©2016 Published in 4th International Symposium on Innovative Technologies in Engineering and Science 3-5 November 2016 (ISITES2016 Alanya/Antalya - Turkey)

Duyguların Sese Yansıtılmasında Konuşmacı Etkilerinin Akustik Parametreler Üzerinden İncelenmesi

*[1]Turgut Özseven ve [2]Muharrem Düğenci
[1]Turhal Meslek Yüksekokulu, Gaziosmanpaşa Üniversitesi, Türkiye
*[2]Mühendislik Fakültesi, Karabük Üniversitesi, Türkiye

Özet

Bu çalışmanın amacı konuşmadan duygu tanımada konuşmacı bağımlılığının duyguyu yansıtmadaki etkilerinin akustik parametreler üzerinden incelenmesidir. Bu amaçla, 7 farklı duygunun aktörler tarafından ifade edildiği veriler üzerinde akustik analiz gerçekleştirilmiştir. Akustik analiz ile elde edilen akustik parametreler konuşmacı bağımlı ve konuşmacı bağımsız sınıflandırmaya tabi tutularak duyguları tanıma başarısı elde edilmiştir. Elde edilen sonuçlara göre konuşmacı bağımlı sınıflandırma başarısı konuşmacı bağımsız sınıflandırma başarısına göre bayanlarda yüksek ve erkeklerde düşüktür. Bayanlar duygularını erkeklere göre seslerine daha fazla yansıtmaktadır.

Anahtar Kelimler: Duygu tanıma, akustik analiz, duygu ve konuşmacı

Abstract

The purpose of this study is to evaluate the effects of reflection of emotion on acoustic parameters of speaker dependent in emotion recognition. For this purpose, acoustic analysis was performed on data from the different emotions expressed by the actors 7. The emotions are classified as speaker dependent and speaker independent using acoustic parameters obtained by acoustic analysis. The according to the results obtained, speaker dependent classification performance is lower in males and high in females compared to the based on speaker independent classification performance. The speech reflection of emotions is more in females than males.

Key words: Emotion recognition, acoustic analysis, emotion and speaker

1. Giriş

Ses, duygunun bir parçasıdır ve duygusal durumun aktarılmasında önemli rol oynamaktadır. Duygularda ortaya çıkan değişiklikler, solunum kalıplarını ve kas gerilimini etkileyerek, ses kalitesinde değişiklikler meydana getirmektedir. İnsan seslerinden psikolojik tanı, duygu durumu tespiti ve kişi tanımada nesnel ve öznel yöntemler kullanmaktadır. Nesnel değerlendirme için akustik analiz tercih edilmektedir.

Akustik analiz ses kayıtlarının sayısal sinyal işleme yöntemleri ile işlenmesi sonucu sese ait çeşitli parametrelerin elde edilmesidir. Çalışmalarda çoğunlukla temel frekans, formant frekansı, jitter, shimmer, sinyal gürültü oranı ve enerji parametreleri kullanılmaktadır. Hangi duyguların hangi akustik parametreler üzerinde etkili olduğunun tespiti yaş ve cinsiyet dahil olmak üzere sesteki bireysel farklılıklardan dolayı zorluk oluşturmaktadır [1]. Duygu tanımadaki bir diğer zorluk ise duyguların konuşmaya tam olarak yansıtıldığı veri elde edilmesidir. Bu nedenle

*Corresponding author: Address: Turhal Meslek Yüksekokulu, Gaziosmanpaşa Üniversitesi, Türkiye, Tokat. E-mail address: turgutozseven@gmail.com

çalışmalarda geçerliliği kabul görmüş EMO-DB [2] ve SUSAS [3] gibi hazır veritabanları kullanılabildiği gibi araştırmacı tarafından toplanan veriler de kullanılabilmektedir. Akustik analiz için Praat [4] ve OpenSMILE [5] gibi hazır araçlar kullanılabildiği gibi araştırmacılar tarafından geliştirilen kodlar da kullanılmaktadır.

Akustik parametreler üzerinden duyguların tespit edilmesi için Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Hidden Markov Model (HMM), Multilayer Perceptron (MLP) ve k-Nearest Neighbor (k-NN) gibi tekil sınıflandırıcılar veya hybrid sınıflandırıcılar kullanılabilir.

Bu çalışmada, duyguların sese yansıtılmasında konuşmacı etkilerini araştırılmıştır. Bu amaçla SVM sınıflandırıcı ile konuşmacı bağımlı ve konuşmacı bağımsız sınıflandırma performansları karşılaştırılmıştır.

Yapılan çalışmalar incelendiğinde elde edilen sonuçlar Tablo 1'de verilmiştir.

Tablo 1. İlgili Çalışmalar

Kullanılan Parametreler	**Parametre Sayısı**	**Classifier**	**Başarı**	**Referans**
F0, Eng, D, Frm, HNR,ZCR, MFCC, FFT	75	SVM	%87.5	[6]
F0, I, SR, D, MFCC	-	SVM	%65.5	[7]
I, SR, F0, D, Jt, Sh, prosody, VQ, LFPC	121	SVM-RBF	%78.3	[8]
MFCC	39	SVM	%67.0	[9]
Eng, F0, MFCC, ZCR, Sp, Sp Eng	95	SVM (SI-CV, 3DEC hier.)	%79.5	[10]
		SVM (SI-CV, DAG)	%77.9	
ZCR, Eng, F0, HNR, MFCC	32	SVM (open set)	%66.2	[11]
		SVM (closed set)	%80.6	
ZCR, Eng, F0, MFCC, Δreg	6552	SVM	%85.2	[12]
ZCR, Eng, F0, MFCC	4368		%86.1	
dd+fea+fun+PCA	37		%80.2	
F0, I, D, Frm, Eng, HNR, Jt, Sh, MFCC	90	SVM	%80.4	[13]
MFCC, LFPC, NPQ, QO, GH, SGP, PSP, QOQ, F0	106	SVM-RBF	%89.0	[14]
I, L, MFCC, ZCR, LSP, F0	-	SVM	%83.1	[15]
F0, I, D, Frm, Eng, HNR, Jt, Sh, MFCC	204	SVM	%80.4	[16]
Pr + Sp + NLD	2641	SVM (Male)	%85.9	[17]
Pr	241		%70.5	
Sp	960		%73.9	
NLD	1440		%77.8	
Pr+Sp	1201		%78.6	

F0: Pitch, Eng: Energy, D: Duration, Frm: Formant Frequency, ZCR: Zero-Crossing Rate, MFCC: Mel-Freqeuncy Cepstrum, FFT: Fast Fourier Transform, Nf: New feature set, Wv: Wavelet, SR: Speech Rate, I: Intensity, LPC: Linear Predictive Coding, HNR: Harmonic to Noise Ratio, Sp: Spectral, Pr: Prosodic, L: Loudness, Jt: Jitter, Sh: Shimmer, Ent: Entropy, vuv: voiced/unvoiced, #: Özellik Seçimi Öncesi, ##: Özellik Seçimi Sonrası

Tablo 1'de verilen literatür taramasına göre SVM sınıflandırıcı ile duygu tanımada %90'lara ulaşan başarı elde edilmiştir. Konuşmacı bağımsız duygu tanımada başarı oranı konuşmacı

bağımlı duygu tanımaya göre daha yüksek başarı oranına sahiptir. Ancak bu konuşmacı bağımlı çalışmalarda konuşmacının cinsiyetine göre farklılık göstermektedir.

Bu çalışmanın ikinci bölümünde materyal ve yöntem, üçüncü bölümünde deneysel sonuçlar verilmiştir.

2. Yöntem ve Materyaller

Bu çalışmada, Berlin Database of Emotional Speech (EMO-DB) veri tabanından temin edilen 535 adet konuşma kaydı kullanılmıştır [2]. EMO-DB farklı duyguların aktörler tarafından ifade edilmesi ile elde edilmiştir. Ses kayıtları 16 kHz örnekleme frekansına sahip olup 16 bit monodur. Kullanılan veriye ait çeşitli bilgiler Tablo 1'de verilmiştir.

Tablo 1. Kullanılan Verinin Dağılımı

Demografik Özellik		**Adet**
Duygu Durumu	Anger	50
	Happiness	25
	Boredom	27
	Disgust	21
	Anxiety/Fear	29
	Sadness	49
Farklı Konuşmacı Sayısı		10
Cinsiyet	Male	92
	Female	109

Konuşma sinyali durağan olmayan bir sinyaldir ama kısa zaman aralıklarında durağan olduğu kabul edilir. Kısa zaman aralıklarını elde etmek için de sinyal çerçeveler bölünmektedir. Çalışmamızda sinyal işleme 20ms çerçeve boyutu, hamming pencereleme ve %50.0 örtüşme ön işleme süreçleri ile gerçekleştirilmiştir. Her çerçeve üzerinden akustik parametreler çıkartılmıştır. Çalışmada kullanılacak akustik parametrelerin elde edilmesi için Praat [4] kullanılmıştır. Çalışmada kullanılan akustik parametreler Tablo 2'de verilmiştir.

Tablo 2. Çalışmada Kullanılan Akustik Parametreler

Akustik Parametre	**Mean**	**Std.Dev.**
F0	✓	✓
F1	✓	
F2	✓	
F3	✓	
Jitter (Local, Rap, ppq5, ddp)	✓	
Shimmer (Local, apq3, apq5, apq11, dda)	✓	
HNR	✓	
Unvoiced_frame	✓	
Voiced_Break	✓	
Intensity	✓	✓
ZCR	✓	

Pitch (F0) gırtlaksı uyarılmanın temel frekansıdır ve ses kıvrımları ve alt gırtlak hava basıncına bağlıdır. Sesin kalınlık ve inceliğini bildirir. Periyot, F0 algısında birincil faktördür. Döngü/saniye olarak ölçülür ve Hertz (Hz) ile ifade edilir. Bu değer ergenlik öncesi kız ve erkeklerde 220-240 Hz civarında iken erişkin erkekler ve kadınlarda sırası ile ortalama 100-150 Hz ve 150-250 Hz arasındadır [18]. Formant (F1, F2, F3) ses yolundaki rezonanstır ve ses yolunun nicel özellikleri ile ilgili spektral bilgi sağlamaktadır. Jitter, periyotlar arası değişikliği gösteren parametredir. F0'daki istem dışı ortaya çıkan düzensizlikleri içerir [19]. Bu parametre ardışık titreşimli döngüleri arasındaki temel frekansın değişimleri olarak tanımlanır [20]. Genlik pikleri arasındaki periyodik varyasyona ise shimmer adı verilir. Bu parametre, ardışık titreşimli döngüler arasındaki gırtlak akış genlik değişiklikleri olarak tanımlanır [21]. HNR, F0 ve onun katları olan harmoniklerin toplam enerjisinin gürültü enerjisine oranıdır. Unvoiced frame, konuşma içerisinde sessiz kalınan bölgelerin oranıdır. Voiced break, konuşmadaki duraksamaların sayısıdır. Intensity, konuşmanın enerjisidir. ZCR, sinyaldeki işaret değişimlerinin oranını gösterir. Ses sinyalinin sıfırdan geçiş sayısı olarak bilinir. Sinüzoidal bir sinyalde her periyotta iki sıfırdan geçiş olduğu için sinyal sıklığı sıfırdan geçiş sayısının yarısı olarak hesaplanır.

SVM istatistiksel öğrenme teorisine dayalı bir yöntemdir. Temel amaç sınıfları birbirinden en iyi şekilde ayıran karar fonksiyonun başka bir ifadeyle hiper-düzlemin tanımlanması esasına dayanır. Literatürde SER sistemlerde en çok kullanılan sınıflandırıcılardan birisidir ve EMO-DB üzerinde (7 duygu) yapılan çalışmalarda en yüksek başarı 93.78% olarak elde edilmiştir [22]. SVM sınıflandırıcı kullanılan başka bir çalışmada 535 veri seti ve 37 özellik kümesi ile 80.2% başarı elde edilmiştir [12].

3. Sonuç ve Öneriler

Akustik parametreler Praat yazılımında elde edilerek Weka üzerinde SVM sınıflandırıcı kullanılmıştır. Sınıflandırma başarıları elde edilirken üç farklı özellik kümesi kullanılmıştır. Bunlardan birincisi 10 aktör tarafından seslendirilen 535 ses doyasına ait akustik parametrelerdir. Bu veri kümesi konuşmacı bağımsız sınıflandırma için kullanılmıştır. İkincisi, bayan bir aktör tarafından 7 duygunun ifade edildiği 72 ses kaydına ait akustik parametrelerdir. Üçüncüsü ise, erkek bir aktör tarafından 7 duygunun ifade edildiği 60 ses kaydına ait akustik parametrelerdir. İkinci ve üçüncü özellik kümeleri ise konuşmacı ve cinsiyet bağımlı sınıflandırma için kullanılmıştır. Tablo 3'de konuşmacı bağımsız ve konuşmacı bağımlı sınıflandırma sonuçları verilmiştir.

Tablo 3. Sınıflandırma Sonuçları

Duygu	Konuşmacı Bağımsız	Konuşmacı Bağımlı	
	Başarı	Erkek	Bayan
Happiness	%43.7	%50.0 ⇧	%72.7 ⇧
Neutral	%70.9	%54.5 ⇩	%60.0 ⇩
Anger	%78.7	%61.5 ⇩	%71.4 ⇩
Sadness	%87.1	%50.0 ⇩	%100 ⇧
Fear	%60.9	%75.0 ⇧	%57.1 ⇩
Boredom	%75.3	%88.9 ⇧	%92.9 ⇧
Disgust	%32.6	%60.0 ⇧	%90.9 ⇧
Ortalama Başarı	%67.1	%64.3 ⇩	%80.3 ⇧

Tablo 3 incelendiğinde bayanlarda konuşmacı bağımlı sınıflandırma başarısı erkeklere ve konuşmacı bağımsız sınıflandırma başarısına göre daha yüksektir. Erkeklerde konuşmacı bağımlı sınıflandırma başarısı bayanlara ve konuşmacı bağımsız sınıflandırma başarısına göre daha düşüktür. Elde edilen bu sonuç bayanların duygularını erkeklere göre seslerine daha fazla yansıttığının bir göstergesidir. Erkekler can sıkıntılarını seslerine daha fazla yansıtırken bayanlar üzüntülerini daha fazla yansıtmaktadır. Erkekler seslerine en az düzeyde üzüntü ve mutluluklarını, bayanlar en az düzeyde korkularını yansıtmaktadır.

References

[1] Zupan B, Neumann D, Babbage DR, Willer B. The importance of vocal affect to bimodal processing of emotion: Implications for individuals with traumatic brain injury. J. Commun. Disord. 2009; 42(1): 1–17.
[2] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. Interspeech 2005; 5: 1517–1520.
[3] Hansen JH, Bou-Ghazale SE, Sarikaya R, Pellom B. Getting started with SUSAS: A speech under simulated and actual stress database. Eurospeech 1997; 97: 1743–46.
[4] Boersma P, Weenink D. Praat: Doing phonetics by computer [Computer program] 2010.
[5] Eyben F, Wöllmer M, Schuller B. Opensmile: The munich versatile and fast open-source audio feature extractor. Proceedings of the international conference on Multimedia 2010; 1459–1462.
[6] Schuller B, Müller R, Lang MK, Rigoll G. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. Interspeech 2005; 805–808.
[7] Shami M, Verhelst W. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Commun. 2007; 49(3): 201–212.
[8] Luengo I, Navas E, Hernaez I. Feature analysis and evaluation for automatic emotion identification in speech. IEEE Trans. Multimed. 2010; 12(6): 490–501.
[9] Vasuki P, Aravindan C. Improving emotion recognition from speech using sensor fusion techniques. TENCON 2012-2012 IEEE Region 10 Conference 2012; 1–6.
[10] Hassan A, Damper RI. Classification of emotional speech using 3DEC hierarchical classifier .Speech Commun. 2012; 54(7): 903–916.
[11] Garg V, Kumar H, Sinha R. Speech based emotion recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers. Communications (NCC), 2013 National Conference on 2013; 1–5.
[12] Chiou BC, Chen CP. Feature space dimension reduction in speech emotion recognition using support vector machine. Signal and Information Processing Association Annual Summit and Conference (APSIPA) 2013; 1–6.
[13] Zhao X, Zhang S, Lei B. Robust emotion recognition in noisy speech via sparse representation. Neural Comput. Appl. 2014; 24(7): 1539–1553.
[14] Kachele M, Zharkov D, Meudt S, Schwenker F. Prosodic, spectral and voice quality feature selection using a long-term stopping criterion for audio-based emotion recognition. 2014; 803–808.
[15] Jin Y, Song P, Zheng W, Zhao L. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition. Acoustics, Speech and Signal

Processing (ICASSP), 2014 IEEE International Conference on 2014; 4808–4812.
[16] Zhao X, Zhang S. Spoken emotion recognition via locality-constrained kernel sparse representation. Neural Comput. Appl. 2015; 26(3): 735–744.
[17] Shahzadi A, Ahmadyfard A, Harimi A, Yaghmaie K. Speech emotion recognition using nonlinear dynamics features. Turk. J. Electr. Eng. Comput. Sci. 2015; 23: 2056–2073.
[18] Sarıca S. Ses analizinde kullanılan akustik parametreler. Tıpta Uzmanlık Tezi, Kahramanmaraş Sütçü İmam Üniversitesi Tıp Fakültesi, Kahramanmaraş, 2012.
[19] Okur KM. CSL ve Dr. Speech ile ölçülen temel frekans ve pertürbasyon değerlerinin karşılaştırılması. KBB Ihtis. Derg. 2001; 8: 152–157.
[20] Farrus M, Hernando J. Using jitter and shimmer in speaker verification. Signal Process. IET 2009; 3(4): 247–257.
[21] Deshmukh O, Espy-Wilson CY, Salomon A, Singh J. Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech. Speech Audio Process. IEEE Trans. On 2005; 13(5): 776–786.
[22] Ivanov A, Riccardi G. Kolmogorov-Smirnov test for feature selection in emotion recognition from speech. Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on 2012; 5125–5128.