Araştırma Makalesi /Research Article

Etnomüzikoloji Dergisi
Ethnomusicology Journal
Yıl / Year: 7 • Sayı / Issue:2
(2024)

ÖĞRENCİ EZGİ VOKAL TEKRARI PERFORMANSLARININ OTOMATİK NOTLANDIRILMASI*

Barış BOZKURT**
Ozan BAYSAL***

Özet

Teknolojideki güncel gelişmeler sebebiyle çevrimiçi müzik eğitiminin giderek yaygınlaştığı bir döneme girmiş bulunuyoruz. Müzik eğitimi için çevrimiçi kaynaklar, dersler ve çevrimiçi eğitim alan müzik öğrencisi sayısı büyük hızla artmaktadır. Çok sayıda müzik öğrencisinin kaydolduğu çevrimiçi derslerde öğrenci müzik icralarının notlandırılması yüksek düzeyde uzman emeği gerektirmektedir. Bu sebeple, görece mekanik olan müzik egzersizlerinin notlandırılması için otomatik sistemlerin tasarımı önem kazanmaktadır. Bu çalışmada, piyano icrası işitilen bir ezginin tekrar edilmesine dayanan öğrenci vokal performanslarını otomatik notlandırılan bir sistem önerilmektedir. Sistem, öğrenci performans kaydı ile referans piyano kaydını karşılaştırarak bir not çıktısı üretir. Süreç, temel frekans serileri ve kroma matrislerinin hesaplanması, zaman hizalaması, mesafe dağılımlarının istatistiksel analizi ve makine öğrenimi ile not tahmini adımlarını içerir. Son adımda kullanılan makine öğrenmesi modeli güdümlü öğrenme yöntemiyle eğitilmiş, bu amaçla eldeki veriler üç ayrı uzman tarafından notlandırılmıştır. Uzman notlandırmaları arasındaki uyum ile eğitilen modelin çıktılarının uzman notlarıyla tutarlılığı ayrıntılı olarak karşılaştırılmış ve sonuçlar sunulmuştur.

Anahtar Kelimeler: Müzik eğitimi, performans değerlendirme

* *Makale Geliş Tarihi: 20 Eylül 2024 – Makale Kabul Tarihi: 10 Kasım 2024*
** Assoc. Prof., Zayed University, College of Interdisciplinary Studies. baris.bozkurt@zu.ac.ae
*** Prof. Dr., Istanbul Teknik Üniversitesi, Türk Musikisi Devlet Konservatuarı, Müzikoloji Bölümü. ozanbaysal@itu.edu.tr

Automatic Assessment for Student Melodic Pattern Imitations

Abstract

Online music education is increasingly gaining attraction globally. The number of music students profiting from online resources, lessons and online music education is growing rapidly. Evaluating student performances in online music classes with high enrollment demands substantial expert involvement. For this reason, the design of automatic systems for the assessment of relatively mechanical musical exercises is becoming crucial. In this study, we propose a system that automatically assesses student vocal performances repeating melodic patterns. The system analyzes both the student's performance and a reference piano recording, producing a grade based on the melodic similarity between the two. The system functions through four primary processes: extracting fundamental frequencies and chroma features, aligning the sequences via dynamic time warping, measuring the distribution of discrepancies, and generating scores using a machine learning algorithm (trained via supervised learning). We provide a study on the consistency between different experts and the outcomes from the machine learning tests for the proposed automated system.

Keywords: Music education, performance assessment

Giriş ve Literatür Özeti

Problem Tanımı

Çevrimiçi müzik eğitimi, hızla büyüyen ve önemli bir ekonomik büyüklüğe ulaşan bir alana dönüşmektedir. Berklee College of Music gibi dünyaca ünlü enstitüler dahil olmak üzere birçok eğitim kurumu çevrimiçi lisans programları oluşturmuş, çok sayıda öğrenci bu programlara kaydolmuş ve çevrimiçi eğitim görmeye devam etmektedir. Birçok teknoloji şirketi müzik eğitimi için mobil uygulamalar tasarlamış ve piyasaya sunmuş durumdadır. Çevrimiçi araçlar ve kaynakları kullanarak müzik öğrenmek isteyenlerin sayısı her geçen gün artmaktadır.

Çok sayıda öğrenciye hizmet veren bu eğitim sistemlerinde tüm süreci insan insana yürütmek yerine, sınırlı insan kaynağını daha etkin kullanmak için belirli noktalarda teknoloji kullanımı bir ihtiyaç haline gelmiştir. Bu konuda teknolojik desteğe ihtiyaç duyulan noktalardan birisi öğrencilerin eğitim planında sunulan egzersizleri yaparken/çalışırken kendilerine otomatik olarak geri bildirimde bulunan sistemlerin tasarlanmasıdır. Otomatik notlama ve geri bildirim seçeneği öğrencilerin derse ilgilerini canlı tutmaya önemli katkı sunmaktadır. Örneğin bilgisayar programlama derslerinde otomatik oluşturulmuş geri bildirimlerin sunulduğu eğitimlerde hiç bildirim sunulmayan eğitimlere göre öğrencilerin derse devam sürelerinin daha fazla olduğu, ilgilerinin daha

uzun süre devam ettiği raporlanmıştır (Galan, Heradio, Vargas, Abad ve Cerrada, 2019). Benzer bir gözlem bilgisayar destekli müzik eğitimi üzerine yapılan bir çalışmada da raporlanmıştır (Zhang ve Yi, 2021). Yazarlar, eğitim sürecine dahil ettikleri bir yazılım aracılığıyla öğrencilere performanslarındaki hataları otomatik olarak görselleştirilip geri bildirim olarak sunan bir sistemi test etmişler ve bu otomatik geri besleme sayesinde müzik öğretiminin etkinliğinin arttığını raporlamışlardır. Birçok etkileşimli çevrimiçi eğitim sistemi, arka planda otomatik notlandırma adımı içeren bu tür bir etkileşime dayanmaktadır.

Çevrimiçi eğitimde, çalınan/işitilen veya notası (veya başka bir sembolik gösterimi) verilmiş bir ezginin öğrenci tarafından sesle/vokalle söyleyerek icrası/tekrarı önemli yer tutmaktadır. Bu egzersiz, öğrencinin frekans algılama ve ezgi hafızası yeteneğini ölçmek için yaygın olarak sınavlarda da kullanılmaktadır. Bu çalışmanın kapsamı bu problemle sınırlandırılmıştır; bu çalışmada amaç, hedeflenen/sorudaki ezgi ve öğrenci icrasındaki ezgi arasında bir uzaklık ölçümü yapan ve bu uzaklıktan yola çıkarak öğrenci icrasına otomatik not veren bir sistem tasarlamaktır. Sistem tasarımı, uzmanlar tarafından notlandırılmış kayıtları içeren bir veri kümesi kullanılarak güdümlü öğrenme yöntemleri ile makine öğrenmesi modellerinin eğitilmesine dayanmaktadır.

Müzik eğitiminde elbette ezgi boyutu dışında ritmik yapı, tını, sanatsal yorum gibi boyutlar da vazgeçilmez derecede önemlidir. Bu çalışmanın sınırlarını belirlerken, ezgi boyutu, bir müzik icrasının görece net tanımlanabilir ve ölçülebilir temel boyutlarından birisi olarak düşünülmüş ve bu sebeple odaklanılmasına karar verilmiştir. Diğer boyutlar (örneğin tını ve sanatsal yorum başarısı) görece tanımlanması ve ölçülmesi daha zor boyutlardır. Teknolojik destek sunulması planlanan uygulama alanı, belirli motifleri tekrar etmeye dayanan ve görece mekanik müzik ezgersizi kategorisine girebilecek eğitim içerikleri olarak sınırlandırılmıştır.

Motiflerin tekrar ediliş şekli açısından iki tür uygulama/koşul düşünülebilir. Birinci tür uygulamada icra belirli bir altyapı üzerine beklenen bir tempoda, altyapı ile uyumlu olarak gerçekleştirilir. İkinci tür uygulamada bir altyapı bulunmamaktadır ve öğrenci serbest tempoda ezgi icrasını gerçekleştirir. Bu iki kategori için otomatik notlandırma problemi farklı zorluklar içerir. Bu çalışmada, konservatuvar giriş sınavlarındaki uygulamaya benzer Görselde ikinci kategoride icraların otomatik notlandırılması

hedeflenmiştir. Bu araştırma problemi, referans kayıt ile öğrenci kaydı arasında zamanda otomatik eşleme işlemleri yapılmasını gerektirdiği için ek bir zorluk içermektedir. Tasarlanan sistemin ilk adımı bu sebeple zamanda eşleme ve kırpma adımlarıdır. Bu adımları takiben, eşlenmiş kayıtlar arasında fark öznitelikleri hesaplanarak makine öğrenmesi modeline girdi olarak sunulmakta, modelin çıktısı olarak da not değeri elde edilmektedir.

Otomatik Vokal Performans Notlandırma Literatür Özeti

İnsan sesiyle yapılan bir müzik performansının bir müzisyen/eğitmen tarafından değerlendirilmesi birçok sübjektif faktörü içerir. Bu faktörlerin çeşitli boyutları çeşitli araştırmalarda ele alınmıştır: şarkı sözlerinin doğru telaffuzu (Jha ve Rao, 2012) şarkıcı formantı[1] seviyesi (Lundy, Roy, Casiano, Xue ve Evans, 2000), ses seviyesinin düzenli kontrolü (Tsai, Ma ve Hsu, 2015), vibrato özellikleri (Nakano, Goto ve Hiraga, 2006) ritim ve tonlama doğruluğu (Lin, Lee, Chen ve Wang, 2014) vb.. Bu çalışmanın ele aldığı boyut olan ezginin doğru icrası (entonasyon doğruluğu), üzerinde en fazla çalışmanın bulunduğu değerlendirme boyutudur. Bu problem ele alınırken, yaygın olarak kayıtlardan hesaplanan temel titreşim frekans (f0) serileri kullanılır (Molina, Barbancho, Gómez, Barbancho ve Tardón, 2013). Ses sinyal analizi perspektifinden bakıldığında, monofonik kayıtlarda temel frekans (f0) tahmini çoğu durumda yüksek güvenilirlikle gerçekleştirilebilmektedir. Bu sebeple, ele alınan problemin, verilen referans ezginin frekans serisi ile icranın frekans serisinin karşılaştırılmasına, diğer bir deyişle ezgi benzerlik düzeyi (İngilizce: melodic similarity) ölçümüne, indirgenmesi literatürde en yaygın kullanılan yöntemdir (Bozkurt, Baysal ve Yüret, 2017).

Müzik bilgi erişim alanında iki ezginin veya iki icranın ezgilerinin benzerlik düzeyinin ölçülmesine birçok farklı uygulamada ihtiyaç duyulmaktadır. Mırıldanarak müzik parçası arama, bir müzik geleneğindeki ezgilerin otomatik karakterizasyonu (örneğin makam/raga'lara özgü ezgi motiflerinin otomatik gruplama yöntemleriyle tespiti), yeni bestelenmiş bir ezginin repertuardaki eserlerle örtüşme düzeyinin ölçülmesi gibi birçok uygulama, ezgiler arası benzerlik/uzaklık ölçümüne ihtiyaç duyar. Bu sebeple, ezgi

[1] Şarkıcı Formantı, orkestra içerisinde insan sesinin duyurulabilmesi amacıyla geliştirilen vokal tekniklerinin kullanılması sonucu olarak ses spektrumunda 3 kHz çevresindeki frekans bandında, normal konuşma ses spektrumuna göreceli olarak enerji artışını açıklamak için kullanılan bir terimdir (Sundberg, 2001).

benzerliği ölçümü literatürü oldukça geniş bir alandır ve büyük oranda ezgi arama hedefine yönelik algoritmaların tasarımına odaklanmaktadır (Gulati, Serra ve Serra, 2015; Typke, 2007). Gulati vd. 2015, melodik benzerlik ölçüm yöntemleri için çok kapsamlı bir karşılaştırmanın sonuçlarını sunmaktadır: Hint Müziği için melodik örüntü bulma ve keşif bağlamında benzerliği hesaplamak için 560 farklı ezgi benzerlik ölçüm tekniği sürümünün[2] karşılaştırılması sonucunda yazarlar, genel olarak, Dinamik Zaman Bükümü (DTW)[3] tabanlı mesafe ölçümlerinin, benzerlik ölçümleri için kullanılan diğer yaklaşımlardan daha iyi performans gösterdiği sonucuna varmışlardır. Burada, yer sınırlılığı sebebiyle, literatür özeti ezgi benzerliği çalışmalarının genel bir özetini sunmak yerine öğrenci ezgi icralarının notlandırılmasına yönelik çalışmaların özetlenmesiyle sınırlandırılmıştır.

İnsan sesiyle ezgi icrası otomatik notlandırması alanındaki ilk çalışmalardan birisi olan (Nakano vd., 2006) ezgide nota geçişlerinde perde aralığı doğruluğu ve vibrato kalitesini ölçmeyi hedeflemektedir. Bunun için frekans serisinin 12 eşit yedirimli ses sistemi nota frekanslarına uzaklıklarının istatistiği ve vibrato kullanılan bölgelerde vibratonun frekans bant genişliği ve saniyedeki vibrato dalgası sayısı öznitelik olarak kullanılarak otomatik sınıflandırma gerçekleştirilmiştir. AIST-HMD veri kümesinden (Goto ve Nishimura, 2005) 600 örnek üzerinde yapılan testlerde başarılı/başarısız sınıflandırmasının %83.5 başarı ile gerçekleştirildiğini raporlamışlardır. Ses kaydı içerisinde özel bir boyutu, bir referansla karşılaştırmadan ele alan bu ilk çalışmaları takiben, literatürdeki çalışmaların, öğrenci icrası ile hedeflenen bir referans nota dizisinin karşılaştırılmasına yöneldiklerini görüyoruz. Otomatik ezgi icrası notlaması literatüründe 2013-2018 arasında bu perspektifi kullanan bir dizi çalışmanın benzer bir yaklaşımı küçük değişikliklerle uyguladığı görülmektedir; (Abeßer, Hasselhorn, Dittmar, Lehmann ve Grollmisch, 2013; Bozkurt, Gulati, Romani Picas ve Serra, 2018; Molina vd., 2013; Schramm, de Souza Nunes ve Jung, 2015) içerisinde önerilen sistemler sırasıyla şu üç adımı içermektedir: i) öğrenci ezgi icrasının otomatik olarak notaya dökülmesi (akustik

[2] Ezgi benzerlik ölçüm tekniği sürüm sayısının bu derece fazla olması, tekniklerin birçok adım içeriyor olması ve her adım için ayarlanabilir bazı parametrelerin bulunmasından kaynaklanmaktadır. Yazarlar birçok alt bileşen için birçok farklı parametreyi optimize edebilmek için çok sayıda olası kombinasyonu birleştirdikleri bir deney gerçekleştirip sonuçları raporlamışlardır.

[3] İki zaman serisi verisinin zamanda eşlenmesi/hizalanması için kullanılan Dinamik Zaman Bükümü tekniği Bölüm 3.3'te özetlenmektedir.

ses dizisinin sembolik diziye (notaya) çevrilmesi), ii) referans nota dizisi ile icra nota dizisi arasında frekans ve süre farklılıklarının istatistiksel parametrelerinden (ortalama fark, farkın standart sapması, vb.) öznitelik vektörünün oluşturulması, iii) öznitelik vektörünü not değerine eşleyen bir makine öğrenmesi modelinin eğitilip kullanılması. Bu çalışmalarda genel mimari benzer olmakla beraber, her adımda kullanılan teknikler büyük farklılıklar içermektedir. Literatürdeki bu çalışmalarda çok farklı veri toplama stratejileri kullanılmıştır ve raporlanan sonuçlar arasında da büyük farklılıklar bulunmaktadır. Örneğin, Molina vd. 2013, sentetik ses verilerinden oluşan bir veri kümesi kullanmışlar, sentetik verileri gerçek ses kayıtlarına rastgele perde/ritim varyasyonları uygulayarak elde etmişlerdir. Bu veri kümesi üzerinde oldukça yüksek bir başarı (tahmin edilen not ile gerçek not değeri arasında 0.97 korelasyon) raporlamışlardır. Schramm vd. 2015, yedi yetişkinden (üç eğitimli ve dört eğitimsiz şarkıcı) 21 seansta topladıkları performans kayıtlarını içeren bir veritabanını kullanmış ve 0.96'lık ikili otomatik sınıflandırma (başarılı/başarısız) doğruluk değeri bildirmişlerdir. Abeßer vd. 2013, Alman okullarında dokuzuncu ve onuncu sınıf öğrencilerinden alınan 617 şarkı kaydını içeren bir veri kümesi kullanmışlardır. Yazarlar, bu veri kümesinde benzer bir metodoloji uygulayarak, %55,7'lik bir ikili sınıflandırma doğruluğu bildirmişlerdir. Tsai vd. 2015, Karaoke uygulamalarından topladıkları veriler (20 Mandarin şarkı klibinin solo vokal bölümleri için 25 şarkıcının kaydı) üzerinde yaptıkları testlerde 0.80 ikili sınıflandırma doğruluğu raporlamışlardır. Schramm vd. 2015, ise önerdikleri sistemin başarısını 0.88 (ikili sınıflandırma) doğruluk skoru ile bildirmiştir. Huang ve Lerch, 2019, pYIN (Mauch ve Dixon, 2014) algoritması ile kestirilen frekans serilerinin zamanda eşlenmesini takiben elde edilen entonasyon, zamanlama, dinamikler ve nota sürekliliği gibi öznitelikleri Temel Bileşen Analizi (Hotelling, 1933) ile birleştirerek otomatik notlandırma performansının artırılmasını hedeflemiştir. Bu çalışma için kullanılan veri kümesi "Florida Bandmasters Association" (FBA) tarafından düzenlenen ortaokul ve lise öğrencilerinin katıldığı seçme sınavlarında 2013, 2014 ve 2015 yıllarında yapılan (saksafon) ses kayıtlarından oluşmaktadır ve sistemin sınıflandırma başarısının gerçek hayatta kullanım için gerekli olandan çok düşük olduğunu raporlamışlardır.

Her çalışmanın farklı veri kümesi üzerinde raporladığı sonuçlar karşılaştırılabilir olmamakta, bu engelin aşılması için açık erişimi olan verilere ihtiyaç duyulmaktadır. Bu yönde atılan ilk adımlardan birisi Bozkurt vd. 2017 ile paylaşılan verilerdir. Kayıtların

kendilerini içermeyen, sadece temel titreşim frekans verilerini ve ikili sınıf etiketlerini (başarılı/başarısız) içeren bu veri kümesi, Pati, Gururani ve Lerch, 2018 gibi bir dizi çalışmada yöntemleri karşılaştırma amaçlı kullanılmış olmakla beraber ses verilerini içermemesi ve sadece başarılı/başarısız etiketleri içermesi itibariyle bu alandaki ihtiyacı karşılamaktan uzaktır.

2016 sonrası çalışmalarda, öğrenci icrasının otomatik notaya dökülmesi adımının atlandığı ve ses kaydından kestirilen frekans serisi ile hedeflenen sembolik verinin karşılık geldiği frekans serisi arasında uzaklığı yapay sinir ağları ile ölçmeyi hedefleyen sistemler de önerilmiştir. Bu çalışmalar, derin öğrenme modellerini temelde üç farklı yaklaşımla uygulamaktadırlar. Birinci tür yaklaşımda frekans serileri doğrudan derin ağa girdi olarak verilerek, ağın performans için verilen not ile girdi arasındaki ilişkiyi eğitim sırasında öğrenmesini sağlamak hedeflenmektedir (Bozkurt vd., 2017; Pati vd., 2018). İkinci tür yaklaşımda, referans kaydın frekans serisi (veya spektrogram gibi başka bir spektral temsili) ile öğrenci icra kaydının frekans serisi arasında dinamik zaman bükümü işlemi için hesaplanan fark matrisinin derin ağa girdi olarak verilmesi tercih edilmektedir (Yang, Wang, Tian, Xu ve Cheng, 2022). Üçüncü tür yaklaşımda ise metrik öğrenme metodları ile spektral temsiller arası uzaklıkların derin ağlarla otomatik öğrenilmeye çalışılmasıdır (Seshadri ve Lerch, 2021; Zhang, Jiang, Jiang ve Peng, 2021). Derin ağ temelli sistemlerinin eğitiminde çok büyük veri kümelerine ihtiyaç duyulmaktadır. Bu modellerin başarıları ancak veri kümesi çok büyük olduğunda standart algoritmaların başarılarını aşabilmektedir.

Özgün Katkı

Bu çalışmada, bir piyano referans kaydını ve öğrenci icrası kaydını girdi olarak alan, kayıtlardan hesaplanan frekans serileri ve kroma özniteliklerini bir arada kullanarak otomatik notlandırma yapan bir sistem tasarlanmış, Python dilinde gerçeklenmiş[4] ve araştırmacılarla açık kaynak kodlu olarak paylaşılmıştır[5]. Önerilen sistemin uzman notlandırmaları ile detaylı karşılaştırılması yapılmış ve uzmanların kendi içlerindeki

[4] "gerçeklenme" ifadesi bu metinde, diğer birçok teknik metinde olduğu gibi, "bir kararı veya planı yürürlüğe koymak, uygulamak" anlamında kullanılan İngilizce terim olan "implementation"ın Türkçe karşılığı olarak kullanılmıştır.

[5] Verileri ve sonuçları da içeren kod kütüphanesi: https://github.com/barisbozkurt/auto-assess-melody-imitation

uyum düzeyine yakın düzeyde uyum içeren notların sistem tarafından otomatik olarak üretilebildiği gözlenmiştir.

Çalışma, şu yönleri ile literatüre katkıda bulunur: öğrenci ezgi vokal tekrarı performanslarının otomatik notlandırılması konusunda açık veri ve açık kaynak kodlu olarak testleri tamamen tekrarlanabilir ilk çalışmadır. Makine öğrenmesi modeline girdi olarak verilen temel titreşim frekans serileri arasındaki fark özniteliklerine kroma temsili fark özniteliklerinin eklemlenmesi, literatürdeki yöntemlerle karşılaştırıldığında yöntemsel farklardan birisidir. Önceki çalışmalarda jürinin verdiği geçti/kaldı şeklinde iki sınıflama problemi olarak ele alınıyor olmasının problemi anlamak için yeterli olmadığına karar verilmiş ve 4 seviyede körleme ve rastgele sırayla dinleyerek notlandırma yapılmıştır. Jürinin bir ekip olarak karar vermesi yerine ayrı ayrı ve rastgele sırayla notlama yapmaları etiketlerin karşılaştırmalı analizi için de imkân sağlamaktadır. Literatürdeki yöntemlerin karşılaştırılması için ihtiyaç duyulan çok geniş bir veri kümesi bu Görselde etiketlenerek paylaşıma açılmıştır.

Makine öğrenmesi modeli olarak literatürde bulunan teknikler kullanılmış, bu yönden özel bir katkı hedeflenmemiştir. Benzer Görselde, dinamik zaman eşlemesi, kroma temsili hesaplanması gibi adımlar literatürde bulunan güncel algoritmalar ve araçlar kullanılarak gerçekleştirilmiş, bu adımlarda yenilik hedeflenmemiştir.

Makalenin Organizasyonu

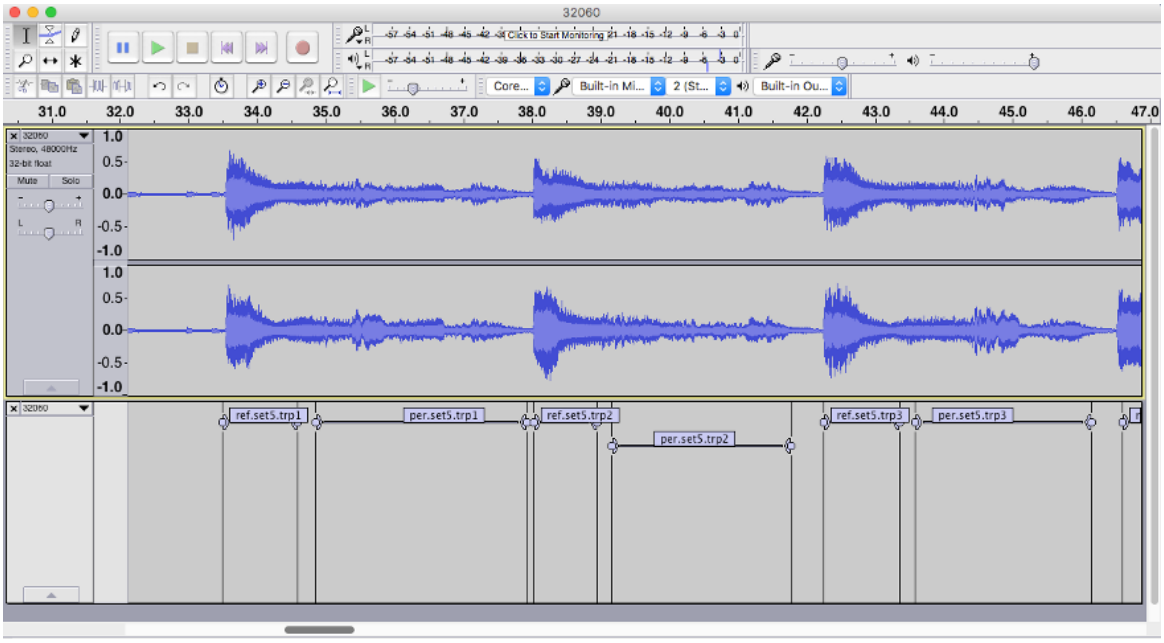
Makalenin ikinci bölümünde öncelikle veri toplama adımlarını açıklanmakta ve veriler etiketleri (uzmanların atadığı/verdiği notlar) üzerinde analiz sonuçları sunulmaktadır. Üçüncü bölümde önerilen sistem detaylı bir Görselde sunulmaktadır. Dördüncü bölüm tartışmalar ve sonuçlara ayırılmıştır.

Veri Toplama, Etiketleme ve Etiketlerin Analizi

Ses Kayıtlarının Toplanması

Bu çalışmada kullanılan veriler 2015 ve 2016 yıllarında konservatuar eleme sınavları kayıtlarından elde edilmiştir. Veritabanı 20 soru setinden 40 farklı melodinin performanslarını içermekte olan 1046 adet aday performans kaydından oluşmaktadır. Her soru setinde ilk melodi majör tonalitede ve 4/4'lük ölçü biriminde, ikinci melodi ise harmonik minör modlarından birinde (genellikle birinci veya beşinci derece eksenli) ve 9/8'lik ölçü birimindedir. Bu araştırmanın amacı doğrultusunda sınav kayıtlarının

bilimsel çalışmalar için kullanılabilmesi için, Etik Kurul onayı ve konservatuvar müdürlüğünden alınan izinler ertesinde öncelikle adayların sınav performanslarına ait video kayıtları anonimleştirilmiş ve ses dosyalarına çevrilmiştir. Yanı sıra, bu ses dosyaları herhangi bir konuşma sesi içermeyecek Görselde tekrar kontrol edilmiş, kayıtların sadece referans piyano sesi ve adayın bir hece ile yaptığı icralardan oluşmasına dikkat edilmiştir. Ardından analiz için Audacity programı kullanılarak, kayıtlarda sadece ilgili bölümleri işleyebilmek için, gerekli bölütleme işaretlemeleri yapılmıştır (Görsel 1). Bölütleme bilgileri kullanılarak kayıtlardan kesitler (ses yükseklik normalizasyon adımlarından sonra) küçük ses dosyaları olarak kaydedilmiş ve uzmanların dinleyerek gerçekleştirecekleri notlandırma işlemleri için hazır hale getirilmiştir.



Görsel 1. Ses Dosyalarının Audacity Programında Soru Tipi, Referans Ses ve Aday Performanslarına Göre İşaretlenmesi

Kayıtların Uzmanlar Tarafından Etiketlenmesi/Notlandırılması

Çalışmada gerçeklenen otomatik notlandırma algoritması kayıt çiftini girdi olarak alıp bir kaydın diğerine olan uzaklığına bağlı olarak notlandırmasını yapmak üzere kurgulanmıştır. Bunun için veri kümesi kayıtları üç uzman tarafından 1-4 skalasında körleme olarak rastgele sırayla notlandırılmıştır. Notlama skalası olarak temsil edilen değerler:

1- Çok başarısız, 2- Ciddi hatalar içeriyor, 3- Küçük hatalar içeriyor, 4- Başarılı

olarak tanımlanmış olup literatürde benzer problemlerde yaygın olarak kullanılmaktadır (Wesolowski, Wind ve Engelhard, 2016).

Notlama öncesi uzmanlar ile bir toplantı yapılmış ve notlama sırasında dikkat edilmesi gereken bazı ölçütler belirlenmiştir. Buna göre, performanslarda,

· Eksik (veya yanlış zamanda) okunmuş her nota için 10 puan,

· Çeyrek-tona yakın entonasyon hatası olan her nota için 5 puan,

· İlgili tonalite/modalitede yanlış okunan (veya eklenen) her nota için 15 puan,

· İlgili tonalite/modalite dışında yanlış okunan (veya eklenen) her nota için 20 puan eksiltilmelidir.

Elde edilen puana göre notlama skalası aşağıdaki gibi olacaktır;

1- Çok başarısız; 0-49 puan,

2- Ciddi hatalar içeriyor; 50-69 puan,

3- Küçük hatalar içeriyor; 70-89 puan

4- Başarılı; 90-100 puan.

Uzmanların notlandırma işlemini farklı zamanlarda gerçekleştirdiklerinde farklı notlar verebilecekleri öngörülmüş ve bu değişimi de gözleyebilmek için iki uzmandan 3 ay sonra tüm veriyi başka bir rastgele sırayla tekrar notlaması talep edilmiştir. Bu Görselde 5 ayrı etiket grubu elde edilmiştir. Bir sonraki bölümde bu 5 etiket grubunun birbiriyle karşılaştırmalı analizini ele alınmaktadır.

Etiketlerin Analizi (Uzman Notlarının Karşılaştırmalı Analizi)

Etiketler arasında uyuşma düzeylerini ölçmek için bütün olası etiket ikileri için Ortalama Mutlak Hata/Fark (OMH)[6], Örtüşme Oranı[7] (tam örtüşme: 1), Krippendorf Alfa ve Pearson korelasyon değerleri ölçülmüş, alttaki tabloda sunulmuştur.

Tablo 1. Uzman etiketleri karşılaştırması. OMH: Ortalama Mutlak Hata

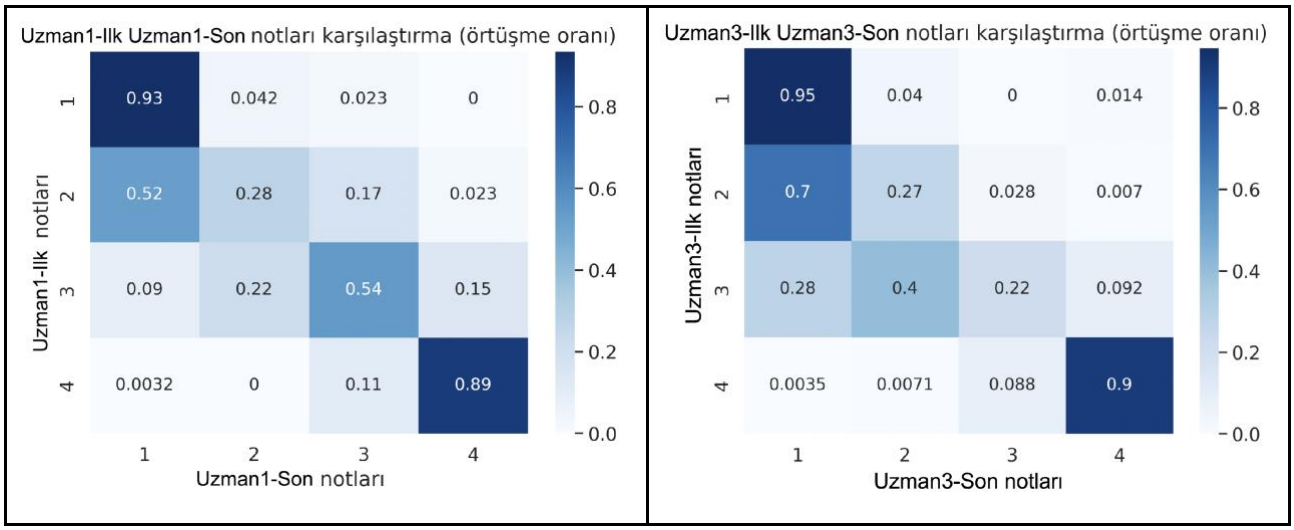
[6] OMH değeri uzmanların aynı kayıt için verdiği notlar arasındaki farkın mutlak değerinin bütün kayıtlar üzerinden ortalaması hesaplanılarak elde edilmektedir.

[7] Örtüşme oranı, uzmanların aynı notu verdikleri kayıt sayısının toplam kayıt sayısına bölünmesiyle elde edilir.

Karşılaştırılan Etiketler	**OMH**	**Örtüşme Oranı**	**Krippendorf Alfa**	**Pearson Korelasyon**
Uzman1-İlk - Uzman2	0.46	0.6	0.45	0.85
Uzman1-İlk - Uzman1-Son	0.36	0.67	0.55	0.87
Uzman1-İlk - Uzman3-İlk	0.37	0.67	0.55	0.86
Uzman1-İlk - Uzman3-Son	0.53	0.56	0.38	0.84
Uzman2 - Uzman1-Son	0.41	0.65	0.5	0.85
Uzman2 - Uzman3-İlk	0.39	0.65	0.51	0.85
Uzman2 - Uzman3-Son	0.35	0.7	0.54	0.86
Uzman1-Son - Uzman3-İlk	0.43	0.64	0.5	0.82
Uzman1-Son - Uzman3-Son	0.46	0.66	0.49	0.82
Uzman3-İlk - Uzman3-Son	0.38	0.67	0.51	0.87

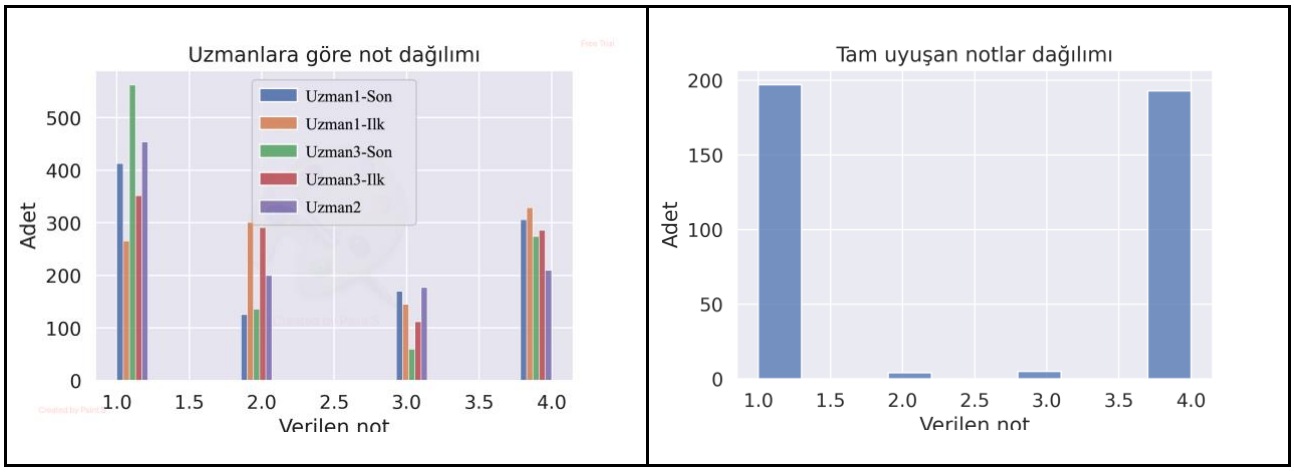
Tablodaki değerleri kısaca özetlemek gerekirse, 0.82 ile 0.87 değerleri arasında değişen Pearson Korelasyonu etiketler arasında güçlü bir pozitif ilişki olduğunu göstermektedir. Krippendorf Alfa değerlerinin ise 0.38 ile 0.55 arası değişim gösterdiği görülmektedir, bu da etiketler arasında makul ve orta dereceli bir uyuma işaret eder. Gerek Ortak Mutlak Hata/Fark (OMH) gerekse Örtüşme Oranı değerlerine bakıldığında karşılaştırılan etiketler arasında ortaklık olmakla birlikte tam bir örtüşmenin olmadığı anlaşılmaktadır.

Etiketleme işlemini birkaç ay sonra tekrarlayan uzmanların ilk verdikleri notlar ile son verdikleri notların örtüşme düzeyi aşağıdaki karıştırma matrislerinde sunulmaktadır. Matrisler satırlar düzeyinde normalize edilmiştir. Örneğin, soldaki matrisin 2. satırı ele alındığında; Uzman1-İlk etiketlerinde 2 notu verilmiş örneklerin 0.52'lik oranlık kısmına sonraki etiketlemede (Uzman1-Son) 1 notu, 0.28'lik oranlık kısmına 2 notu, 0.17'lik oranlık kısmına 3 notu ve 0.023'lük oranlık kısmına 4 notu verilmiştir. Uzmanların tekrar etiketleme yaptıklarında, daha önce 2 veya 3 notu verdikleri örneklere verdikleri notları büyük oranda değiştirdikleri görülmektedir.



Görsel 2. Uzmanların tekrar notlandırma sonrası notlarının karşılaştırması

Toplam 1046 öğrenci performansı üzerinden yapılan notlamalarda 5 etiket grubunda da aynı not verilen örneklerin sayısı 399 olup bu notların çok büyük kısmı 1 ve 4 notlarıdır. Tablo 1, Görsel 2 ve 3 ele alındığında, çok başarılı ve çok başarısız icra örneklerinin etiketlenmesi işlemini jürilerin uyumla ve güvenle gerçekleştirdikleri ancak ara notlarda uyum düzeyinin oldukça düştüğü ve jürilerin aynı örneği tekrar etiketlediğinde kararlarını değiştirmeye eğilimli oldukları gözlenmiştir. Örneğin Görsel 2-sağ'da sunulan Uzman3-ilk ve Uzman3-son etiketleri karşılaştırıldığında, aynı uzman (Uzman3) tarafından ilk etiketlemede 2 notu verilen örneklere birkaç ay sonra ikinci kez not verildiğinde %70'ine 1, %27'sine 2, %2.8'ine 3 notu verildiği gözlenmiştir. Benzer Görselde ilk etiketlemede 3 notu verilen örneklerin %28'ine ikinci notlandırmada 1, %40'ına 2, %22'sine 3 ve %9.2'sine 4 notu verilmiştir. Tüme etiket verilerine ve detaylı karşılaştırma sonuçlarına Bölüm 1.3'te verilen link üzerinden erişilebilir.



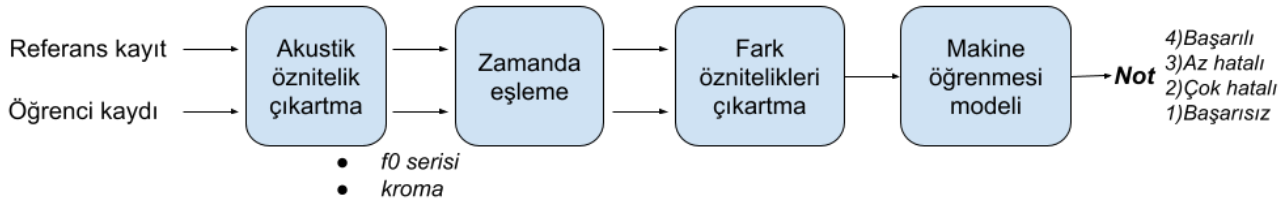
Görsel 3. Not dağılımları

Yukarıdan da anlaşılacağı üzere "Başarılı" (4) ve "Çok Başarısız" (1) performanslar konusunda uzmanlar arasında (ve aynı uzmanın farklı zamanlarda verdiği notlar arasında) güçlü bir uyum söz konusu olmakla birlikte ara notlar (2 ve 3) dahil edildiğinde zayıf bir uyum gözlenmektedir. Bu icralardaki hataların ne kadar "ciddi hata" veya ne kadar "küçük hata" olarak değerlendirileceği konusunda uzmanların not verirken, ezgideki notaların temel frekansı dışındaki faktörlerden etkilenmiş olabilecekleri düşünülmektedir. Bu konuya sonuç bölümünde kısaca değinilecektir.

Önerilen Otomatik Notlandırma Sistemi

Sistem Mimarisi ve Kullanılan Kütüphaneler

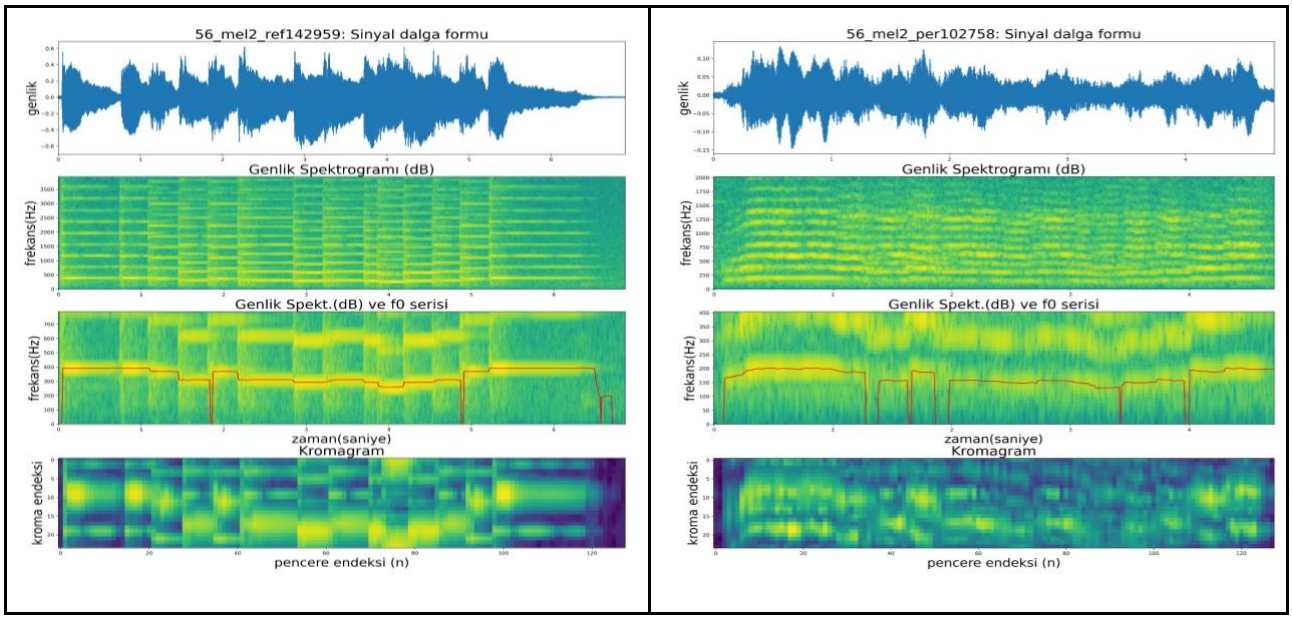
Önerilen sistem aşağıdaki Görselde özetlenmiştir. Sistem Python dilinde gerçeklenmiş olup açık kaynak kodlu olarak Bölüm 1.3'te verilen link üzerinden paylaşılmaktadır. Kodlar hazırlanırken librosa, dtw-python, crepe, resampy, scipy, libfmp, sklearn, xgboost kütüphanelerinden, kayıtlardan öznitelik hesaplama ve makine öğrenmesi testlerin gerçekleştirilmesi adımlarında, faydalanılmıştır. Aşağıda her bir bileşen alt başlıklarda ele alınmıştır.



Görsel 4. Önerilen sistemin mimarisi

Akustik Öznitelik Çıkartma/Hesaplama

Sistemde akustik öznitelikler olarak temel titreşim frekans (f0) serisi ve kroma özniteliği kullanılmaktadır. Görsel 5'de bir kayıt ikilisi (sol kolonda bir referans piyano kaydı ve sağ kolonda bir öğrenci performans kaydı) için sinyal dalga formu ve çıkartılan/hesaplanan öznitelikler görselleştirilmiştir.



Görsel 5: Kayıtların dalga formları ve elde edilen öznitelikler. Yukarıdan aşağı: 1) sinyal dalga formu, 2) genlik spektrumu (0 - 2 kHz), 3) genlik spektrumu (0 - 400 Hz) ve Crepe kütüphanesi ile hesaplanan temel titreşim frekans serisi (kırmızı çizgi), 4) Librosa kütüphanesi kullanılarak hesaplanan 24 boyutlu kromagram. Sol kolon: referans olarak kullanılan piyano kaydı, Sağ kolon: öğrenci performans kaydı

Kayıttan temel titreşim frekans serisi hesaplanması

Problemde ele alınan kayıtlar monofonik olup referans kayıtlarda piyano, öğrenci kayıtlarında ise insan sesi bulunmaktadır. Müzik bilgi erişim uygulamalarında ezgiye dair bilgiye ihtiyaç olduğunda en yaygın kullanılan temsillerden birisi temel titreşim frekansının zamanla değişimine karşılık gelen melografdır (Carmi-Cohen, 1964). Güncel sistemlerde, bu temsili elde etmek için sinyalin küçük kesitlerinde tekil frekans değerleri kestirilip arka arkaya dizilerek frekans serileri (frekans değerlerini ardışık olarak içeren vektörler) elde edilir. Konuşma ve müzik ses işlemede en çok çalışılan konulardan birisi olan bu alanda onlarca farklı teknik önerilmiş ve kullanılmıştır. Bu çalışmada en güncel araçlar içerisinde üç tanesi seçilerek, öncelikle iKala veri kümesi (Chan, Yeh, Fan, Chen, Su, Yang ve Jang, 2015) üzerinde performansları, müzik bilgi erişim alanındaki standart metrikler (*mir_eval* Python kütüphanesi[8] (Raffel, McFee, Humphrey, Salamon, Nieto, Liang ve Ellis, 2014) kullanılarak karşılaştırılmıştır. Bu araçlar şunlardır:

[8] https://craffel.github.io/mir_eval/

- Essentia kütüphanesi[9] (Bogdanov, Wack, Gómez, Gulati, Herrera, Mayor, Roma, Salamon, Zapata, Boyer, Mayor ve Serra, 2013) içerisindeki Melodia algoritması/aracı (Salamon ve Gomez, 2012).
- Librosa kütüphanesi[10] içerisindeki pYIN algoritması/aracı (Mauch ve Dixon, 2014)
- Crepe algoritması/aracı[11] (Kim, Salamon, Li, ve Bello, 2018)

Ham frekans doğruluğu (İngilizce: "raw pitch accuracy") kriteri[12] dikkate alınarak gerçekleştirilen karşılaştırma sonucu Crepe algoritmasının/aracının performansının en yüksek olduğunun gözlenmesi üzerine çalışmada bu algoritmanın kullanılmasına karar verilmiştir. Bu algoritma/araç kullanılarak veri kümesindeki tüm kayıtlar işlenmiş ve frekans serileri elde edilmiştir.

Kayıttan kroma özniteliği hesaplanması

Ezgi analizinde diğer yaygın kullanılan öznitelik kroma özniteliğidir. Kroma özniteliği, bir ses sinyal kesitinde her bir notaya düşen ortalama enerjiyi temsil eden çok boyutlu bir vektördür (Ewert, 2011). Örnekle açıklayacak olursak, kroma vektörü içerisinde her bir nota için bir toplam enerji değeri bulunmaktadır. Örneğin Do notasının enerji değeri, sinyalin spektrumunda her oktavdaki Do notasına karşılık gelen dar frekans bölgelerindeki enerjiler toplanarak elde edilir. Müzik bilgi erişim uygulamalarında yaygın olarak Batı popüler ve klasik müziği nota uzayı dikkate alındığı için bir oktavda 12 nota olduğu varsayılır ve kroma vektörleri 12 boyutlu olarak hesaplanır. Bu çalışmada daha detaylı entonasyon temsili gerekliliği sebebiyle oktavın logaritmik 24 eşit parçaya bölündüğü varsayılmış ve 24 boyutlu vektörler elde edilmiştir. Analizi yapılan ses sinyalinin küçük kesitleri için elde edilen kroma öznitelikleri yan yana dizildiğinde, kromagram adı verilen spektrograma benzer bir zaman-frekans temsili elde edilir. Görsel 5'in en alt bölümünde iki kaydın kromagram görüntüsü sunulmuştur. Bu çizimler hemen üst bölümdeki frekans serisi ve spektrogram ile karşılaştırıldığında kromagramın tek oktavlık bir alanla sınırlandırılmış ve temel titreşim frekans bilgisi ile yüksek derecede ilişkili bir spektrogram temsili olduğu görülmektedir. Kromagram temsili, ses içeriğindeki notaların enerjisine dair bir temsil olduğu için yaygın olarak otomatik akor tespiti

[9] https://essentia.upf.edu/
[10] https://librosa.org/
[11] https://github.com/marl/crepe
[12] https://craffel.github.io/mir_eval/#mir_eval.melody.raw_pitch_accuracy

(McVicar, Santos-Rodríguez, Ni ve De Bie, 2014), otomatik müzik-nota senkronizasyonu/hizalaması (Kirchhoff ve Lerch, 2011) gibi uygulamalarda yaygın olarak kullanılan bir özniteliktir. Çalışmada kroma özniteliklerini hesaplamak için Librosa kütüphanesi içindeki chroma_stft[13] fonksiyonu kullanılmıştır.

Zamanda Eşleme/Hizalama

Eldeki problem, farklı uzunlukta iki kaydın ezgi benzerliğinin ölçülmesini gerektirmektedir. Bunun için bir önceki adımda hesaplanan farklı uzunluktaki öznitelik vektörlerinin ilgili bölgelerinin birbiriyle eşleştirilmesi gerekmektedir. Farklı uzunluklardaki parametre serilerinin zamanda eşlenmesi işlemi yaygın bir Görselde literatürde dinamik zaman bükümü (dynamic time warping (DTW)) algoritmaları kullanılarak gerçekleştirilir. Bu, verilen iki seriyi farklı bölgelerinde esneterek veya daraltarak serileri zamanda eşlemeyi sağlayan bir algoritmadır. Görsel 6'da, kullanılan veri çiftinden birisinde DTW eşleme öncesi ve sonrası parametre serilerini sunulmuştur. DTW algoritması parametre serilerinin bütün karşılıklı elemanları arasındaki uzaklıklardan oluşan bir maliyet matrisinde en az toplam maliyet için ilk serinin hangi elemanıyla ikinci serinin hangi elemanı arasında eşleme yapılması gerektiğini bulan bir dinamik programlama algoritmasıdır (Giorgino, 2009). Matematiksel olarak, DTW bir eniyileme problemi olarak Eşitlik (1)'de verildiği gibi tanımlanabilir:

$$DTW_q(x, x') = \min_{P \in A(x,x')} \left(\sum_{(i,j) \in P} d(x_i, x_j')^q \right)^{\frac{1}{q}} \quad (1)$$

Burada, K uzunluğundaki hizalama yolu P, K dizin çiftlerinin $((i_0, j_0),\ldots\ldots,(i_{K-1}, j_{K-1})$ bir dizisi ve $A(x, x')$ tüm kabul edilebilir yolların bir kümesidir. Yolun kabul edilebilir olarak değerlendirilmesi için şu koşulları sağlaması gerekir:

- Zaman serilerinin sırasıyla başlangıç ve bitişleri birlikte eşleşmelidir:
  - $P_0 = (0, 0)$
  - $P_{K-1} = (n-1, m-1)$ k
- Dizi hem i, hem de j'de monotonik artandır ve tüm zaman serisi dizinlerinin/endekslerinin en az bir kez bulunması gereklidir:
  - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
  - $j_{k-1} \leq j_k \leq j_{k-1} + 1$

[13] https://librosa.org/doc/main/generated/librosa.feature.chroma_stft.html

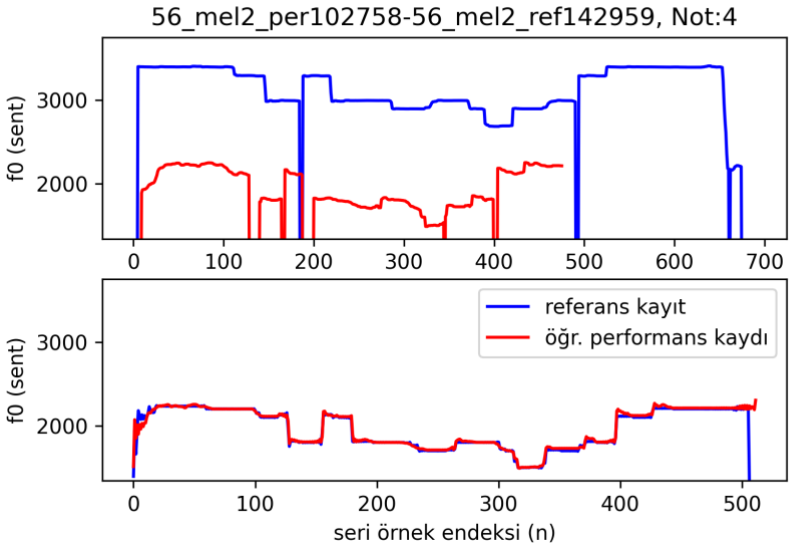
Elemanlar arası uzaklıklar hesaplanırken, eleman değerleri tekil sayılar olduğu durumda (frekans serileri arasında eşleme yapıldığı durumda) mutlak fark kullanılması yaygındır ve biz de bu fark hesabını kullandık. 24 boyutlu kroma vektörleri arasındaki DTW maliyet matris değerleri hesabında çok farklı uzaklık ölçütleri kullanılabilmekle beraber, bu çalışmada, yaygın kullanılan L1 uzaklığı kullanılmıştır[14]. 24 boyutlu kroma özelliklerinin, L1 uzaklığı kullanarak, dinamik zaman bükümü ile eşlenmesi için ise libfmp Python kütüphanesi (Müller ve Zalkow, 2021) kullanılmıştır.

$a = (a_1, a_2, \ldots, a_n)$ ve $b = (b_1, b_2, \ldots., b_n)$ R^N'de tanımlı 2 vektör olmak üzere L1 (Manhattan) Uzaklığı şu Görselde hesaplanır:

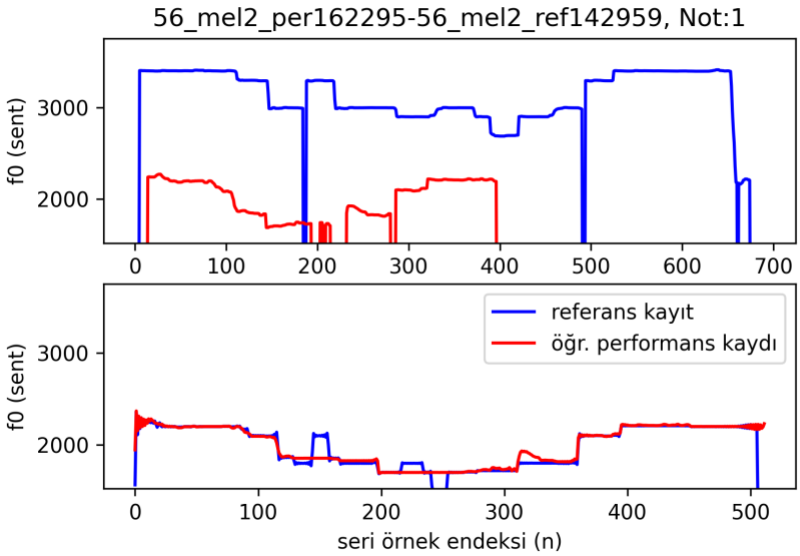
$$d_1(a, b) = \| a - b \|_1 = \sum_{i=1}^{N} |a_i - b_i| \quad (2)$$

Bu problemde, öğrencinin görece uzun bir ezgiyi icra etmesi gerektiğinde kısa nefes boşlukları bırakmasında sakınca bulunmamaktadır. Yine, referans kayıtlarında da yer yer sesler arasında küçük boşluklar bulunabilmektedir. Bu boşlukların zamanda eşleme işlemini etkilememesi için öncelikle frekans serisinde bulunan sıfır değerleri serilerden çıkartılmakta (boşluklar dışarıda bırakılmakta) ve uzun olan seri, kısa olanın boyutuna gelecek Görselde tekrar örneklenmektedir (tersi işlem (kısa serinin uzun ile aynı boyuta getirilmesi) sinyalin ani değiştiği bölgelerde ek bileşenler getirebileceği için bu tercih yapılmıştır). Giorgino, 2009 tarafından paylaşılan *dtw-python* pakedi kullanılarak zamanda eşleme işleminin sonuçlarına dair aşağıda iki örnek sunuyoruz. İlk örnek Görsel 5'te sunulan kayıtlara ait olup bu örnekte öğrenci performansı tam not (4) almıştır. İkinci örnekte öğrenci performansı en düşük notu (1) almıştır.

[14] L2 uzaklığı, Kosinüs uzaklığı ve Hamming uzaklığı ile de testler gerçekleştirilmiş, kullanılan uzaklık tanımına bağlı olarak testler sonucu elde edilen farklar az ve L1 uzaklığı lehine olması sebebiyle bu metne o testlerin açıklanması ve sonuçlarının tartışılması dahil edilmemiştir.



Görsel 6. Görsel 5'de sunulan referans ve öğrenci performans kayıtlarının frekans serileri zamanda eşlenmeden önce (üst Görsel) ve eşlendikten sonraki hali (alt Görsel). Bu örnekte öğrenci tam not almıştır.



Görsel 7. Referans ve öğrenci performans kayıtlarının frekans serileri işlenmeden önce (üst Görsel) ve eşlendikten sonraki hali (alt Görsel). Bu örnekte öğrenci en düşük notu almıştır

Görsel 6 incelendiğinde, eşlenme işlemi sonrası tam not alan örneğin, frekans serisinin referans frekans serisiyle büyük örtüşmeye sahip olduğu görülmektedir. Görsel 7'deki düşük not alan performansın frekans serisinin referans frekans serisiyle örtüşme düzeyi ise daha düşüktür. Eşleme adımını takiben eşlenmiş seriler arasındaki farkın istatistiksel dağılımı hesaplanıp makine öğrenmesi modeline girdi olarak verilmektedir. Farkın istatistiksel dağılımının hesaplaması işlemi bir sonraki bölümde açıklanmaktadır.

Fark Öznitelikleri Çıkartma/Hesaplama

Makine öğrenmesi modeline girdi olarak kullanılan öznitelikler şunlardır; eşleme sırasındaki hesaplanan maliyet değeri (Eşitlik-1'de hesaplanan değer), eşleme sırasında her bir serinin uzunluklarının değişim oranı ve eşleştirme sonucu elde edilen serilerin farkından ölçülen istatistiksel dağılım değerleri. Bu değerler hem frekans serileri için, hem de kroma matrisleri için hesaplanmaktadır.

Elde edilen zamanda eşlenmiş serileri arasındaki fark L1 uzaklığı ile hesaplanarak bir fark serisi elde edildikten sonra şu istatistiksel parametreler hesaplanır: Ortalama fark, farkın standart sapması, serinin son %10 luk dilimindeki ortalama fark (bu öznitelik, icranın finalde doğru kapanış yapmasının notlamayı doğrudan etkilediği düşünülerek eklenmiştir), sınır değerleri {0, 25, 50, 75, 100, 125, 150, 200, 1200} sent olan 8 boyutlu bir histogram vektörü. Frekans serileri ve kroma serileri üzerinden bu öznitelikler yan yana dizilerek 28 boyutlu öznitelik vektörü elde edilmektedir. Bu vektör, etiket verisiyle birleştirilerek, her satırında bir örneğer dair özniteliklerin ve etiket değerinin bulunduğu standart tablo türü makine öğrenmesi verisi olarak kaydedilmektedir.

Makine Öğrenmesi Modeli

Tablo şeklinde veriler üzerinden gerçekleştirilen makine öğrenmesi uygulamalarında karar ağaçları temelli modellerin güncel derin öğrenme modelleri ile karşılaştırıldığında bile tercih edilir olduğu bilinmektedir (Shwartz-Ziv ve Armon, 2022). Bu sebeple, fark özniteliklerini girdi olarak alıp icraya verilecek notu tahmin eden model için karar ağaçları temelli modeller ve diğer sınıflandırıcı modellerini temsil için birer model seçilmiştir. Tüm testler paylaşılan açık kaynak kodlu araçlar sayesinde tekrarlanabilir formdadır. Yeni modellerin testlere dahil edilmesi birkaç kod satırı değişimi ile gerçekleştirilebilir veya PyCaret (https://pycaret.org/) türü araçlarla çok fazla sayıda model çeşitli hiper parametre kombinasyonları ile test edilip daha ideal bir kombinasyonun bulunması mümkündür. Bu metinde, yer darlığı sebebiyle, temsili modeller kullanarak (ve modelleri varsayılan parametreleriyle kullanarak) sınırlandırılmış testlerin sonuçlarını sunuyoruz.

İcralara otomatik not verme işlemi hem otomatik sınıflandırma hem de otomatik regresyon problemi olarak ele alınabilir. Bu sebeple bu iki yaklaşımla iki ayrı test

gerçekleştirilmiş ve sonuçlar alt bölümlerde sunulmuştur. İki yaklaşımda elde edilen sonuçların karşılaştırılabilir olmasını sağlamak için aynı makine öğrenmesi performans metrikleri kullanılmış ve raporlanmıştır. Testlerde hem çapraz doğrulama hem de baştan ayrılan test kümesi üzerinde sonuç raporlama gerçekleştirilmiş ancak yer darlığı sebebiyle bu metinde sadece test kümesinde elde edilen sonuçlar sunulmuştur. Çapraz doğrulama deney sonuçlarına www.github.com[15] üzerinden erişilebilir.

Verilerin eğitim ve test kümelerine bölünmesi işlemi soru (ezgi örüntüsü) düzeyinde gerçekleştirilmiş, böylece makine öğrenmesi modellerinin soruları ezberleme ihtimali için önlem alınmıştır (diğer bir deyişle eğitim için kullanılan ezgi soruları ile test için kullanılan ezgi soruları farklıdır). Çapraz doğrulamada da doğrulama kümesi soru bazında ayrımla seçildiği için çapraz doğrulama sonuçlarının ortalaması ile test kümesi üzerinde raporlanan sonuçlar büyük oranda örtüştüğü gözlenmiştir. Uzmanlardan alınan beş etiket kümesinin kesişimi olarak ayrıca üç ve üçten fazla uzmanın aynı notu verdiği örnekler alınmış ve bu notu kullanarak "çoğunluk etiketi" isimli bir etiket kümesi oluşturulmuştur. Bu küme makine öğrenmesi testlerine altıncı küme olarak eklenmiştir.

Notlandırma işleminin otomatik sınıflandırma problemi olarak ele alınması

İlk olarak problem otomatik sınıflandırma problemi olarak ele alınmış, her bir etiket kümesi için ayrı ayrı yürütülen makine öğrenmesi testleri sonuçları Tablo 2'de sunulmuştur. Her bir testte, bir etiket kümesinin verileri eğitim ve test olarak ayrılmakta, ayrılan eğitim kümesiyle sistem eğitilmekte ve geride kalan test kümesi üzerinde sistemin performansı ölçülmektedir.

Tablo 2. Otomatik sınıflandırma testleri performans listesi

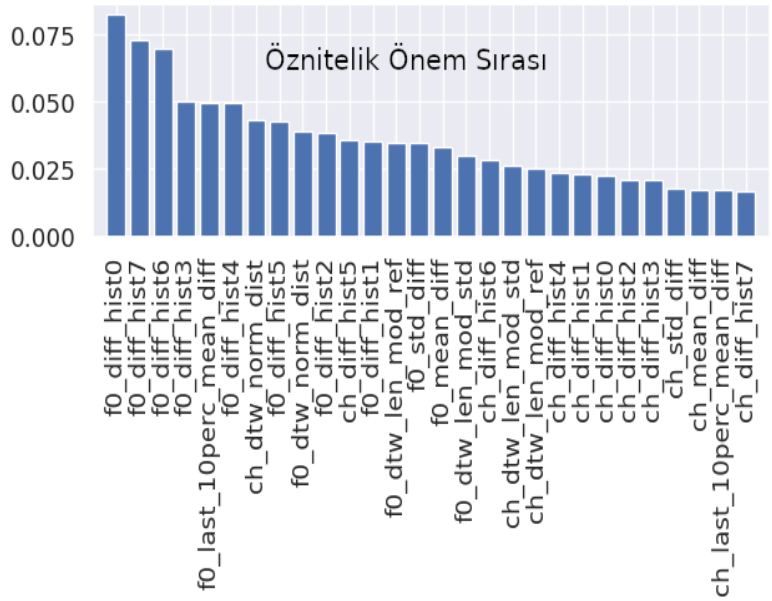
Veri	Model	OMH	Doğruluk	Ağırlıklandırılmış F1
Uzman1-ilk	LogisticRegression	0.47	0.61	0.59
	RandomForest	0.48	0.60	0.58
	XGBoost	0.50	0.58	0.55
	SVM	0.51	0.56	0.55
Uzman1-son	LogisticRegression	0.49	0.60	0.58

[15] Hakem değerlendirmesi sürecinde içeriği anonimleştirmek için doğrudan link verilmemiştir. Makale kabul edilirse basıma hazır sürümde tam link yerleştirilecektir.

	RandomForest	0.55	0.56	0.52
	XGBoost	0.56	0.55	0.51
	SVM	0.57	0.55	0.52
Uzman3-ilk	LogisticRegression	0.60	0.53	0.51
	RandomForest	0.62	0.52	0.48
	XGBoost	0.63	0.51	0.47
	SVM	0.65	0.49	0.47
Uzman3-son	LogisticRegression	0.61	0.50	0.49
	RandomForest	0.65	0.50	0.43
	XGBoost	0.69	0.49	0.43
	SVM	0.66	0.50	0.46
Uzman2	LogisticRegression	0.56	0.54	0.53
	RandomForest	0.58	0.54	0.52
	XGBoost	0.63	0.50	0.48
	SVM	0.60	0.52	0.51
Çoğunluk-etiketi	LogisticRegression	0.63	0.50	0.49
	RandomForest	0.65	0.51	0.48
	XGBoost	0.65	0.50	0.46
	SVM	0.66	0.47	0.45

Tablo 2'deki değerler sınıflandırıcı bazında gruplandığında ve çeşitli etiket kümeleri üzerinde elde edilen değerin ortalaması ele alındığında, ortalama 0.56 OMH değeri (OMH standart sapma: 0.07) ile en iyi performansı Logistic Regression modelinin sağladığı görülmektedir.

Makine öğrenmesi modeli olarak karar ağaçları kullanıldığı durumda girdi olarak kullanılan öznitelikler için bir sınıflandırmada etkinlikleri açısından önem sıralamasını otomatik olarak elde etmek mümkün olmaktadır. Görsel 8'de bu Görselde elde edilen öznitelik önem sırası sunulmaktadır.



Görsel 8. RandomForest modeli kullanılarak[16] hesaplanan öznitelik önem sırası

Görsel 8 incelendiğinde, 28 boyutlu öznitelik vektöründe sınıflandırmaya en fazla etkisi olan özniteliklerin frekans serilerinin farkından elde edilen histogram değerleri olduğu görülmektedir. Genel olarak kroma özniteliklerinin katkısı, frekans seri özniteliklerinin katkılarından düşüktür.

Notlandırma işleminin otomatik regresyon problemi olarak ele alınması

İkinci olarak problem otomatik regresyon problemi olarak ele alınmış, her bir etiket kümesi için ayrıca yürütülen makine öğrenmesi testleri sonuçlar Tablo 3'te sunulmuştur. Regresyon işlemi, tahmin edilen not değerini sürekli bir değer olarak ele aldığı için, 3.2 gibi ara değerler de sunar. Doğruluk ve Ağırlıklandırılmış F1 metriklerini hesaplamak için bu değer en yakın tam sayıya yuvarlanarak tam sayıya yuvarlanmış nota değerleri elde edilmiş, bu değerlerle ayrıca "OMH-Yuvarlanan Değerler" de hesaplanıp sunulmuştur. Yuvarlama işleminin OMH değerini bir miktar iyileştirdiği (düşürdüğü) gözlenmiştir.

[16] Hesaplama yöntemi için bakınız:
https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Tablo 3. Otomatik sınıflandırma testleri performans listesi

Veri	Model	OMH	OMH-Yuvarlanan Değerler	Doğruluk	Ağırlıklandırılmış F1
Uzman1-ilk	Linear regression	0.61	0.60	0.44	0.43
	Random forest	0.51	0.44	0.60	0.61
	XGBoost	0.52	0.45	0.59	0.60
	Adaboost	0.67	0.66	0.38	0.30
Uzman1-son	Linear regression	0.64	0.62	0.41	0.40
	Random forest	0.53	0.49	0.56	0.57
	XGBoost	0.55	0.50	0.55	0.56
	Adaboost	0.69	0.69	0.32	0.25
Uzman3-ilk	Linear regression	0.65	0.61	0.45	0.42
	Random forest	0.61	0.57	0.50	0.51
	XGBoost	0.63	0.59	0.49	0.50
	Adaboost	0.68	0.64	0.42	0.36
Uzman3-son	Linear regression	0.68	0.65	0.42	0.36
	Random forest	0.58	0.55	0.52	0.53
	XGBoost	0.61	0.57	0.50	0.50
	Adaboost	0.70	0.67	0.38	0.30
Uzman2	Linear regression	0.68	0.65	0.41	0.37
	Random forest	0.60	0.57	0.48	0.49
	XGBoost	0.61	0.56	0.51	0.51
	Adaboost	0.67	0.67	0.37	0.29
Çoğunluk-etiketi	Linear regression	0.69	0.67	0.39	0.35
	Random forest	0.62	0.59	0.49	0.51
	XGBoost	0.64	0.60	0.48	0.49
	Adaboost	0.73	0.70	0.34	0.26

Tablo 3'teki değerler model bazında gruplandığında, yuvarlanmış not değerleri ile yapılan ölçümler ve çeşitli etiket kümeleri üzerinde elde edilen değerin ortalaması ele alındığında, ortalama 0.54 OMH değeri (OMH standart sapma: 0.06) ile en iyi performansın Random Forest modeli olduğu görülmektedir.

Tartışma ve Sonuçlar

Bu çalışmada, piyano sesiyle kaydedilmiş bir ezginin insan sesiyle icrası/tekrarı sonrasında iki kaydı girdi olarak alarak otomatik notlandırma işlemi yapan bir sistem sunulmuştur. Tanımlanan problemde sinyal işleme açısından şu iki temel zorluk bulunmaktadır. Birinci zorluk birbirine karakter olarak benzemeyen iki öğenin karşılaştırılıyor oluşudur: referans kayıt (piyano sesi) ile performans kaydı (insan sesi), entonasyon özellikleri açısından karakter olarak oldukça farklıdır. Görsel 5'te örneği sunulduğu gibi, piyano kayıtları frekans serileri yan yana dizilmiş düzlükler Görselde bir karaktere sahip iken (Görsel 5 sol kolon), insan sesi frekans eğrileri net düzlükler içermeden sürekli yayvan geçişler içeren eğriler şeklindedir (Görsel 5 sağ kolon). İkinci zorluk da icranın serbest tempoda ve serbest uzunlukta gerçekleştiriliyor olması, bu sebeple, iki kaydın karşılaştırılabilir hale getirilmesi için önce zamanda bölgesel olarak esnetme-daraltma işlemlerinin tam otomatik yapılmasını gerekli kılmasıdır. DTW algoritmaları iki seriyi birbirine eşleme konusunda oldukça yaygın ve başarıyla kullanılan algoritmalardır. Ancak eldeki problemde, sistemin hatalı olarak notlandırdığı örnekler özellikle incelendiğinde, DTW işleminin zamanda eşleme sırasında gereğinden fazla iyi eşleştirme yaparak öğrencinin hatalarını temizlediği gözlenmiştir. Bu problemin aşılması için DTW algoritmasına ek kriterler eklenerek çok sayıda deneyler yapılmış ama bir iyileştirme gözlenemediği için bu metinde bu konudaki deneylere yer verilmemiştir. Bu işlem (zamanda eşlemeyi sağlayacak ama bunu sınırlandırarak, hataları yok edecek düzeyde esnetme-daraltmaların önüne geçecek kriterleri tanımlama), sistemin iyileştirme potansiyelinin olduğu bir boyuttur.

Önerilen sistem çeşitli makine öğrenmesi modelleri kullanılarak test edilmiştir. Sunulan test kombinasyonunda en iyi değerler, sınıflandırma problemi olarak ele alındığı durumda 0.56 OMH, regresyon problemi olarak ele alındığında 0.54 OMH olarak ölçülmüştür. Aynı metrik (OMH) kullanılarak uzmanların etiketleri birbiriyle ikili olarak karşılaştırıldığında ortalama 0.40 OMH (standart sapma: 0,04) değeri hesaplanmaktadır. Notların/etiketlerin 1-4 arasında olduğu düşünüldüğünde, önerilen sistemin uzmanlar arası (ve aynı uzmanın farklı zamanlarda yaptığı notlandırmalar arası) uyum düzeyine yaklaşan bir performansa sahip olduğu görülmektedir.

Bu deneylere ek olarak, literatür özetinde bahsi geçen derin öğrenme modelleri ile denemeler yapılmış ancak elde edilen başarı oranlarının metinde sunulan sistemden daha düşük olduğu gözlenmiştir. Bunun altında yatan temel sebebin veri kümesi büyüklüğü olduğu düşünülmektedir çünkü derin ağ modelleriyle yapılan testlerin çoğunda öğrenmenin neredeyse hiç gerçekleşmediği gözlemlenmiştir. Metnin bütünlüğü ve uzunluğu gözetilerek derin öğrenme modelleriyle yapılan (ve daha düşük skor elde edilen) bu deneylere metin içerisinde yer verilmemiştir.

Tasarlanan sistem, kayıtlarda sadece ezgi boyutunu işlemeyi hedefler. Burada tartışmaya açık noktalardan birisi uzmanların not verirken ezgideki notaların temel frekansı dışındaki faktörlerden etkilenip etkilenmediği bilinmemesidir. Bu konu hakkında yapılan son dönem çalışmalarında temel frekans dışında perde zarfı, tını ve oktav farkları gibi değişkenlerin de algılanan perdeyi etkilediği saptanmıştır (Köker, 2017; Oxenham, 2012; Zarate, Ritson ve Poeppel, 2013). Bu tip faktörler icra sırasındaki entonasyon farklarının bazen kabul edilebilir, bazen de hata olarak algılanmasına sebep olabilmektedir - hele ki bu entonasyon açısından mükemmele yakın bir icranın ardından gelen bir icra ise oluşturacağı tezatlık etkisinden dolayı hata olarak algılanmaya daha müsaittir. Bu tip durumlar, uzmanların farklı zamanlarda farklı sıralamalarda puanladıkları icracılar için "entonasyon" (-5puan) ve "yanlış okunan nota" (-15puan) değerlendirmelerinde karar değişikliklerine neden olmuş olabilir. Bu çalışmada da uzman notları karşılaştırıldığında, ara notlar (2 ve 3) dikkate alındığında uzmanlar arası (ve aynı uzmanın çeşitli zamanlarda verdiği notlar arasında) uyumun oldukça düşük olduğu görülmüştür. Bir başka deyişle, "Başarılı" ve "Çok Başarısız" performansların belirlenmesinde güçlü fikir birliği vardır, bunların dışındaki icralardaki hataların ne kadar "ciddi hata", ne kadar "küçük hata" olarak değerlendirilmesi yukarıda bahsedilen başka faktörlerden etkisi altında da gerçekleşmiş olabilir. Önerilen sistem, bir uzmanın verdiği notlar üzerinden notlamayı öğrenmektedir (ve yine o uzmanın notlandırdığı diğer örnekler üzerinde test edilmiştir). Eldeki etiketlerin/notların sorgulamaya açık olduğu durumda makine öğrenmesi testleriyle elde edilen sonuçların yorumlanması da zorlaşmaktadır. Bununla beraber, sistem performansı ile uzmanlar arası uyum düzeyinin yakın oluşu sistemin gerçek hayat uygulamalarında başarıyla kullanılma potansiyeli olduğunu düşündürmektedir. Otomatik sistemin kullanıldığı durumda uzman zaman maliyeti olmayacağı için öğrenciye

çok daha fazla soru sorulup bu işlem sonucunda daha güvenli bir değerlendirme notu elde edilebilir.

Teşekkür

Bu çalışma, TÜBİTAK tarafından '1001 - Bilimsel ve Teknolojik Araştırma Projelerini Destekleme Programı' kapsamında desteklenmiştir (Proje no:121E198).

Kaynakça

Abeßer, J., Hasselhorn, J., Dittmar, C., Lehmann, A., ve Grollmisch, S. (2013, October). Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils. *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)* (pp. 975-988).

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., Boyer, H., Mayor, O., ve Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. *Proceedings of the International Society for Music Information Retrieval (ISMIR), Curitiba, Brazil.*(pp.493-498).

Bozkurt, B., Baysal, O., ve Yüret, D. (2017, September). A dataset and baseline system for singing voice assessment. *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR), Matosinhos, Portugal* (pp. 25-28).

Bozkurt, B., Gulati, S., Romani Picas, O., ve Serra, X. (2018). MusicCritic: a technological framework to support online music teaching for large audiences. *Proceedings of the 33rd World Conference on Music Education (ISME), Baku, Azerbaijan.*

Chan, T. S., Yeh, T. C., Fan, Z. C., Chen, H. W., Su, L., Yang, Y. H., ve Jang, R. (2015, April). Vocal activity informed singing voice separation with the iKala dataset. *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 718-722).

Carmi-Cohen, D. (1964). An investigation into the tonal structure of the maqamat. *Journal of the International Folk Music Council*, *16*, 102-106.

Ewert, S. (2011). Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. *Proceedings of the Conference of the International Society for Music Information Retrieval (ISMIR)*.

Galan, D., Heradio, R., Vargas, H., Abad, I., ve Cerrada, J. A. (2019). Automated assessment of computer programming practices: The 8-years uned experience. *IEEE Access*, *7*, 130113-130119.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software*, *31*, 1-24.

Goto, M., ve Nishimura, T. (2005). AIST humming database: Music database for singing research. *IPSJ SIG Notes (Technical Report)(Japanese edition)*, *2005*(82), 7-12.

Gulati, S., Serra, J., ve Serra, X. (2015, April). An evaluation of methodologies for melodic similarity in audio recordings of Indian art music. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 678-682).

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Huang, J., ve Lerch, A. (2019). Automatic Assessment of Sight-reading Exercises. *Proceedings of the Conference of the International Society for Music Information Retrieval (ISMIR)*. (pp. 581-587).

Jha, M. V., ve Rao, P. (2012, February). Assessing vowel quality for singing evaluation. *Proceedings of the National Conference on Communications (NCC)* (pp. 1-5).

Kim, J. W., Salamon, J., Li, P., ve Bello, J. P. (2018, April). Crepe: A convolutional representation for pitch estimation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 161-165).

Kirchhoff, H., ve Lerch, A. (2011). Evaluation of features for audio-to-audio alignment. *Journal of New Music Research*, *40*(1), 27-41.

Köker, O. (2017). *The Acceptable Pitch Range (s) For Single Note Repetitions In Music Aptitude Examinations* (Masters Thesis). İstanbul Teknik Üniversitesi, Graduate School.

Lin, C. H., Lee, Y. S., Chen, M. Y., ve Wang, J. C. (2014, September). Automatic singing evaluating system based on acoustic features and rhythm. *Proceedings of the International Conference on Orange Technologies* (pp. 165-168). https://doi.org/10.1109/ICOT.2014.6956625

Lundy, D. S., Roy, S., Casiano, R. R., Xue, J. W., ve Evans, J. (2000). Acoustic analysis of the singing and speaking voice in singing students. *Journal of voice*, *14*(4), 490-493.

Mauch, M., ve Dixon, S. (2014, May). pYIN: A fundamental frequency estimator using probabilistic threshold distributions. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 659-663). https://doi.org/10.1109/ICASSP.2014.6853678

McVicar, M., Santos-Rodríguez, R., Ni, Y., ve De Bie, T. (2014). Automatic chord estimation from audio: A review of the state of the art. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(2), 556-575. https://doi.org/10.1109/TASLP.2013.2294580

Molina, E., Barbancho, I., Gómez, E., Barbancho, A. M., ve Tardón, L. J. (2013, May). Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 744-748). https://doi.org/10.1109/ICASSP.2013.6637747

Müller, M., ve Zalkow, F. (2021). libfmp: A Python package for fundamentals of music processing. *Journal of Open Source Software*, *6*(63), 3326. https://doi.org/10.21105/joss.03326

Nakano, T., Goto, M., ve Hiraga, Y. (2006). An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. *Proceedings of the Ninth International Conference on Spoken Language Processing*.

Oxenham, A. J. (2012). Pitch perception. *Journal of Neuroscience*, *32*(39), 13335-13338. https://doi.org/10.1523/JNEUROSCI.3815-12.2012

Pati, K. A., Gururani, S., ve Lerch, A. (2018). Assessment of student music performances using deep neural networks. *Applied Sciences*, *8*(4), 507. https://doi.org/10.3390/app8040507

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., ve Ellis, D. P. W. (2014, October). MIR_EVAL: A Transparent Implementation of Common MIR Metrics. *Proceedings of the Conference of the International Society for Music Information Retrieval (ISMIR)*.

Salamon, J., ve Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE transactions on audio, speech, and language processing*, *20*(6), 1759-1770. https://doi.org/10.1109/TASL.2012.2188515

Schramm, R., de Souza Nunes, H., ve Jung, C. R. (2015, October). Automatic Solfège Assessment. *Proceedings of the Conference of the International Society for Music Information Retrieval (ISMIR)*. (pp. 183-189).

Seshadri, P., ve Lerch, A. (2021). Improving music performance assessment with contrastive learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.2108.01711

Shwartz-Ziv, R., ve Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84-90. https://doi.org/10.48550/arXiv.2106.03253

Sundberg, J. (2001). Level and center frequency of the singer's formant. *Journal of voice*, *15*(2), 176-186. https://doi.org/10.1016/S0892-1997(01)00019-4

Tsai, W. H., Ma, C. H., ve Hsu, Y. P. (2015). Automatic Singing Performance Evaluation Using Accompanied Vocals as Reference Bases. *J. Inf. Sci. Eng.*, *31*(3), 821-838.

Typke, R. (2007). *Music retrieval based on melodic similarity* (Doctoral dissertation, Utrecht University). https://dspace.library.uu.nl/handle/1874/19776

Wesolowski, B. C., Wind, S. A., ve Engelhard Jr, G. (2016). Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted Rasch partial credit model. *Music Perception: An Interdisciplinary Journal*, *33*(5), 662-678. https://doi.org/10.1525/mp.2016.33.5.662

Yang, W., Wang, X., Tian, B., Xu, W., ve Cheng, W. (2022). A Multi-stage Automatic Evaluation System for Sight-singing. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2022.3168132

Zarate, J. M., Ritson, C. R., ve Poeppel, D. (2013). The effect of instrumental timbre on interval discrimination. *PloS one*, *8*(9), e75410. https://doi.org/10.1371/journal.pone.0075410

Zhang, H., Jiang, Y., Jiang, T., ve Peng, H. (2021). Learn by Referencing: Towards Deep Metric Learning for Singing Assessment. *Proceedings of the Conference of the International Society for Music Information Retrieval (ISMIR)*. (pp. 825-832).

Zhang, Y., ve Yi, D. (2021). A new music teaching mode based on computer automatic matching technology. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(16), 117-130.

Extended Abstract

Automatic Assessment Of Student Vocal İmitation Performances Of Melodic Patterns

Online music education has witnessed remarkable growth, with renowned institutions offering online degree programs to meet the increasing demand for remote learning. As student enrollment continues to rise, incorporating technology becomes imperative to optimize limited human resources and provide effective feedback to a large number of students. Automatic grading and feedback systems play a crucial role in maintaining student engagement and progress. This study focuses on the automatic assessment of vocal performances by measuring the distance between a target melody and a student's rendition, utilizing supervised learning methods and a dataset of expert-rated recordings.

The study specifically targets exercises involving the repetition of simple musical phrases, which fall under the category of relatively mechanical musical exercises. The proposed system employs machine learning techniques to predict grades based on feature differences between time-aligned recordings.

The literature on automatic vocal performance grading primarily emphasizes the evaluation of pitch accuracy, using the extracted fundamental frequency (f0) series from monophonic recordings. Among various approaches, Dynamic Time Warping (DTW)-based distance measures have demonstrated superior performance in measuring melodic similarity. Previous studies have explored statistical analysis, automatic transcription, and deep learning models to predict grades. However, deep learning models often encounter limitations due to the size of available datasets. To contribute to the field, this study presents an open-source system for vocal performance grading, incorporating pitch series and chroma features. The system

generates grades that closely align with expert consensus, utilizing a blind and randomized grading process for reliable comparative analysis.

The proposed system includes four blocks. The first block computes the fundamental frequency series and chromagram of a given pair of recordings (a reference recording and a student performance recording). The second block applies alignment in time using a dynamic time warping algorithm. The third block computes distances between aligned features and their statistics to form a feature vector. The final block is a machine learning model that takes in the feature vector and estimates a grade (in range 1-4 referring to perfect performance, good performance with minor errors, performance with major errors and completely off) for the student performance. This system has been trained and tested on audio data collected during conservatory entrance examinations with 5 individual sets of expert annotations (running tests separately for each annotation set). The paper presents a study of inter expert grades comparison as well as comparison of grades assigned by the same expert to the same performances at different times.

The proposed system exhibits promising performance, achieving accuracies of 0.56 OMH for classification tasks and 0.54 OMH for regression tasks. The agreement among expert ratings, measured using the OMH metric, averages at 0.40 OMH, indicating a level of agreement similar to that observed among experts. Although deep learning models were explored, they yielded lower success rates primarily due to limitations imposed by the dataset size.

Factors beyond the fundamental frequency, such as pitch envelope, timbre, and octave differences, were found to be influencing pitch recognition in previous studies. Similar factors may have influenced the expert evaluations in this study, as the agreement among experts was lower for intermediate grades, implying subjective evaluations of error severity. Nevertheless, the proposed system aligns with expert consensus, laying a solid foundation for further system development.

In conclusion, this study introduces an automated system for scoring vocal performances using piano-to-voice recordings in the context of online music education. The system effectively addresses challenges associated with the dissimilarity between recordings and the need for precise time alignment and segmentation. Its performance is promising compared to expert ratings, highlighting its potential for real-life applications. The incorporation of factors beyond the fundamental frequency improves the system's alignment with expert consensus, enhancing its reliability. The automated system offers a cost-effective solution, significantly augmenting the evaluation process in music education and performance training. The provision of a shared dataset and open-source code promotes reproducibility and facilitates comparative analysis.

A key challenge in this task is the alignment of recordings, performing time-stretching and time-compression operations to facilitate comparisons. While dynamic time warping based approaches are used in many similar alignment tasks successfully, for this specific task, too good alignment applying extreme stretching to student performance results in performance reduction for automatic assessment (assigning a high grade to a low quality performance). One potential direction of research is to study this phenomenon and use additional criteria in alignment to avoid extreme modifications in student performance.

Future research may focus on expanding the dataset size and exploring additional dimensions of musical performance evaluation, further enriching the field of automatic scoring systems in music education.