İnsan Sesinden Duygu Çıkarma
Emotion Recognition from the Human Voice

Cevahir Parlak
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
cevahirparlak@yahoo.com

Banu Diri
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
banu@ce.yildiz.edu.tr

***Özetçe*—İnsan konuşması kişiler arasındaki iletişimin en önemli aracıdır ve binlerce yıl eskiye dayanır. Bilindiği gibi insan sesi de parmak izi gibi biyometrik bir özellik olmakla birlikte ayrıca, konuşan kişinin o an ki duygusunu da taşımaktadır. Dolayısıyla canlı konuşmalardan elde edilen ses verileri duygu belirlemede metinlere göre daha gerçekçi özellikler barındırabilmektedir. Bu çalışmada, konuşmacıdan bağımsız olarak bir konuşmanın kızgın, mutlu, nötür ve üzgün mü olduğunun tespiti yapılmaktadır.**

Anahtar Kelimeler—Duygu Analizi; Duygu Madenciliği; Prozodi; Ses Perdesi.

***Abstract*—Speech is the most important communication tool between humans and is thousands of years old. As known human voice is a biometric feature like fingerprint and carries the emotional state of the speaker. Therefore, speech data acquired from live talks may have more realistic emotional features than textual data. In this work, we will try to determine whether a speaker independent speech is angry, neutral, happy or sad.**

Keywords—Emotion Analysis; Emotion Mining; Prosody; Pitch.

I. GİRİŞ

Bugüne kadar elde edilen verilerin büyük kısmı metin tabanlı olmasına rağmen, gelişen teknoloji sayesinde artık insan konuşmaları da bir veri kaynağı olarak kullanılabilmektedir [1]. Konuşma tanıma teknolojisinin yaklaşık 50 yıllık bir geçmişinin olmasına karşılık insan sesinden duygu çıkarımı çalışmaları henüz yenidir ve büyük ilgi görmektedir. Duygu çıkarımı için gerekli verileri elde etmek önemli bir sorun teşkil etmektedir. Çünkü, çok fazla duygu türü bulunmakta ve bunları yansıtan konuşmaları doğal haliyle elde etmek kolay olmamaktadır. Bazı çalışmalar ise, sadece konuşmanın olumlu veya olumsuz olup olmadığının tespitine yöneliktir.

Çağrı merkezi uygulamaları, konuşma örneklerinden duygu çıkarımının kullanıldığı en popüler alanlardır ve kızgın ve nötür ayrımı bu uygulama alanı için oldukça önemlidir [2,3]. Zira operatörlerin müşterileri ikna edebilmesi şirketler açısından çok önem taşımaktadır. Müşterilere gerekli ve yeterli ikna edici yanıtları veremeyen şirketler, müşteri kaybına kadar gidebilecek kayıplara uğrayabilirler. Duygu çıkarma sistemleri bir yandan operatörlerin müşterilere verdiği cevapların yeterince tatmin edici olup olmadığını kontrol ederken, diğer taraftan müşterilerin aldıkları cevap karşısında ikna olup olmadıklarını tespit etmek için kullanılabilir. Eğer, operatör müşteriyi ikna edemiyorsa bu operatörün çağrılıp eğitime tabi tutulması gerekebilir. Müşteri ikna olmamışsa bunun nedenleri araştırılıp çözüm bulmaya çalışılır. Çağrı merkezlerindeki on binlerce konuşmanın insan eliyle değerlendirilmesi mümkün değildir. Bu nedenle otomatik duygu çıkarımı uygulamaları kullanılarak aşırı duygu yüklü konuşmalar diğerlerinden ayrılarak elenir ve bunlar üstünde çalışılarak çözüm aranır.

Bilgisayarlı öğrenme sistemlerinde, yalan makinelerinde, sesli e-posta sistemlerinde ve ses terapisinde de duygu çıkarımı uygulamaları kullanılabilinir [3]. Son yıllarda otomobillerde de duygu çıkarımı uygulamaları kullanılmaya başlanmıştır. Buradaki amaç sürücünün duygu haliyle performansı arasındaki ilişkiyi gösteren verileri gözlemlemektir [3]. Son yıllarda duygu çıkarımı uygulamalarının oyunlarda ve insan-robot uygulamalarında kullanımı için büyük bir potansiyel gözükmektedir [3].

Makalenin ikinci bölümünde duygu çıkarımı üzerine yapılan çalışmalardan, üçüncü bölümde kullanılan veri setleri ve duygu çıkarımının temel özelliklerinden, dördüncü bölümde bu çalışmanın işleyiş şeklinden, beşinci bölümde SoundGarden programının özellikleri ve seçilen parametrelerinden, altıncı bölümde deneysel sonuçlardan, yedinci bölümde de tartışma ve sonuçlardan bahsedilmektedir.

II. İLGİLİ ÇALIŞMALAR

İnsan sesinden duygu çıkarma henüz yeni bir araştırma konusu olmakla birlikte özellikle son on yılda bu alandaki çalışmalar büyük bir ivme kazanmıştır. Pek çok duygu veri seti oluşturulmuş ve bu veri setleri üzerinde değişik duyguların tespiti için uygulamalar geliştirilmiştir [1,2,3,4,5].

Duygu çıkarımının uygulama alanlarından olan çağrı merkezlerinde genellikle kızgın ve nötür konuşmaların tespiti gerekir. Bir elektrik şirketinin çağrı merkezinin kayıtlarından alınan ve kızgın ile nötür konuşmaları ayıran bir çalışmanın [2] sonuçları Çizelge I'de verilmiştir. Bu veri setinde 2 erkek 9 kadın olmak üzere toplam 11 konuşmacı ve 190 kızgın, 201 nötür olmak üzere 391 ses kaydı bulunmaktadır. Bu kayıtların en

978-1-4673-5563-6/13/$31.00 ©2013 IEEE

uzunu 13.60 saniye, en kısası 0.52 saniye ve ortalama uzunluğu da 3.25 saniye olarak verilmiştir. Veriler 22050 Hertz ile örneklenmiş ve MPEG Layer 3 formatıyla kaydedilmiştir [2].

ÇİZELGE I. Bir elektrik şirketinin çağrı merkezindeki konuşmaların kızgın ve nötür sınıflandırılması.

		Tahmin	
		Kızgın	**Nötür**
Gerçek	**Kızgın**	% 76.3	% 23.7
	Nötür	% 16.1	% 83.9

Bazı çalışmalarda kişinin konuşmasının olumlu veya olumsuz olup olmadığının tespitine çalışılır. Bu amaçla yapılan oldukça geniş kapsamlı bir çalışma 84 katılımcı ile gerçekleştirilmiş ve gittikleri restorantlarla ilgili düşüncelerini sözlü olarak ifade etmeleri istenmiştir. Sözlü konuşmaların yanı sıra metinsel kaynaklardan da yararlanılmış ve metin tabanlı sınıflandırma ile akustik sınıflandırmadan elde edilen veriler kaynaştırılmıştır [1]. Bu çalışmada konuşmacı bağımsız uygulama yeterli başarıyı gösterememiş, konuşmacı bağımlı çalışma da ise %72.6'lık bir başarı oranı yakalanmıştır. Aynı çalışmada insanlar tarafından aynı veri seti üstünde yapılan değerlendirmede %84'lük bir başarı gözlemlenmiştir.

Berlin EmoDB [5], 5 kadın, 5 erkek aktörden alınan, içerisinde 7 farklı duyguyu (kızgın(127), neşeli(71), üzgün(62), korku(69), canı sıkkın(81), tiksinti(46), nötür(79)) barındıran 5 kısa ve 5 uzun cümleden oluşan Almanca bir veri setidir. Pan [6], Berlin EmoDB üstünde mutlu, nötür ve üzgün için yaptığı sınıflandırmada %95.1'lik başarı oranı elde etmiştir. Aynı çalışma mutlu, üzgün, nötür, sıkkın, tiksinti olmak üzere 5 duygu üzerinde yapıldığında, enerji ve ses perdesi özellikleri de kullanıldığında %66, LPCMCC özellikleri kullanıldığında %70, her iki özellik birden kullanıldığında ise %82'lik başarı oranı elde edilmiştir. Wu [7]'nun, Berlin EmoDB üstünde yaptığı çalışma ise, 7 farklı duyguyu sınıflandırmış ve %88'lik bir başarı elde etmiştir.

Bir diğer veri seti FAU Aibo [8] iki farklı okulda, yaşları 10-13 arasında değişen, 51 çocuğun (30 kız, 21 erkek) Aibo adlı bir robotla etkileşimi sırasında kaydedilen, doğal duygu yüklü Almanca dilindeki öfke, olumlu, empatik, nötür ve diğerleri olmak üzere beş duygu sınıfını içeren, toplam uzunluğu 9 saat olan bir veri setidir. Bu veri setinde 5 farklı duygu sınıfı için Kockman [8] %41.7, Bozkurt [8] %41.59'luk bir başarı almışken, olumlu-olumsuz duyguya göre yapılan sınıflandırmada ise Dumouchel [8] %69.72, Kockman [8] %68.3, Bozkurt [8] %67.9'luk sınıflandırma başarısı elde etmiştir.

III. VERİ SETİ OLUŞTURMA

Duygu çıkarımı için son yıllarda yapılan çalışmalar bu konuda bazı veri kümelerinin oluşturulmasını sağlamıştır [4,5]. Bu veri setlerinin bazıları ücretli iken, bazılarına ücretsiz erişim de mümkündür. Bu çalışmada kendi veri kümemizi oluşturmak için televizyon kanallarından elde edilen konuşma örneklerini kullandık. Televizyon kanallarından elde edilen konuşmalarda nötür konuşma kaynağının çok fazla olmasına rağmen (ana haberlerdeki konuşmalar gibi) duygu yüklü konuşma elde edebilmek pek kolay olmamaktadır. Bazen duygu yüklü konuşmaların arasına müzik veya değişik gürültüler girebilmekte bu da verinin kullanılabilir olmasına engel olmaktadır. Ayrıca, kişilerden ses verisini elde ederken kişinin ilgili duyguyu vermesi pek kolay değildir.

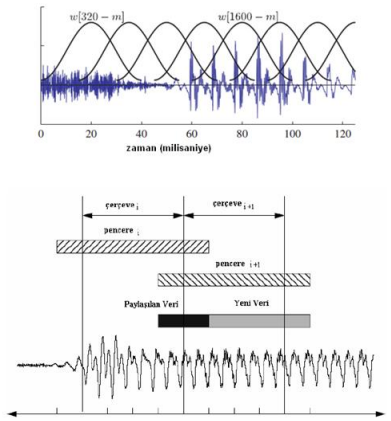
IV. AKUSTİK ÖZELLİKLERDEN DUYGU ANALİZİ

İnsan sesinden duygu çıkarımında kullanılacak özelliklerin en önemlisi f0 denilen temel frekanstır. Temel frekans, f0, konuşmanın ses perdesi özelliklerinden birisidir. Bilindiği gibi erkeklerin sesi kadınlara göre daha kalındır. Bunu sağlayan özellik temel frekanstır, erkeklerde 80-160 Hz, kadınlarda 150-250 Hz, çocuklarda ise 200-400 Hz arasında değişmektedir. Temel frekans dışında konuşma hızı, konuşma kalitesi, enerji gibi özellikler de duygunun tespitinde kullanılabilecek özellikler arasındadır. Bu özelliklerin kullanılmasıyla daha iyi sonuçlar elde edilebilir. Bu çalışmada f0'ın konuşma sinyali boyunca gösterdiği değişim ve sessiz duraklamalar göz önüne alınmış, diğer özelliklerin kullanılması ise ileriki bir çalışma olarak bırakılmıştır. Çalışmamızda konuşma sinyalinden elde edilen özellikler arasında filtre bankaları, spektral frekanslar, enerji, sıfır geçiş oranı, düşük ve yüksek frekans enerji ve temel frekans bulunmaktadır.

Bu çalışmada sadece f0, duygu tahmini için kullanılmıştır. Konuşma sinyali Şekil II'de gösterildiği gibi 32 milisaniyelik parçalara ayrılmış ve her 16 milisaniyede bir Şekil II'de gösterildiği gibi üst üste bindirmeli olarak analiz edilmiştir. Kullanılan program bu aralıkların değiştirilmesine de olanak vermektedir.

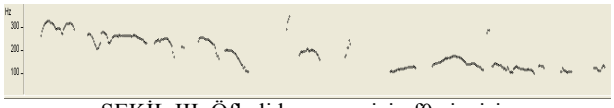
Bilindiği gibi genel olarak konuşma tanıma ile ilgili çalışmalar konuşmacı bağımlı ve konuşmacı bağımsız olarak ikiye ayrılmaktadır. Konuşmacı bağımsız modellerde istenen başarı pek yakalanamamıştır. Bizim çalışmamız konuşmacı bağımsız bir modeldir ve iyi bir modelle konuşmacı bağımsız çalışmalarda da yeterli bir başarı elde edilebilir.

Bu amaçla, elimizdeki veri kümesi incelenmiş ve kızgın konuşmalarla nötür konuşmalar arasındaki farklar belirlenmeye çalışılmıştır. Maksimum f0 değeri 400 Hz olarak kabul edilebilir.

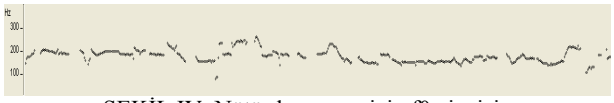


ŞEKİL II. Üst üste bindirmeli pencere uygulaması [9,10].

F0 tespiti için otokorelasyon, Yin Metodu ve Harmonik metodlar yaygın olarak kullanılmaktadır [11,12,13]. Bu çalışmada harmoniklerden yararlanılmıştır. Sesli ve sessiz harflerin ayırt edilebilmesi için sıfır geçiş oranı ve kısa zaman enerji değerleriyle beraber bazı spektral özelliklerden yararlanılmıştır. Şekil III ve Şekil IV'te kızgın ve nötür konuşmalar için f0 çizgileri gösterilmektedir.



ŞEKİL III. Öfkeli konuşma için f0 çizgisi.



ŞEKİL IV. Nötür konuşma için f0 çizgisi.

Çalışmamızın temel dayanak noktalarını özetlemek gerekirse;

1. Kızgın ve mutlu konuşmalarda standart sapma yüksektir ve f0 genel olarak yüksek bir değerde seyreder.
2. Nötür, kızgın ve mutlu konuşmalarda duraklamalar azdır.
3. Üzgün konuşmalarda duraklamalar fazladır ve f0 düşük bir değerde seyreder.

Bu özellikleri kullanarak sınıflandırma yapmaya çalışılmıştır. Sınıflandırmamızda mümkün olduğunca az ve konuşmacı bağımsız özellikleri kullanmaya özen gösterdik. Özellik sayımız az olduğundan sınıflandırıcı olarak basit bir karar ağacının kullanılması uygun görülmüştür.

IV. SOUNDGARDEN UYGULAMASI

SoundGarden programı ses işleme için tasarlanmış bir uygulama olup, bu çalışmada duygu çıkarımı için kullanılmıştır. Program herhangi bir yardımcı araç kullanmadan ses dosyaları üzerinde işlem yapabilmekte ve gerekli ses özelliklerini çıkarabilmektedir. İşlenecek ses dosyaları mono formatta MS Wave dosyalarıdır. Stereo veya diğer formatlar kullanmanın ne konuşma ne de duygu tanıma için bir avantajı yoktur. İnsan sesi için örnekleme frekansının 16000 Hz olarak seçilmesi uygundur. Programa girilecek veriler *Set Parameters* düğmesiyle varsayılan değerler olarak yüklenmiştir. (Pencere uzunluğu:512 örnek=31.25ms, Sıfır geçiş uzunluğu 512 örnek=31.25ms, Bindirme uzunluğu:256 örnek, En büyük frekans şiddetinin en küçüğe oranı:20, En düşük f0 oranı:12, En düşük frekans şiddeti (1/1000):150, Maksimum f0 değeri:400Hz, Minimum f0 değeri:75 Hz). Pencere uzunluğu ve sıfır geçiş eşit alınmalıdır. Pencere uzunluğu ve bindirme uzunluğu değiştirildiği takdirde farklı sonuçlar elde edilebilir. Çerçeve fonksiyonu olarak Dikdörtgen, Hamming veya Blackman seçilebilir. Çalışmamızda dikdörtgen çerçeve kullanılmıştır.

V. DENEYSEL SONUÇLAR

Deneysel sonuçlar için iki farklı veri seti kullanılmıştır. Birincisi, bizim oluşturduğumuz veri seti (Emo1), diğeri literatürde kullanılan Berlin EmoDB veri setidir. Emo1 143 adet, farklı kişilere ait kızgın ve nötür ses verisinden oluşmaktadır. Bunlardan 32 tanesi kızgın, 111 tanesi de nötür sınıfına ait verilerdir. Kızgın veriler üzerinde programımız çalıştırıldığında 32 verinin 26 tanesini doğru olarak tahmin etmiştir. Kızgın etiketli verilerin azlığı nedeniyle eğitim imkanı olmamasına rağmen, nötür veriler üstünde eğitim imkanı bulunmaktadır. Nötür eğitim verileri üstünde 13/16'lık bir performans elde edilmiştir. Aynı algoritma Emo1 nötür test veri setine uygulandığında da 75/95 oranında bir performans elde edilmiştir. Bu değerler Çizelge III'te gösterilmiştir.

ÇİZELGE III. Emo1 Veri seti için sonuçlar.

EMO1	Başarı Oranı
Kızgın	% 82
Nötür	% 79

Berlin EmoDB'de konuşmacı bağımsız yaptığımız testlerde kızgın etiketli veriler için %82, nötür veriler için %76 başarı elde edilmiştir (Çizelge IV). Berlin EmoDB üstünde mutlu, nötür ve üzgün sınıflandırması sonuçlarımız ise Çizelge V'de gösterilmiştir.

ÇİZELGE IV. Berlin EmoDB için kızgın, nötür sonuçları.

Berlin EmoDB	Başarı Oranı
Kızgın	% 82
Nötür	% 76

Berlin EmoDB veritabanı genellikle kısa cümlelerden oluşmaktadır ve bu veritabanında sadece f0'ın standart sapmasının FFT çözünürlüğü ile karşılaştırılması kullanılarak sonuç belirlenmeye çalışılmıştır.

ÇİZELGE V. Berlin EmoDB için mutlu, nötür, üzgün sonuçları.

Berlin EmoDB	Mutlu	Nötür	Üzgün
Mutlu	% 87.32	% 8.45	% 4.23
Nötür	% 13.92	% 78.48	% 7.59
Üzgün	% 0	% 22.58	% 77.42

Çalışmamızın konuşmacı bağımsız bir çalışma olduğu ve insanların duygu ifadesinin de çok değişken tabiatı göz önünde bulundurulduğu takdirde bu sonuçlar oldukça tatmin edicidir. İleriki çalışmaların odak noktası daha iyi bir f0 tespit algoritması geliştirmek ve enerji, konuşma hızı, konuşma kalitesi, konuşmadaki stres gibi özelliklerinde kullanılmasıyla daha iyi performans elde edilmesi olmalıdır.

VI. TARTIŞMA VE SONUÇ

Bu çalışmada bugüne kadar yapılanlardan oldukça farklı bir yol izlenmiş ve MFCC vektörlerinin kullanılmasına gerek görülmemiştir. MFCC vektörlerinin kullanımı sonuçları kullanıcı bağımlı veya en azından veri seti bağımlı hale getirebilir. Sinyalin enerjisi ile duygu tahmini arasında bir ilişki beklenmemekle birlikte özellikle düşük frekanslardaki enerjinin yüksek frekanslardaki enerjiye oranı konuşmadaki stresi belirleme açısından bazı duygularda belirleyici olabilir. Konuşma hızı (speaking rate) olarak bilinen özellik sesli kısımların toplam uzunluğa oranı olarak belirlenebilir [2]. Çalışmamızda konuşma hızıyla ilgili herhangi bir değerlendirmeden yararlanılmamıştır. Ancak kızgın konuşmaların genel olarak nötürden daha hızlı, üzgün konuşmalarında nötürden daha uzun olduğu belirgindir. Konuşmacı bağımsız bir konuşma hızı analizi ve konuşma stres analizi için biraz daha araştırmaya ihtiyaç bulunmaktadır. Bu çalışmada konuşma sinyali içindeki duraklamalar hesaplanmış ve bu duraklamaların özellikle nötür ve üzgün ayrımında çok belirleyici olduğu görülmüştür. Bu hesaplama için yeni bir yöntem kullanılmıştır. Ayrıca, konuşmanın başındaki ve sonundaki boşluklarında tespit edilmesi gereklidir. Çünkü, bu boşluklar sadece kayıt esnasında oluşan ve herhangi bir özellik taşımayan boşluklardır. Veri setleri orijinal haliyle işlenmiştir ancak sıfır geçiş oranının iyi tespit edilebilmesi açısından DC değerin elenmesi gereklidir. Ayrıca minimum f0 değerinin altındaki frekanslar da bir yüksek geçiren filtre ile silinebilir. Cümlelerin içindeki sessiz kısımların tespitinde ses hazırlık zamanı olarak bilinen (Voice Onset Time) ve c, ç, b, p, t, d, k ve g gibi harflerde bulunan özelliğin sessiz kısımlardan ayırt edilebilmesi oldukça önemlidir. Genel olarak f0 algılama algoritmaları hataya meyillidir. Özellikle n, m, l, r, z, y, v, j gibi bazı sessiz harflerin de seslilik özelliği gösterebilmesi bu algoritmaların yanılmasına sebep olabilmektedir. Sesli, sessiz harf ayırımı ve ses hazırlık zamanı tespiti konuşma tanımadaki ciddi problemlerdendir. Sesli sessiz ayırımı için sıfır geçiş oranı, kısa zaman enerji ile birlikte bazı spektral özellikler kullanılmıştır. Bazı çalışmalardaki özellikle %90'ın üstündeki başarı oranının ciddi bir eleştiriye ihtiyacı var gibi gözükmektedir. Bunun yanı sıra duygu veri setlerinin gerçekçiliklerinin ve doğruluklarının da incelenmesinde fayda vardır. Sadece 2 kişiyle yapılan bir duygu veri setinin 7 milyar insanı modelleyemeyeceği aşikardır. Sonuç olarak şunun bilinmesi gerekir ki insan sesi oldukça değişken bir olgudur ve ses teknolojileri şu anda dünya üstündeki en karmaşık teknolojidir. Bu nedenle çok daha geniş kapsamlı araştırmalara ve gelişmiş metotlara ihtiyaç vardır ve bu araştırmalarda da %75 seviyelerindeki konuşmacı bağımsız bir başarı oranı oldukça iyi bir hedef olarak belirlenmelidir.

KAYNAKÇA

[1] Mairesse,F., Polifroni, J., Di Fabbrizio, G., "Can Prosody Inform Sentiment Analysis? Experiments On Short Spoken Reviews", Nokia Research.

[2] Morrison, D., Wang, R., Silva, L.C., Xu, W.L., "Real-time Spoken Affect Classification and its Application in Call-Centres", *Information Technology and Applications*, 2005.

[3] Ramakrishnan, S., "Recognition of Emotion from Speech: A Review", *International Journal of Speech Technology*, v:15, Issue 2,pp 99-117, 2012.

[4] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A., "Acoustic Emotion Recognition: A Benchmark Comparison of Performances", *ASRU*, 2009.

[5] http://pascal.kgw.tu-berlin.de/emodb/

[6] Pan, Y., Shen, P., Shen, L., "Speech Emotion Recognition Using Support Vector Machine", *International Journal of Smart Home*, v:6, no. 2, April, 2012.

[7] Wu, S., Falk, T.H., ve Chan, W.P., "Automatic speech emotion recognition using modulation spectral features", *Speech Communication*,53(5), 768-785, 2011.

[8] Bozkurt, E., Erzin, E., Erdem, Ç., Erdem, E., "Interspeech 2009 Duygu Tanıma Yarışması Değerlendirmesi", *SIU*, 2010.

[9] Rabiner, L. R. ve Schafer, R. W., "Introduction to Digital Speech Processing", Foundations and Trends in Signal Processing, v:1, 1–194, 2007.

[10] Picone, J., "Fundametals Of Speech Recognition: A Short Course", *Institute for Signal and Information Processing*, 1996.

[11] Schuller, B., Batliner, A., Steidl, S. ve Seppi, D., "Recognising Realistic Emotions And Affect in Speech: State Of The Art And Lessons Learnt From The First Challenge," *Speech Communication*, v:53, no. 9-10, s. 1062–1087, 2011.

[12] Tavares, T. F., Arnal Barbedo, J.G., Lopes, A., "Performance Evaluation of Fundamental Frequency Estimation Algorithms", Proceeding of The International Workshop on Telecommunications, IWT, 2009.

[13] Seo, N., "ENEE632 Project4 Part I: Pitch Detection", 2008.