Performance Analysis of MFCC Features On Emotion Recognition from Speech

Mesut Durukal (Corresponding author)
Department of Electronics Engineering, Gebze Technical University
41400, Kocaeli, Turkey
E-mail: mesut.durukal@tubitak.gov.tr

A. Koksal Hocaoglu
Department of Electronics Engineering, Gebze Technical University
41400, Kocaeli, Turkey

Abstract

This work presents MFCC-based emotion recognition from speech. For this purpose, features of labeled speech signals are extracted and the classifier is trained. Then, test data is classified using its features and the classification performance is measured. In this work, MFCC (Mel-frequency Cepstrum Coefficients) features are extracted for training and recognition. In addition to investigation of success rates for different emotion classes, comparison of results to other results obtained with additional features are also analyzed.

Keywords: emotion recognition from speech, performance analysis on emotion recognition, emotion classification.

MFCC Özniteliklerinin Konuşmadan Duygu Tanıma Üzerindeki Performans Analizi

Özet

Bu çalışma, konuşma işaretinin incelenerek kişinin duygu durumunun tanımlanması konusunu ele almaktadır. Bu amaç için öncelikle etiketlenmiş olan konuşma verilerinin öznitelikleri çıkarılarak sınıflandırıcı eğitimi yapılmakta; sonrasında ise test verileri kullanılarak sınıflandırma performansı ölçülmektedir. Öznitelik olarak MFCC (Mel-frequency Cepstrum Coefficients) katsayıları alınmış ve farklı duygu sınıfları için tanıma başarı sonuçları irdelenip diğer özniteliklerle yapılan çalışmalarla kıyaslanması yapılmıştır.

Anahtar kelimeler: konuşmadan duygu tanıma, duygu tanıma performans analizi, duygu sınıflandırma.

1. Giriş

1.1. Genel Bakış

İş gücü ve zamanın çok önemli duruma geldiği dünyada, işler olabildiğince otomatik olarak yapılmak istenmektedir. Bu yolla harcanan emek asgari düzeye indirgenmekte ve işler hızlı ve verimli bir şekilde yapılmaktadır. Aynı durum, iletişim için de geçerli olduğu için konuşma tanıma ve işleme üzerinde birçok çalışma yapılmıştır.

Çağrı merkezleri gibi konuşma tabanlı çalışılan yerlerde ses ve konuşmacı durumuna göre otomatik tanımlamalar yapılarak verim artırılması hedeflenmektedir. Son yıllarda bu alanda yapılan çalışmaların odak noktalarından bir tanesi de konuşmacının duygu durumudur. Konuşmacının duygu durumunun (mutlu, sinirli, üzgün, sıkılgan, vb.) tespit edilmesi, uygulama alanına göre çok önemli bir girdi olarak kullanılabilmektedir.

1.2. Kaynak Tarama

Konuyla ilgili Rao et al. [1] MFCC özniteliklerini kullanarak Gaussian Mixture Model sınıflandırıcısı ile test verilerini dört ayrı duygu sınıfına ayrıştırmış ve farklı setler için %68'den başlayarak %85'e varan başarı oranı elde etmişlerdir. Yine MFCC öznitelik tabanlı yapılan bir diğer çalışmada [2] ise beş duygu sınıfı için yapay sinir ağı sınıflandırıcısı ile % 79 oranında sınıflandırma başarısına ulaşılmıştır.
MFCC haricinde başka özniteliklerle çalışan sistemler de geliştirilmiştir. LPCC (Linear Prediction Cepstral Coefficients) öznitelikleriyle yapay sinir ağı üzerinden modellenen yapı [3] örneklerden bir tanesidir.
Ses işaretleri için en sık kullanılan özniteliklerden biri de tonlama ve stres (prosodic features) özellikleridir. Bu öznitelikleri de işin içine katarak yapılan KNN sınıflandırmalarında Rong ve ekibi [4] 5 sınıf için, Zhang ve ekibi [5] ise 4 sınıf için ise % 72 başarı oranı yakalamışlardır.
HMM (Hidden Markov Model) [6], KNN (K-Nearest Neighbor) [7] ve SVM (Support Vector Machine) [8],[9] gibi algoritmalar kullanan bazı diğer sistemlerde ise %56'dan %87'ye kadar tanıma oranları elde edilmiştir. Bu çalışmalarda LFPC, LPCC, MFCC, prozodi ve kalite öznitelikleri kullanılmıştır.
Yapılan çalışmalarda karşılaşılan zorlukları gidermek için farklı yöntemler geliştirilmiştir. Aşılması gereken en yaygın problemlerden biri, farklı duygu sınıflarının benzer özniteliklere sahip olmasının sebep olduğu performans düşüklüğüdür. Mutlu sınıfına ait seslerle sinirli etiketli seslerin ortak olarak yüksek enerjiye sahip olması gibi durumlar başarı oranını düşüren etkenler arasında gösterilebilir.
Önerilen çözümlerden biri sınıfları gruplayıp toplam kategori sayısını düşürmektir. [10] Enerji odaklı düşünerek mutlu ve sinirli duygularını yüksek, üzgün ve sıkılgan duyguları ise düşük sınıfında toplamak örnek bir uygulamadır.

1.3. Kapsam ve Avantajlar

Bu çalışmada; öznitelik olarak MFCC, sınıflandırıcı olarak kNN ve SVM, veritabanı olarak ise diğer çalışmalarla kıyaslamayı kolaylaştırmak açısından yaygın olarak kullanılan Berlin Duygu Tabanlı Konuşma Veritabanı [11] kullanılmıştır.
2. Bölümde uygulanan yöntemler açıklanmakta, 3. Bölümde veri seti tanımlanmakta ve test sonuçları verilmekte, 4. Bölümde ise sonuçlarla ilgili çıkarımlar yapılmakta ve çalışmanın özeti sunulmaktadır.
Bu çalışmanın sağladığı temel avantajlar benzer özniteliklere sahip duygu sınıfları arasında gruplamalar yapılarak sınıflandırıcının geliştirilmesi, farklı duygu sınıfları için tanıma başarı sonuçlarının irdelenip diğer özniteliklerle yapılan çalışmalarla kıyaslanmasının ortaya konması ve bu sayede MFCC özniteliklerinin konuşmadan duygu içeriğinin tespitine olan katkısının göz önüne çıkarılmasıdır.

2. Yöntem

Konuşmadan duygu tanıma için öncelikle ses işaretleri üzerinde ön işleme yapılarak iyileştirildikten sonra öznitelik çıkarımı yapılarak model eğitimi yapılır. Daha sonra etiketlenmesi istenen veri eğitilmiş model kapsamında sınıflandırılır.

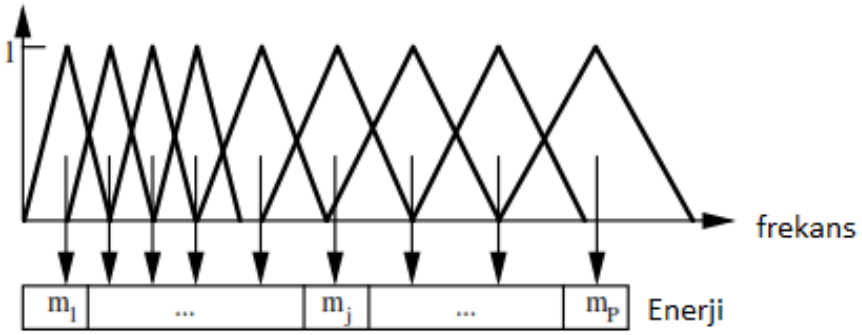
2.1 Veri ve Önişleme

Bu çalışma kapsamında Berlin Duygu Tabanlı Konuşma Veri tabanı [11] kullanılmıştır. Bu sette 10 farklı metinden oluşan ve 10 aktör (5 bay, 5 bayan) tarafından okunan 535 konuşma bulunmaktadır. Konuşmalar; mutlu, sinirli, üzgün, korkmuş, sıkılmış, iğrenmiş ve normal duygularla okunmuş durumdadır.
Ses işaretlerinden öznitelik çıkarımından önce sessizlikten arındırma işlemi uygulanır. Bu işlem için her bir örnekleme penceresindeki işaretin seviyesi kontrol edilip eşik değerinin altında kalan bölümler belirlenir ve bunlar sınıflandırma maksadıyla kullanılmaz. İşaretin ortalama değerden farkının standart sapma değerine bölümünün 0,3'ten düşük olduğu pencereler atılarak sessiz bölümler arındırılmış olur.

2.2 Öznitelik Çıkarımı

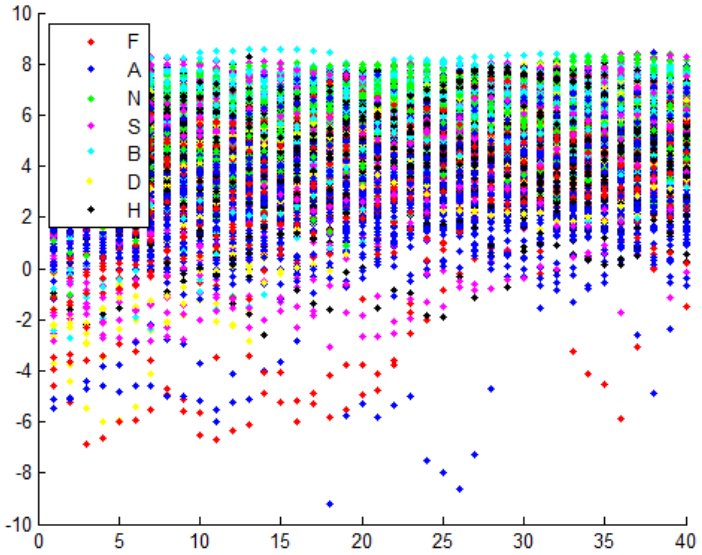
Cümle, kelime gibi uzun ses bölümlerinin genel özelliklerinden çıkarılan öznitelikler prozodi nitelikleridir. Temel frekans (pitch), enerji, ritim, bant genişliği gibi özellikler bazı prozodi niteliklerindendir [1]. Araştırmalar, duygu tanıma için en önemli özellikler arasında sesli kısımların oranı ve maksimum yoğunluk olduğunu göstermiştir [5].
Ses işaretleri için yaygın olarak kullanılan sistem (spectrum) özniteliklerinin yanında kaynak işaretinden çıkarılan kaynak (source) öznitelikleri de tanıma sistemlerine girdi sağlayan bir diğer parametre setidir. Ses rengi, boğumlanma, armonik/gürültü oranı, stres gibi bazı parametreler ise sesin kalite özniteliklerini oluşturur [5].

MFCC öznitelikleri, temel ses özelliklerini sentez vektörlerine çevirir ve spektral özellikler içinden en sık kullanılanıdır [4]. Bu katsayı seti, işaretin logaritmik güç spektrumunun Fourier dönüşümü hesaplanarak elde edilir.



Şekil 1. Mel-scale süzgeç bankası. [12]

MFCC katsayılarının frekans dönüşümü hesaplanırken dikkat edilen noktalardan biri pencerelerin genişliklerinin hassasiyet önemine göre seçilmiş olmasıdır. En düşük frekans aralıklarında dar pencere genişliği seçilirken yüksek frekanslar için pencereler genişletilmektedir. Bu çalışmada 12 adet MFCC katsayı seti çıkarılarak ilgili işlemler yapılmıştır.
Şekil 2'de model eğitiminde kullanılmak üzere çıkarılan MFCC katsayı setinin ilk vektörlerinin dağılımı gösterilmiştir. 'F' korku, 'A' sinirli, 'N' normal, 'S' üzgün, ' B' sıkılmış, 'D' iğrenme, 'H' ise mutlu duygularını temsil etmektedir. Aynı simgeler sonuç tabloları için de geçerlidir.
Özniteliklerin dağılımı incelendiğinde genelde aynı duygu sınıflarına ait değerlerin benzer aralıklarda yer aldığı görülebilir. Sınıflar birbirinden ne kadar uzakta yer alırsa ayrıştırmanın da o kadar kolay olacağı söylenebilir.



Şekil 2. Özniteliklerin görünümü.

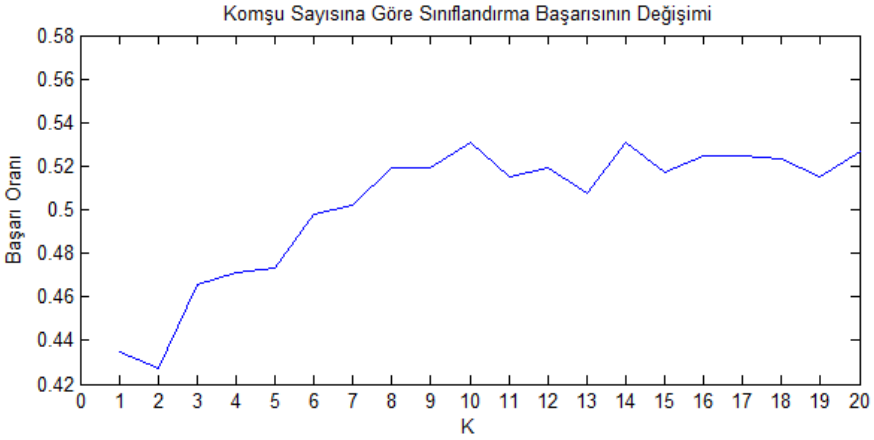
2.3 Sınıflandırma
Sınıflandırıcı eğitimi için kullanılan öznitelikler, belli algoritmalarla seçilerek hem daha etkin kullanılabilecek olanlar girdi olarak sağlanmış olur hem de veri üzerinde boyut azalımı yapılmış olur. Bu amaç için kullanılan yöntemlerden biri Chi-square, kazanç oranı, bilgi kazanımı, tutarlılık gibi süzgeçlerden birini kullanarak yapılan filtreleme işlemidir [5]. PCA (Principal Component Analysis), MDS (Multi-Dimensional Scaling), ISOMap gibi algoritmalar ise en yaygın kullanılan boyut azaltıcı

yöntemlerdendir [4]. Bu çalışma kapsamında sadece MFCC öznitelikleriyle çalışıldığı için bu set üzerinde bir boyut azaltma uygulaması gerçekleştirilmemiştir.
Çıkarılan öznitelikler her bir test verisini eğitim seti içerisindeki en yakın komşularının çoğunluk oylamasına göre etiketleyen KNN [13] ve eğitim seti üzerinde yapılan optimizasyon çalışmalarına göre çıkarılan destek vektörleriyle karar oluşturan SVM [14] sınıflandırıcıları ile etiketlenerek her biri için elde edilen sonuçlar irdelenmiştir.

3. Sonuçlar

3.1 Optimizasyon İşlemleri

Uygulanan sessizlikten arındırma işleminin etkilerini görmek için diğer parametreler sabit tutularak hem sessiz bölümleri atılmış ses işaretleriyle hem de orijinal işaretlerle sonuçlar çıkarılmıştır. Orijinal işaretlerle %43 başarı oranı yakalanırken sessiz pencerelerin atıldığı işaretlerle bu oran %53'e çıkmıştır. Bu nedenle, çalışmanın devamında konuşma işaretlerine sessiz pencerelerin dikkate alınmasını engelleyen ön işleme adımı uygulanarak sonuçlar elde edilmiş ve raporlanmıştır.
KNN algoritmasında test verisinin en yakınındaki eğitim seti örneklerinin etiketlerine bakıldığı için; kontrol edilen en yakın komşu sayısı, performans üzerinde doğrudan etkilidir. Şekil-3'te seçilen komşu sayısının performans üzerindeki etkisi verilmektedir. En yakın komşu sayısının 1'den 10'a kadar arttıkça performansı artırdığı, 10'dan sonra ise etkisinin nispeten istikrarlı hale geldiği Şekil-3'te gözlemlenebilir. Bu nedenle, çalışmanın devamında k sayısı 10 olarak seçilmiştir.



Şekil 3. Kontrol edilen komşu sayısının etkisi.

3.2 Doğruluk Tabloları

Optimize edilen değişkenlerle elde edilen doğruluk tablosu Tablo 1'de gösterilmiştir.

Tablo 1. 7 duygu sınıfı için doğruluk tablosu

	F	A	N	S	B	D	H
F	32	9	13	6	6	1	1
A	5	113	1	0	0	1	2
N	5	1	58	3	7	2	1
S	1	1	19	40	3	0	0
B	5	2	41	11	17	1	0
D	15	9	14	4	4	8	2
H	12	22	6	1	3	2	10

Veri seti sınırlı olduğu için test işlemi için Jackknife metodu uygulanmıştır. Jackknife, yeniden örnekleme metotları içerisinde ilk denenen yöntemlerden biridir. [15] Belirli sayıda örnek içeren setin bir bölümü ayrılarak geriye kalan örnekler gözlem verileri olarak tutulur ve böylece Jackknife işlemi tamamlanır. %5 Jacknife uygulamasında her bir döngüde veri setinin farklı bir %5'lik kısmı test için ayrılarak kalan kısmı eğitim için kullanılır. Böylelikle 20 döngü sonucunda bütün set kapsanmış olarak işlem tamamlanır ve başarı sonucu her bir döngüde elde edilen tanıma oranlarının ortalaması alınarak hesaplanır.

Farklı duygu setleriyle oluşturulan alt kümeler için KNN ve SVM sınıflandırıcıları ile elde edilen sonuçlar Tablo 2'de gösterilmiştir.

Tablo 2. Sonuç tablosu

Sınıflandırıcı	Duygu Seti	Başarı
KNN	A, S, H, N	79 %
	A, S, H, F	73 %
	Y(A,H), N, K(S,B)	77 %
	A, S, H, F, N	66 %
	A, S, H, D, B, F	59 %
	A, S, H, N, D, B, F	53 %
SVM	A, S, H, D, B, F	57 %
	A, S, H, N, D, B, F	51 %

3.3 Sonuçların Analizi

Kaynak tarama bölümünde anlatılan çalışmaların özetlenmiş hali Tablo 3'de sunulmuştur. Elde edilen sonuçlar ve geçmiş çalışmalar incelendiğinde bazı çıkarımlar yapılabilir.

Tablo 3. Referans Çalışmalar

Ref.	Set	Boyut Azaltma	Sınıflama	Duygu #	%
[1]	MFCC	-	GMM	4	68
[2]			BPNN	5	79
[3]	LPCC	-	ANN	5	46
[4]	MFCC DFT, Prozodi	ERFTrees	KNN	5	72
[5]	Kalite, Prozodi	Relief	KNN	4	72
[10]		Doğrusal Regresyon	NN	3	59
[6]	LPCC	Vektör Basamaklama	HMM	6	56
	MFCC				59
[8]	MFCC, Kalite, Prozodi	-	1NN	7	63
		SVM SFFS			75
[9]		-	SVM		79

Öncelikle sadece MFCC katsayı seti ile yapılan bu çalışmada daha fazla duygu sınıfı için dahi LPCC [3],[6] ile elde edilen başarı oranının geçilmiş olduğu görülebilir.

Sadece MFCC vektörlerini ele alan diğer çalışmalarla kıyaslandığında, yapılan çalışmanın seçilen alt kümeye göre %68 ile %85 arasında sonuçlar veren sisteme [1] göre sonuçların daha dar bir aralıkta değişkenlik göstermesi de göz önünde bulundurularak daha kararlı olduğu söylenebilir. Yapay sinir ağı (BPNN) [2] kullanılan çalışmada nispeten daha az sayıda okuyucuya bağlı (6 aktörle) gerçekleştirilen sistemde ise daha yüksek bir oran elde edilmiştir.

Dar kapsamlı ayrıştırmalar için (4 sınıf) [5] de, MFCC vektörlerinin prozodi ve kalite vektörlerine göre daha başarılı çalıştığı söylenebilir. Aynı şekilde sınıfların gruplanarak [10] ayrıştırma kategorilerinin düşürüldüğü durumlarda kalite ve prozodi öznitelikleri %59 başarı sonucu verirken MFCC öznitelikleri için %77 oranında başarılı tanımlama sonucu elde edilmiştir.

Aynı alt kümede yapılan sınıflandırmalarda HMM [6] ve KNN sınıflandırıcılarının % 59, SVM sınıflandırıcısının ise % 57 sonucunu verdiğine dayanarak sınıflandırıcının kullanılan öznitelik seti kadar başarılı tanıma oranı üzerinde etkili olmadığı çıkarımı yapılabilir.

MFCC katsayı setinin LPCC, prozodi ve kalite vektörlerine göre daha başarılı sonuçlar verdiğinin tespitine ek olarak en yüksek oranda tanımlama başarısının, MFCC özniteliklerinin diğer özniteliklerle beraber kullanıldığı çalışmalarda elde edilmiş olduğu görülebilir. Beş sınıf [4] için performans çok az iyileşmiş olsa da; tam set üzerinde fark daha net görülebilir. 7 duygu üzerinden yapılan sınıflandırmalarda

([8], [9]) bu özniteliklerin beraberce kullanılmalarının başarı oranını %79'a kadar çıkarıldığı görülmektedir.
Yani aslında prozodi ve kalite öznitelikleri tek başına çok yüksek performans göstermemelerine rağmen MFCC tabanlı çalışan tanıma sistemlerine eklendiklerinde başarı oranını artıran etkenler olarak değerlendirilebilirler.

4. Değerlendirme

Bu çalışmada özetle yalnızca MFCC öznitelikleri kullanılmış, LPCC, prozodi ve kalite özniteliklerini kullanan diğer çalışmalarla kıyaslandığında daha başarılı sonuçlar elde etmiştir. Yapılan önişleme ve "k" parametre optimizasyonu ile sonuçlar iyileştirilmiş ve toplam set için % 53 oranında sınıflama başarısı elde edilmiştir. Yapılan iyileştirmelerle % 43'ten, % 53'e çıkan sonuç %10 civarında performans artırımını göstermektedir. Sınıflandırıcının etkisinin gözlemlenmesi için aynı koşullarla KNN ve SVM sınıflandırıcılarının performansları kıyaslanmış ve seçilen sınıflama metodunun performans üzerinde alt küme ve öznitelik kadar etkili olmadığı görülmüştür.

Kaynaklar

[1] Rao, K. S., Kumar, T. P., Anusha, K., Leela, B., Bhavana, I. and Gowtham, S. V., "Emotion Recognition from Speech", *International Journal of Computer Science and Information Technologies (IJCSIT)*, Vol. 3 (2), 2012.

[2] Gilke, M., Kachare, P., Kothalikar, R., Rodrigues, V. P. And Pednekar, M., "MFCC-based Vocal Emotion Recognition Using ANN.", *International Conference on Electronics Engineering and Informatics (ICEEI): IPCSIT vol. 49, IACSIT Press*, 2012

[3] Pathak, S. and Arun K., "Recognizing emotions from speech." *Electronics Computer Technology (ICECT)*, 2011, *3rd International Conference on. Vol. 4. IEEE,* 2011.

[4] Rong, J., Gang L., and Yi-Ping P. C., "Acoustic feature selection for automatic emotion recognition from speech.", *Information processing & management 45.3: 315-328*, 2009.

[5] Zhang, S. and Zhijin Z., "Feature selection filtering methods for emotion recognition in Chinese speech signal.", *Signal Processing, 2008. ICSP 2008. 9th International Conference on. IEEE*, 2008.

[6] Nwe, T. L., Say W. F. and Liyanage C. S., "Speech emotion recognition using hidden Markov models.", *Speech communication* 41.4:603-623, 2003.

[7] Dellaert, F., Thomas, P. and Alex W., "Recognizing emotion in speech." *Spoken Language, ICSLP 96. Proceedings., Fourth International Conference on*. Vol. 3. IEEE, 1996.

[8] Schuller, B., Müller, R., Lang, M. K. and Rigoll, G., "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles.", *INTERSPEECH*, 2005.

[9] Giannoulis, P. and Gerasimos P., "A hierarchical approach with feature selection for emotion recognition from speech." *LREC.*, 2012.

[10] Tato, R., Santos, R., Kompe, R. and Pardo, J. M., "Emotional space improves emotion recognition.", *INTERSPEECH*, 2002.

[11] Berlin Emotional Speech Database: http://pascal.kgw.tu-berlin.de/emodb/index-1024.html

[12] Young, S. J., Evermann, G., Gales, M. J. F., "The HTK Book", 2006

[13] Blitzer, J., Kilian Q. W. and Lawrence K. S., "Distance metric learning for large margin nearest neighbor classification." *Advances in neural information processing systems,* 2005.

[14] Campbell, W. M., Sturim, D. E., Reynolds, D. A., and Solomonoff, A., "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation." *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. IEEE International Conference on*. Vol. 1., 2006.

[15] MacKinnon, D. P., Lockwood, C. M. and Williams, J., "Confidence limits for the indirect effect: Distribution of the product and resampling methods." *Multivariate behavioral research* 39.1:99-128, 2004.