif}@ce.yildiz.edu.tr

Özetçe

Bu çalışmada, başlıca nedenleri ses teli felci, gırtlak kanseri ve ses tellerindeki organik lezyonlar olan kronik ses bozukluklarından ötürü sesini kaybetmiş hastaların fısıltı benzeri konuşmalarını yeniden yapılandırarak normal sese yakın hale getiren özgün bir sistem geliştirilmiştir.

Önerilen sistemde KUDÖK (Karma Uyarımlı Doğrusal Öngörüm Kodlaması, MELP-Mixed Excitation Linear Predictive Coding) yöntemi ile disfonik konuşmadan elde edilen akustik özellikler kullanılarak ötümsüz fonem içeren çerçeveler tespit edilmiş, bu çerçeveler dışındaki çerçeveler için formant frekansları ile perde bağlantısı kullanılarak perde oluşturulmuş, formant yapısı değişikliği yapılmış ve seslilik eklenmiştir. Perde üretimi için perde-formant frekansı ilişkisinden yararlanılmıştır. Değiştirilmiş akustik özellikler ile yeni sesin sentezi için KUDÖK kullanılmıştır.

Geliştirilen sistem ile elde edilen sentetik sesin kalitesinin incelenmesi için spektral uzaklık hesabı yapılmış ve öznel dinleyici testlerine başvurulmuştur. Testler sonunda sentetik sesin özellikle tanınabilirlik, doğal sese yakınlık ve tonlama açısından disfonik sese göre tercih edilebilir olduğu saptanmıştır.

Abstract

In this study, a novel system, delivering synthetic speech with the quality near to natural, is designed and implemented by reconstructing dysphonic speech of patients that have lost their voice totally due to apoplectic chordae vocalis, organic lesions of vocal cords or partial laryngectomy.

In the proposed system, MELP (Mixed Excitation Linear Prediction) is used for synthesizing the normal speech. The unvoiced phonemes are determined in dysphonic speech and synthetic pitch is generated by using pitch-formant frequency relation and then formant distortion modification is applied and voicing is added for the phonemes other than unvoiced phonemes. MELP will be used for synthesizing enhanced speech by using modified acoustic features.

Spectral distance measurement is made and subjective listening tests are applied for assessing the produced synthetic speech by our proposed system. Our tests show that, synthetic speech produced this way is preferable when compared to dysphonic speech especially in terms of timbre, recognition and naturality.

1. Giriş

Gırtlakta oluşan enfeksiyonlar, gırtlak kanseri ve ses tellerindeki organik lezyonlar nedeniyle her yıl yüzbinlerce insan konuşma yeteneğini kaybetmektedir. Gırtlak kanserinin tedavisinde uygulanan larinjektomi (gırtlağın cerrahi yolla alınması) operasyonundan sonra hastalarda sesin düzeltilmesi için kullanılan üç yöntem vardır[1]. Birinci yöntem yapay gırtlak kullanmaktır (alaryngeal konuşma). İkinci yöntemde hasta boynun alt kısmındaki delikten aldığı havayı, ağızdan çıkarmayı öğrenerek ses çıkartır (özefageal konuşma). Üçüncü yöntemde ise larinjektomi sonrası, gırtlağa bir protez yerleştirilir (Trakeo-Özefageal konuşma). Ses tellerinde lezyon oluşumu bulunan ve konuşma ve solunum fonksiyonlarının devamı için kısmi larinjektomi geçirmiş hastalarda ise bu yöntemler çoğu zaman uygulanamamaktadır.

Alaryngeal ve özefageal konuşmanın iyileştirilmesine yönelik sinyal işleme tabanlı çalışmalar son 15 yılda hızlanmış ve çoğunluğu ses değişikliği tekniklerine dayanan birçok sistem tasarlanmıştır. Fakat literatürde bu yöntemlerin uygulanamadığı hasta grubunun çıkardıkları doğal sesi iyileştirmeye yönelik bir çalışma mevcut değildir.

Aguilar ve Nakano, örüntü tanıma yaklaşımı temelli bir alaryngeal konuşma iyileştirme sistemi sunmuşlardır[2]. Önerilen sistemde, konuşmanın sesli kısımları yapay sinir ağları yöntemleri ile bulunur ve normal konuşmanın ilgili segmentleri ile değiştirilir. Bi ve Qi, var olan konuşmacı dönüştürme algoritmalarını alaryngeal konuşmayı iyileştirmek için kullanmışlardır[3]. Spektral bozulmaları vektör nicemleme ve doğrusal çok değişkenli regresyon tabanlı konuşmacı dönüştürme metotlarını kullanarak düzeltmeyi hedeflemişlerdir. Pozo ve Young [4], gırtlaksal dalga formu kullanarak ve seğirmeyi azaltarak trakeo-özefageal konuşmayı yeniden sentezlemişlerdir. Qi ve Winberg [5] ise kadın trakeo-özefageal konuşmacıların seslerini gırtlaksal dalga formu ile yeniden sentezlemiş ve perde yumuşatması yapmışlardır. Sawada ve Takeuchi [6], perde periyodu kestirimi ve spektral değişiklik yöntemlerini kullanarak, özefageal konuşmanın iyileştirilmesi ve sinirsel felçli hastaların konuşmalarının düzeltilmesi üzerine çalışmışlardır.

Bu çalışmada, sesini kaybetmiş hastalardan alınan fısıltı benzeri disfonik konuşmayı yeniden yapılandırarak normal konuşma tonuna benzer bir tonda sentetik konuşma üreten özgün bir sistem geliştirilmiştir.

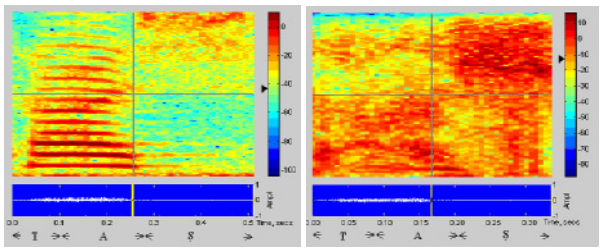
Çalışmanın 2. bölümünde disfonik ses ve normal sesin farklılıkları incelenmiş, 3. bölümde disfonik sesten normal ses

978-1-4244-4436-6/09/$25.00 ©2009 IEEE

elde edilmesi adımları anlatılmış, dördüncü bölümde deneysel sonuçlara yer verilmiş, sonuç bölümünde elde edilen sonuçlar değerlendirilmiştir.

2. Normal Ses ve Disfonik Sesin Akustik Farklılıkları

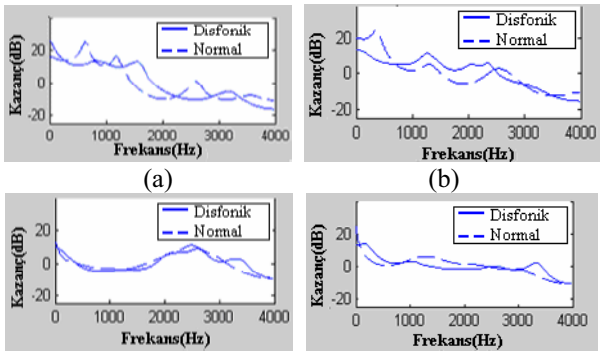
Disfonik ses seslilik, perde ve formant yapısı bakımından normal sese göre farklılık göstermektedir. Disfonik seste algılanabilir bir perde periyodu bulunmaz ve tamamen gürültülüdür. Şekil 1'de disfonik konuşmacı ve normal konuşmacı tarafından söylenmiş /taş/ kelimesi spektrogramları görülmektedir.



(a) (b)

Şekil 1: /Taş/ kelimesi için (a) Disfonik Ses (b) Normal Ses Spektrogramları

Kronik disfonik sesin ve normal sesin formant frekansları ile bant genişliklerinin birbirinden farklılık gösterdiklerini kaydeden çalışmalar mevcuttur [6]. Şekil 2'de görüldüğü gibi, disfonik sese ait ötümlü fonemlerin formant bant genişlikleri genellikle daha geniş ve formant frekansların daha büyüktür. Buna karşın ötümsüz fonemlerin formant yapısında bir bozulma olmadığı görülmektedir.



(a) (b)

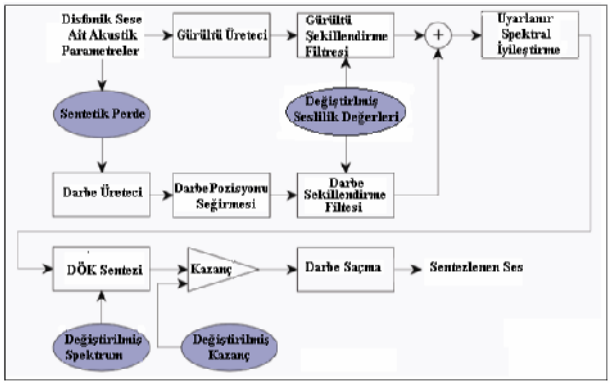
(c) (d)

Şekil2: Ötümlü fonemlerden (a) /A/ , (b) /r/ , ötümsüz fonemlerden (c) /k/ (d) /f/ fonemine ait spektrumlar

Formant yapısındaki farklılıklara ek olarak, disfonik seste algılanabilir bir perde periyodu ve seslilik de bulunmamaktadır.

3. Disfonik Konuşmanın Yeniden Yapılandırılması

Bu çalışmada geliştirilen disfonik ses iyileştirme sistemi, giriş sinyali olan disfonik sesin KUDÖ kodlayıcı kullanılarak analiz edilmesi ile başlar. Analiz sonunda elde edilen ÇSF(Çizgisel Spektrum Frekansları) kullanılarak, giriş sinyalindeki ötümlü sessiz ve sesli içeren çerçeveler tespit edilir. Şekil 3'te maviyle renklendirilmiş işlemler ile bu çerçeveler için kazanç, seslilik, perde ve formant yapısı değişikliği yapılır. Son adımda ise ötümlü sessiz ve sesli içeren çerçevelerden, uyarlanmış parametreleri kullanılarak, ötümsüz sessiz içeren çerçevelerden ise orijinal parametreleri koruyarak sentetik ses elde edilir.



Şekil3 : Disfonik Ses İyileştirme Sistemi Blok Diyagramı

3.1. Disfonik Ses Veri Tabanı Oluşturma

Bu çalışmada, ses teli felçli 3 hastaya ve kısmi larinjektomi geçirmiş 2 hastaya ait sesler incelenmiştir. Her hastadan alınan farklı uzunluklarda, fonetik olarak düzgün dağılımlı, toplamda 50 kelimeden oluşan cümle ve isim tamlaması 8kHz örnekleme frekansında kaydedilmiştir.

3.2. Ötümsüz Fonemlerin ve Tespiti

Konuşma analizine dayanan pek çok sistemde ötümlü-ötümsüz sınıflandırmasına gerek duyulmaktadır. Ötümlü-ötümsüz sınıflandırması için literatürde en sık kullanılan yöntemler perde analizi, özilinti fonksiyonu ve sıfır kesişidir [7]. Aguilar ve Nakano [2], alaryngeal konuşmada ötümlü-ötümsüz sınıflandırması için perde değerinden yararlanmışlardır. Ancak disfonik seste alaryngeal konuşmanın aksine algılanabilir bir perde değeri bulunmamaktadır. Yapılan literatür taramalarında disfonik ses veya fısıltılı ses için ötümlü-ötümsüz sınıflandırması yapan bir çalışmaya rastlanmamıştır.

Bu çalışmada, disfonik sesteki sesli çerçeveler için ötümlü-ötümsüz fonem sınıflandırması yapılmıştır. Sesli çerçevelerin tespiti için giriş sinyalinin ortalama kazanç değerinin 0.6 katı, eşik değeri olarak kullanılmıştır. Oluşturulan sistemin testi için hazırlanan disfonik ses veritabanı kısıtlı sayıda örnek içerdiğinden, eğitim seti, disfonik sesler yerine özellik olarak benzerlik gösteren fısıltılı seslerden oluşturulmuştur. Eğitim seti, bu sınıflara ait her ses için ortalama ÇSF'nı içermektedir. Sınıflandırılacak çerçeveye ait ÇSF kullanılarak, k-en yakın komşuluk yöntemi ile çerçevenin ait olduğu sınıf tespit edilmiştir.

Test sonuçları incelendiğinde tüm fonem grupları için ortalama sınıflandırma başarısının %85 olduğu gözlenmiştir.

3.3. Bant Geçişli Seslilik Ekleme

Yapılan çalışmalarda normal konuşmanın sesli kısımlarında, en düşük dört veya beş frekans bandına (0–4 veya 0-5 kHz)

ait seslilik değerlerinin 1, disfonik konuşmanın sesli kısımlarında ise 0 olarak hesaplandığı gözlenmiştir. Normal sese yakın bir sentetik ses elde edilebilmesi için, ötümlü seslerde (0 – 5 kHz) frekans bandı için darbe uyarımı üretilmelidir. Bu amaçla, disfonik sesler için hesaplanan bant içi seslilik değerleri uyarlanmalıdır. Geliştirilen sistemde, yukarıda anlatılan gözlemlere dayanılarak, ötümlü fonem bilgisi içeren çerçeveler için en düşük dört frekans bandına ait seslilik değeri 1 olarak alınmış, diğer çerçeveler için ise, tüm frekans bantlarına ait seslilik değerleri 0 olarak sabitlenmiştir.

3.4. Perde Ekleme

Kronik disfonik konuşmada algılanabilir bir perde periyodu değeri görülmemektedir. Disfonik sesten normal ses elde edilebilmesi için, kişinin normal konuşması esnasında gözlenebilecek perde periyodu değerlerine uygun, sentetik perde değerleri üretilmelidir. Önerilen sistemde, perde ve kazanç değerleri arasında gözlemlenen ilinti kullanılarak, aşağıdaki eşitlik ile perde tahmini yapılmıştır [8].

$$p^{n}_{yeni} = ((k^{n} - k_{ortalama}) * \beta) + p_{referans} \quad (1)$$

Burada p^{n}_{yeni}, tahmin edilen perde değerini, k^{n} n numaralı çerçeveye ait kazanç değerlerini, $k_{ortalama}$, ortalama kazanç değerini, $p_{referans}$ ise referans perde değerini göstermektedir. $p_{referans}$, sentezlenecek sesin tonunu ayarlamak için, β ise dinamik erimi kontrol etmek için kullanılır. Önerilen sistem, referans perde değerini, otomatik olarak hesaplamaktadır. Fısıltılı seste perde değeri bulunmadığından, disfonik ses sahibine en uygun referans perde değerinin hesaplanması için, fısıltılı /a/ foneminin formant frekanslarından yararlanılmıştır.

Formant frekansları ile perde değerleri arasındaki bağıntıyı inceleyen pek çok çalışma bulunmaktadır [9,10]. Farklı konuşmacıların fısıltılı ve normal sesle söyledikleri /a/ fonemleri incelendiğinde perde değeri azaldıkça, formant frekanslarının arttığı gözlenmiştir.

En yüksek perde periyodu değerine sahip konuşmacının 2. formant frekansı ile perde periyodu değeri f^{buyuk} ve p^{buyuk}, en düşük perde periyodu değerine sahip konuşmacının 2. formant frekansı ile perde periyodu değeri ise f^{kucuk} ve p^{kucuk} olmak üzere, iyileştirilecek sese ait /a/ fonemi spektrumunun 2.formant frekansı f_2 ile ifade edilirse, bu sesin iyileştirilmesi için kullanılacak referans perde değeri aşağıdaki gibi bulunabilir.

$$a = (p^{buyuk} - p^{kucuk}) / (f^{kucuk} - f^{buyuk}) \quad (2)$$

$$perde_{referans} = (f^{kucuk} - f_2) * a + p^{kucuk} \quad (3)$$

Önerilen sistemde, karşılaşılması muhtemel en büyük ve en küçük perde değerine sahip konuşmacılara ait perde ve 2.formant frekansları kullanılmıştır. Buna göre, f^{buyuk}=897, p^{buyuk}=89 ve f^{kucuk}=1788 , p^{kucuk}=20 olarak alınmıştır.

3.5. Formant Yapısı Değiştirme

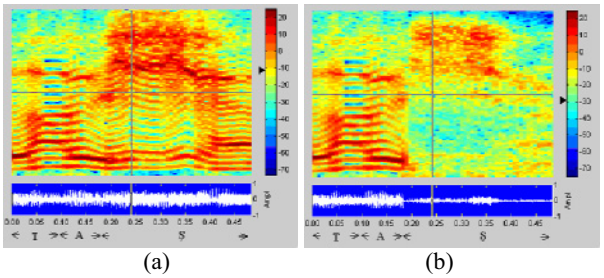
Disfonik ve normal seslendirilmiş ötümlü fonemler incelendiğinde, disfonik sese ait formant tepelerinin özellikle ilk iki tanesinin formant bant genişliklerinin daha geniş ve formant frekanslarının daha büyük olduğu gözlenmiştir. Disfonik konuşmanın anlaşılırlığını arttırmak için spektral yumuşatma [8], formant bant genişliği ve frekansı değişikliği yapılmıştır [11].

3.6. Deneysel Sonuçlar

Sentezlenen ses spektrumlarının normal sese yakınlığının nesnel olarak ölçülebilmesi için log spektral uzaklık hesabı yapılmıştır. Ancak spektral farkın hesaplanması, sentezlenen sesin normal sese yakınlığının tespiti için tek başına yeterli değildir. Bu eksikliğin giderilmesi için öznel dinleme testlerine de başvurulmuş, sentetik ses ve normal ses spektrogramları incelenmiştir.

Disfonik ses ve iyileştirilmiş ses ile normal ses arasındaki ortalama spektral uzaklıklar hesaplandığında, disfonik sese yapılan modifikasyonlar sayesinde ötümlü fonemlerde yaklaşık %25'lik spektral iyileştirme sağlandığı görülmüştür. Ötümsüz fonemlerde ise akustik özellikler korunarak sentezlenen sesin, akustik özellikler modifiye edilerek sentezlenen sese göre yaklaşık %11 oranında normal sese daha yakın olduğu gözlenmiştir.

Disfonik sesin normal sesten farklılıklarının görsel olarak algılanabilmesi için spektrogramlardan faydalanılmıştır. Şekil 4'de kısmi larinjektomili hastaya söyletilmiş /Taş/ kelimesi (Şekil 1) için tüm fonemlerin modifiye edilmesiyle elde edilmiş sentetik konuşma spektrogramı ve yalnızca ötümlü fonemlerin modifiye edilmesiyle elde edilmiş sentetik konuşma spektrogramı görülmektedir.



(a) (b)

Şekil 4: Disfonik /Taş/ kelimesi için (a) Bütün fonemlerin akustik özellikleri değiştirilerek elde edilen spektrogram (b) Ötümlü fonemlerin akustik özellikleri değiştirilerek elde edilen spektrogram

Şekil 3'ten anlaşılabileceği gibi, değiştirmeler sayesinde, sentetik sese periyodiklik kazandırılmıştır. Ayrıca sadece ötümlu fonemlerin değiştirilmesiyle elde edilen spektrogram normal sese daha yakındır.

3.6.1 Algısal Değerlendirme

Sentezlenen sesin, disfonik sese göre ne derecede tercih edilir olduğunun tespiti için öznel dinleme testlerine başvurulmuştur. 3'ü ses teli felci ve 2'si kısmi larinjektomili

5 hastanın 10'ar fonem dengeli cümlesi, bu sesler için sentezlenen düzeltilmiş sesler ve aynı cümleler için normal konuşmacılardan alınan ses örnekleri kullanılmıştır. Bu örnekler 5 dinleyiciye dinletilerek, sesleri vurgu/tonlama, tanınırlık, anlaşılırlık ve doğallık açısından 1:5 aralığında (1 çok kötü ve 5 çok iyi olmak üzere) puanlamaları istenmiştir. Bu örnekler için elde edilen ortalama puanlar aşağıdaki tabloda özetlenmiştir.

Tablo 1: Disfonik ve Düzeltilmiş Konuşmanın Karşılaştırılması

	Disfonik Konuşma	**Düzeltilmiş Konuşma**
Tonlama	0,3	3,3
Tanınabilirlik	1,5	2,6
Anlaşılırlık	2,4	2,7
Doğallık	1,1	3,2

Tablo 1'den de anlaşılabileceği gibi, özellikle tonlama ve doğallık açısından, sentetik ses, disfonik sese göre daha yüksek tercih edilebilirliktedir. Ancak yapılan formant değişiklikleri, fonem bağımlı olmadığından anlaşılırlıkta aynı başarı sağlanamamıştır.

3.6.1 Algısal Değerlendirme (2)

Disfonik sesten normal ses elde edilebilmesi için yapılan değişikliklerin, daha kaliteli bir sentetik ses üretilmesi için geliştirilmeleri gerektiği açıktır. Sistemin iyileştirilmesi için ileride yapılacak çalışmalara ışık tutması açısından, gerçekleştirilen modifikasyon yöntemlerinin başarılarının ayrı ayrı değerlendirilebilmesi gerekmektedir. Bu amaçla, normal bir konuşmacıdan alınmış ses örneklerine perde, seslilik ve formant değişiklikleri uygulanmış, sonuçlar yine orjinal ses ile karşılaştırılmıştır.

Üretilen sesin anlaşılırlığını en çok etkileyen değişiklik, formant yapısı değişikliğidir. Orjinal değerler yerine sentetik perde değerleri kullanılması, sesin tonlama ve doğallık bakımından aldığı puanlarda önemli bir düşüşe neden olmaktadır. Bölüm 3.4'de anlatıldığı gibi, sentetik perde üretilmesi için kullanılan β değişkeni dinamik erimi ifade etmektedir. β arttırıldığında üretilen sesin titreşim yeteneği de artmakta ancak bu durumda anlaşılırlık azalabilmektedir. Normal konuşmacıdan alınan ses örneğinin, sistemin ürettiği sentetik seslilik değerleri ile yeniden sentezlenmesi ile elde edilen sesin kalitesinin ise, orjinal ses örneğine çok yakın olduğu saptanmıştır.

4. Sonuçlar

Bu çalışmada, gırtlak kanseri ve ses teli felci gibi rahatsızlıklar nedeniyle, kronik ses kısıklığı problemi olan hastaların konuşmalarını, konuşmanın sürekli yapısını bozmadan normal sese yaklaştıran KUDÖK tabanlı bir sistem tasarlanmıştır. Sistemin başarısının ölçülmesi için log spektral uzaklık hesabı yapılmış ve öznel dinleme testlerine başvurulmuştur. Spektral uzaklık hesabı kullanılarak, sentetik sesin normal sese, disfonik sese göre ortalama %25 oranında daha yakın olduğu gözlenmiştir. Geliştirilen sistemde disfonik ses için üretilen perde değerleri kazanç parametresine bağlı olarak değişmektedir. Perde üretiminin sesin kazanç parametresi ile birlikte başka parametrelerinin de göz önünde bulundurularak yapılması sentetik sesin mekanikliğinin azaltılmasına ve kulağa daha doğal gelmesine yardımcı olabilir.

Bu çalışmada geliştirilen sistem, elde edilen sentetik sesin anlaşılabilirliği arttırılarak, ilerki çalışmalarda bir gömülü sistem uygulaması ile hastaların günlük hayatlarında kullanabilecekleri taşınabilir bir cihaz haline getirilebilir. Böylece geliştirilecek sistem, önerilen mekanik ve tıbbi çözümlerde olduğu gibi hastanın uzun süren eğitimler almasını gerektirmeyecek, enfeksiyon riski taşımayacak, kullanımı kolay olacak ve geçici ses kayıplarında da kullanılabilecektir.

5. Teşekkür

Bu çalışmada, ihtiyaç duyulan tıbbi bilgi ve hastalardan alınan ses verisi İstanbul Üniversitesi Cerrahpaşa Tıp Fakültesi, Kulak Burun Boğaz ve Baş Boyun Cerrahisi Anabilim Dalı'ndan temin edilmiştir.

6. Kaynakça

[1] AKIN, İ. "Total Larenjektomi Sonrası Ses Restorasyonu", *K.B.B. ve Baş Boyun Cerrahisi Dergisi, Cilt 2 Sayı; 2,* 1994.

[2] Aguilar G., Nakano-Miyatake M., "Alaryngeal Speech Enhancement Using Pattern Recognition Techniques", *IEICE - Transactions on Information and Systems,Volume E88-D ,Issue 7, pp. 1618-1622*, 2005.

[3] Bi N. and Qi Y., "Speech conversion and its application to alaryngeal speech enhancement", *Proc. ICSP96.pp.1586-1589*, 1997.

[4] Pozo A. and Young S., "Continuous Tracheoesophageal Speech Repair", *EUSIPCO*, 2006.

[5] Qi Y., Weinberg B. and Bi N., "Enhancement of female esophageal and tracheoesophageal speech", *Journal of the Acoustical Society of America, vol. 98, pp. 2461-2465*, 1995.

[6] Sawada H. , Takeuchi N., Hisada A., "A Real-time Clarification Filter of a Dysphonic Speech and Its Evaluation by Listening Experiments",*International Conference on Disability, Virtual Reality and Associated Technologies (ICDVRAT2004), pp. 239-246,* 2004.

[7] Bishnu S. Atal, Lawrence R. Rabiner. : "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition". *In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. assp-24, no. 3, june,* 1976.

[8] Morris R.W.1; Clements M.A., "Reconstruction of speech from whispers", *Medical Engineering and Physics, Volume 24, Number 7, pp. 515-520(6), September,* 2002.

[9] Thomas, I. B. , "Perceived pitch of whispered vowels," *J. Acoust. Soc. Am., vol. 46, no.2, pp. 468,* 1969.

[10] Higashikawa M., Nakai K., Sakakura A. and Takahashi H., "Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study", *Journal of Voice, pp 155-158*, 1996.

[11] McLoughlin I. V. and Chance R. J., "LSP-based speech modification for intelligibility enhancement," in *Proceedings 13th International Conference on DSP, vol. 2, pp.591-594*, 1997.