Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 7(2) (2024) 102-115
DOI: 10.38016/jista.1400047

Olay Kamerası ile Verimli Konuşma Sesi Tespiti için Zamansal Evrişimsel Ağlar

Arman Savran[1*]

[1*] Bilgisayar Mühendisliği Bölümü, Yaşar Üniversitesi, İzmir, Türkiye

arman.savran@yasar.edu.tr

Öz

Konuşma sesi tespiti (KST), insan bilgisayar arayüzleri için yaygın olarak kullanılan gerekli bir ön-işlemedir. Karmaşık akustik arka plan gürültülerinin varlığı, büyük derin sinir ağlarının ağır hesaplama yükü pahasına kullanımlarını gerekli kılmaktadır. Görü yoluyla KST ise, arka plan gürültüsü problemi olmadığından, tercih edilebilen alternatif bir yaklaşımdır. Görü kanalı, ses verisine erişimin mümkün olmadığı durumlarda ise zaten tek seçenektir. Ancak, genelde uzun süreler aralıksız çalışması beklenen görsel KST, video kamerası donanım ve video verisi işleme gereksinimlerinden dolayı önemli enerji sarfiyatına sebep olur. Bu çalışmada, görü yoluyla KST için, nöromorfik teknoloji sayesinde verimliliği geleneksel video kameradan oldukça yüksek olan olay kamerasının kullanımı incelenmiştir. Olay kamerasının yüksek zaman çözünürlüklerinde algılama yapması sayesinde, uzamsal boyut tamamen indirgenerek sadece zaman boyutundaki örüntülerin öğrenilmesine dayanan son derece hafif fakat başarılı modeller tasarlanmıştır. Tasarımlar, zamansal alıcı alan genişlikleri gözetilerek, farklı evrişim genleştirme tiplerinin, aşağı-örnekleme yöntemlerinin ve evrişim ayırma tekniklerinin bileşimleri ile yapılır. Deneylerde, KST'nin çeşitli yüz eylemleri karşısındaki dayanıklıkları ölçülmüştür. Sonuçlar, aşağı-örneklemenin yüksek başarım ve verimlilik için gerekli olduğunu ve bunun için, maksimum-havuzlamanın adımlı evrişim yöntemiyle aşağı-örnekleme yapmaktan daha üstün başarım elde ettiğini göstermektedir. Bu şekilde üstün başarımlı standart tasarım 1.57 milyon kayan nokta işlemle (MFLOPS) çalışır. Evrişim genleştirmesinin sabit bir faktörle yapılıp aşağı-alt örnekleme ile birleştirilmesiyle de, benzer başarımla, işlem gereksiniminin yarıdan fazla azaldığı bulunmuştur. Ayrıca, derinlemesine ayrışım da uygulanarak işlem gereksinimi 0.30 MFLOPS'a, yani standart modelin beşte birinden daha aşağısına indirilmiştir.

Anahtar kelimeler: Konuşma Sesi Tespiti, Olay Kamerası, Verimli, Görsel Konuşma, Genleştirilmiş Evrişim, Ayrılabilir Evrişim

Temporal Convolutional Networks for Efficient Voice Activity Detection with Event Camera

Abstract

Voice activity detection (VAD) is a widely used essential pre-processing for human-computer interfaces. The presence of complex acoustic background noise requires the use of large deep neural networks at the expense of heavy computational load. Visual VAD is a preferable alternative approach since there is no background noise problem. Also, the video channel is the only option when access to audio data is impossible. However, visual VAD, which is generally expected to operate continuously for long periods of time, causes significant energy consumption due to the requirements of video camera hardware and video data processing. In this study, the use of the event camera, whose efficiency is much higher than the traditional video camera thanks to neuromorphic technology, was examined for VAD through vision. Thanks to the event camera's detection at high time resolutions, the spatial dimension is completely reduced and extremely lightweight but successful models that work only in the time dimension have been designed. Designs are made with combinations of different types of dilated convolution, down-sampling methods, and separable convolution techniques, taking into account temporal receptive field sizes. In the experiments, the robustness of VAD against various facial actions was measured. The results show that down-sampling is necessary for high performance and efficiency, and for this, max-pooling achieves superior performance than down-sampling with stepwise convolution. This high-performance standard design operates at 1.57 million floating point operations (MFLOPS). By performing dilated convolution with a constant factor and combining it with down-subsampling, it was found that the processing requirement was reduced by more than half, with similar performance. Additionally, by also applying depthwise separation, the processing requirement was reduced to 0.30 MFLOPS, less than one-fifth of the standard model.

Keywords: Voice Activity Detection, Event Camera, Efficient, Visual Speech, Dilated Convolution, Separable Convolution

* Sorumlu yazar.
E-posta adresi: arman.savran@yasar.edu.tr

Gönderme : 4 Aralık 2023
Revizyon : 8 Nisan 2024
Kabul : 18 Nisan 2024

1. Giriş (Introduction)

Konuşma sesi tespiti (KST), işitsel veya görsel konuşma tabanlı arayüzler ve sahne analizi için önemli bir ön-işlemedir. Ses-temelli KST, konuşma tanıma, konuşmacı tanıma, konuşma sesi iyileştirme, konuşmacı günlüğü çıkarma, komut-kontrol gibi uygulamalarda (Wang vd., 2023, Korkmaz ve Boyacı, 2023, Zhang vd., 2016, Çubukçu vd., 2015) yaygın kullanılsa da bazı sınırlamaları vardır. Arka plan gürültüsü başlıca bir güçlüktür. Büyük derin sinir ağları (DSA'lar) çeşitli zorlu gürültü koşullarında etkili çözüm sağlayabilse de ağır hesaplama yükü ve enerji tüketimine neden olurlar (Zhang vd., 2016). KST birçok uygulamada ön-işleme olarak çalışması gerektiğinden yüksek verimlilik önem arz eder. Öte yandan, görü yoluyla KST dudak hareketlerinden konuşmayı saptarken akustik gürültünün hiçbir etkisi yoktur. Teknik nedenler veya gizlilik ihtiyacından dolayı işitsel kanalın bulunmaması ise, işitsel VAD için bir kısıtlamadır; bu durumlarda tek seçenek görü yoluyla tespit olabilir. Bahsedilen nedenlerden dolayı, görsel-işitsel veya yalnızca görsel kipte çalışan KST yöntemleri üzerine araştırmalar yapılmaktadır (Ariav vd., 2018, Guy vd., 2020).

Bu araştırmalar video kameralar kullanarak oldukça başarılı sonuçlar elde edebilmişlerdir. Ancak, genelde uzun süreler aralıksız çalışması beklenen görsel KST, video kamerası donanım ve video verisi işleme gereksinimlerinden dolayı önemli enerji sarfiyatına sebep olur. Nöromorfik mühendislik ilkeleriyle geliştirilen ve yeni ortaya çıkan olay kameralarının, geleneksel video kameralara göre üstün olduğu yönleri, KST için daha başarılı ve daha verimli çözümlerin geliştirilmesini sağlayabilir. Olay kamerasının, robotik uygulamalarda, uçan gözlerde (drone), otonom araçlarda veya mobil cihazlarda geleneksel kameralara göre avantajlı bir alternatif olduğu veya video kamerasını tamamlayıcı bir görü kanalı olduğu zaten birçok defa gösterilmiştir (Gallego vd., 2022). Olay kamerasının başarısı, basit bir deyişle, "akıllı piksel" temelli bir algılama tekniğinden kaynaklanmaktadır. "Akıllı piksel" tabiri, üzerine düşen ışık şiddetinin değişimini, diğer piksellere bağlı olmadan kendi başına saptayabilen pikseli ifade eder. Geleneksel kamera piksellerinde böyle bir mekanizma yoktur, sadece ışık yoğunluğu sayısal olarak örneklenir. Bu saptama olayı, bir piksel-olayı olarak kodlanarak arabirim üzerinden aktarılır. Olay kamerasına özgü bu asenkron tetiklenen piksel-olayları sayesinde çok yüksek zamansal çözünürlük, düşük gecikme, düşük güç gereksinimi ve yüksek dinamik aralık gibi önemli avantajlar elde edilir. Bu avantajlar, olay kamerası temelli KST ön-işlemesi sayesinde, daha verimli ve başarımlı yeni nesil görsel veya görsel-işitsel uygulamaların geliştirilmesinin önünü açabilir.

Görsel KST problemi ile yakından ilgili olarak, insan yüzü işleme alanında olay kamerası ile yapılan çalışmalar mevcuttur. Örneğin, olay kamerası kullanılarak gerçekleştirilen otomatik dudak okumanın, geleneksel kamera başarımını geride bırakabileceği gösterilmiştir (Tan vd., 2022). Bu sonuç, dinamik örüntülerin yüksek zaman çözünürlüğü ile algılanması sayesinde gerekli bilginin korunması argümanı ile açıklanabilir. Örneğin, gerçek hayattaki böyle bir uygulamada, bir KST ön-işleme biriminin devreye girmesiyle, gereksiz yere işlemci meşgul eden ve enerji harcayan DSA hesaplamalarının yapılmasının önüne geçilebilir ve konuşma dışındaki görsel aktivitelerin yanlış dudak okumaya neden olması önlenebilir. Başka bir uygulama olarak, örneğin, konuşma dinamikleri öğrenilmesi yoluyla kişi tanıma ele alınabilir. Olay kamerasının konuşma dinamiği temelli kişi tanımadaki avantajları gösterilmiştir (Moreira vd., 2022). Konuşmacı sesi iyileştirme (Arriandiaga vd., 2021) gibi çalışmalarla örnek uygulamalar çoğaltılabilir. KST, benzeri bütün uygulamalarda gerekli bir ön-işlemedir ve ayrıca, akustik ses işleme gibi olay kamerasının kullanılmadığı uygulamalarda da aktivasyon ön-işlemesi olarak görev alabilir.

Bu çalışma, bu tür sistemleri olanaklı kılabilmek için, kaynak gereksinimi çok düşük seviyede olan ve buna rağmen etkili olarak çalışılabilen yöntemlerin geliştirilmesi üzerinedir. Görü verileri, sahnedeki aktiviteye uyumlu olarak yüksek zaman çözünürlüklerinde seyrek yapıda oluştuğundan, sadece zaman boyutundaki değişim örüntülerinin öğrenilmesi yoluyla KST geliştirilmesi hedeflenmiştir. Burada ana fikir, uzamsal boyutu tamamen indirgeyerek işlem yükünden büyük oranda tasarruf edebilmektir. Ağız bölgesindeki piksel-olaylarının uzamsal boyutu tamamen indirgenerek ve sadece zamansal eksende örüntü tanıma yaparak, çok düşük gereksinimli fakat yüksek başarımlı sınıflandırıcılar hedeflenmiştir. Konuşma dinamiği örüntü özniteliklerini öğrenme yoluyla çıkarmak için ise zamansal evrişimli ağlar uygulanmıştır. Bunun sebebi, öz yinelemeli ve dönüştürücü sinir ağlarına kıyasla çok daha verimli çalışmaları ve çok büyük olmayan veri kümelerinde en iyilemesi nispeten kolay olmasıdır (Bai vd. 2018). Farklı evrişim tekniklerini kullanan tasarımlar olay kamerası temelli KST problemi için sınanmıştır. Aşağı-örnekleme, genleştirme, derinlemesine ve gruplamalı ayrılabilir evrişim teknikleri kullanılarak ve zamansal alıcı alanları analiz edilerek tasarımlar yapılmıştır. Zaman ekseninde maksimum havuzlama yoluyla aşağı-örnekleme yapılarak çeşitli yüz eylemlerine karşı gürbüz çalışan ve yüksek başarımlara ulaşan bir standart tasarım önerilmiştir. Evrişim genleştirme ve

derinlemesine ayırma yapılarak da bu standart yönteme göre beş kattan fazla işlem kazancı sağlanmıştır.

Makalenin geri kalanında, önce Bölüm 2'de ilgili çalışmalara yer verilir. Bölüm 3'te, uzamsal alanı ağız bölgesinde indirgeyen piksel-olayı temsil modeli ve önerilen çeşitli zamansal evrişim teknikleri ve tasarımları anlatılır. Bölüm 4'te, verimlilik artırma tasarımlarının başarım ve işlem yükleri sunularak başarım ve verimlilik açılarından en iyi modeller saptanır ve konuşma dışındaki yüz dinamiklerine karşı dayanıklıkları ayrı ayrı ölçülür. Son olarak Bölüm 5'te, elde edilen bulgular özetlenerek ana sonuçlar verilir.

2. İlgili Çalışmalar **(Related Work)**

Geleneksel kameralarla yapılan önceki KST çalışmaları akustik ortam gürültüsüne karşı dayanıklılığı artırmak için, birçok defa görsel-işitsel çözümler önermişlerdir (Ariav vd., 2018, Ghaemmaghami vd., 2015). Ancak, işitsel veriler olmadığı durumlarda, KST için tek seçenek video verileri olabilir. Görü yoluyla KST için, Patrona vd. (2016) optik akış ve görüntü gradyanı tanımlayıcıları ile görsel-kelime-çantasına dayalı bir teknik önermiştir. Guy vd. (2020) özyinelemeli ağları doğrudan yüz nirengi noktalarına uygulamış ve optik akış örüntülerini evrişimli ağlar ile modellemişlerdir.

Yakın çekim yüz sahnelerinde yapılan KST çalışmalarının yanında, birden fazla insan vücudunu ve diğer ön plan nesnelerini içeren, arka plan karmaşıklığı fazla olan çok geniş açılı sahnelerde konuşma ile ilgili yüz ve vücut kısımlarını saptayan KST çalışmaları da vardır (Sharma vd., 2019, Shahid vd., 2021). Bu kapsamdaki bütünleşik mekansal lokalizasyon ve KST problemi daha zorlu olduğundan, çok daha karmaşık modellerin kullanılmasını gerektirir. Ek olarak bunların uygulama alanları bu makalenin hedeflerinden daha farklıdır. Dolayısıyla, böyle bütünleşik problemler bu çalışmanın kapsamı dışındadırlar.

Diğer taraftan, nöromorfik sensörler, kendilerini sahne aktivitesine uyarlayarak ve sıkıştırılmış seyrek algılama gerçekleştirerek çok yüksek zaman çözünürlüğü, enerji verimliliği ve yüksek dinamik aralığı avantajları sunarlar (Gallego vd., 2022). Geleneksel sensör, tüm piksellerdeki ışık yoğunluğunu eş zamanlı olarak örneklediğinden bu özelliklerden mahrumdur. Nöromorfik sensör ise, bir piksel üzerindeki ışık yoğunluğunda bir miktar değişiklik tespit ettiği anda, diğer piksellerle eş zamanlı olmayan bir piksel-olayı oluşturur. Bu yeni algılama teknolojisi birçok başarılı uygulamanın geliştirilmesini sağlamıştır. Örneğin, nesne sınıflandırma görevlerinde (Deng vd., 2022, Kim vd., 2022, Schaefer vd. 2022, Gehrig vd., 2019), el hareketi işaretlerinin tanınmasında (Amir vd., 2017), yürüme biçimi tanımada (Wang vd., 2019, Wang vd., 2022), nesne saptamada (Li vd., 2022, Schaefer vd., 2022, Perot vd., 2020) ve izlemede (Zhang vd., 2022), otonom sürüş için direksiyonu dönmesini tahmin etmek amacıyla (Maqueda vd., 2018), kamera poz takibi (Gallego vd., 2018) ve optik akış tahmini için (ParedesValles vd., 2021, Gehrig vd. 2019) kullanılmıştır. Ayrıca, olay kamerası verilerinden video geri-çatım işleminin başarılı bir şekilde yapılabileceği gösterilmiş (Zhu vd., 2022, ParedesValles vd., 2021, Rebecq vd., 2019) ve, standart videolarda hareket bulanıklığını gidermek için olay kamerası kullanımı (Tulyakov vd., 2022, Pan vd., 2019) önerilmiştir.

Konuşma artikülasyonlarının dinamik örüntülerinin taşıdığı bilgi seviyesi son derece yüksek olabildiğinden, olay kamerasının zamansal çözünürlük avantajından faydalanmayı ilke edinen çeşitli konuşma işleme araştırmaları yapılmıştır. Çalışmaların birçoğu, başarımı artırmak için işitsel ve görsel sinyallerin birleşimine odaklanmıştır. Neil vd. (2016) ve Li vd. (2019) görsel-işitsel konuşma tanıma için DSA'lar aracılığıyla görsel olay verilerini ses kipiyle birleştirmiş; Savran vd. (2018) bir görsel uzam-zamansal filtreyi işitsel DSA ile birleştirerek gürültülü akustik ortamlarda konuşma sesi algılayıcı başarım ve verimliliğini artırmış; Arriandiaga vd. (2021) konuşmacıyı ayırarak konuşma iyileştirme yapmak amacıyla yüz nirengi noktalarında optik akış kestirimi yapmıştır. Yakın zamanda, Tan vd. (2022) en güncel video dudak okuma yöntemlerini geride bırakan üstün dudak okuma başarımları elde etmiş ve, Savran (2023a) ses aktivitesi bulma için tamamen evrişimsel DSA önermiştir. Bunların dışında, göz kırpma saptama (Lenz vd., 2020, Ryan vd., 2021), yüz pozu hizalama (Savran ve Bartolozzi, 2020, Savran, 2023, Savran, 2023b), yüz bulma (Barua vd., 2016), kimlik tanıma (Moreira vd., 2022) ve ifade tanıma (Berlincioni vd., 2023) gibi görsel konuşmanın yanı sıra çeşitli olay kamerası temelli yüz işleme çalışmaları da mevcuttur.

3. Yöntem (Methodology)

Önerilen görsel KST, Şekil 1'de gösterilmektedir. Önce piksel-olayları verisi kullanılarak ağız bölgesi olay yoğunluğu kestirimi yapılır, böylece konuşma sesi ile ilgili görsel veriler çok düşük boyutlu ve DSA işlemlerine uygun bir forma indirgenir. Bu temsil biçimi Bölüm 3.1'de anlatılmaktadır. Sonra evrişimsel sinir ağı gövdesinde öznitelikler çıkarılır ve baş kısmında tahmin yapılır. DSA'nın gövde kısmı, Bölüm 3.2 ve Bölüm 3.3'te anlatılan teknikler uygulanarak modellenir. Bölüm 3.2'de, alıcı alanı büyütürken karmaşıklığı fazla artırmayan model mimarileri ve Bölüm 3.3'te de, işlem yükünü azaltan modellerler açıklanmaktadır. Ağ mimari tasarımı ise Bölüm 3.4'te yapılır.

3.1. Görsel Olay Yoğunluğu Çıkarımı (Visual Event Intensity Extraction)

Bir piksel-olayı, sensör pikselindeki logaritmik ışık yoğunluğu değişimi belli bir eşik değerini aştığı anda, diğer sensör piksellerinden bağımsız olarak, yani asenkron olarak, bir tetiklenme sonucu oluşur (Gallego vd., 2022). Piksel-olayı, değişimin pozitif veya negatif yönde olduğunu belirten ikili polarite değişkeni p, sensör düzlemindeki konum (x,y) ve, zaman etikteki t bilgilerini içerir. i indeksli piksel-olayı bir $e^i = (x^i, y^i, t^i, p^i)$ çok-öğelisi ile ifade edilir. Dudaklar ana konuşma artikülatörleri olduğundan, ses aktivitesinin dinamik görsel örüntülerini elde etmek için ağız bölgesindeki piksel olaylarının yoğunluğu zaman ekseni boyunca hesaplanır. Bu temsilin yalnızca iki kanalı vardır; biri pozitif, diğeri negatif değişim kutbu içindir. Böylece, ağız bölgesi üzerindeki uzamsal alan tamamen indirgenerek zaman ekseninde değişen sadece iki boyutlu bir gösterim elde edilir.

Ağız bölgesi, yüz nirengi noktaları kullanılarak çıkarılır. Ağız merkez noktası bu bölgenin merkezi olarak alınır. Dikdörtgen biçiminde bir referans koordinat sistemi üzerinde bölge şablonu tanımlanır. Bu çalışmada, dikdörtgen uzunluk-genişlik oranı ¾ olarak belirlenmiştir. Ancak, ağız bölgesinin iki boyutlu izdüşümü değişen yüz pozu nedeni ile biçim değiştirdiği için, şablonun değişen poza göre hizalanması gerekir. Hizalama için, göz ve ağız merkez noktalarını temel alan 2 boyutlu afin poz kestirimi yapılır (Savran ve Bartolozzi, 2020). Her an için değişen bu afin dönüşüm, dikdörtgen şablonu her seferinde o anki poza uygun bir dörtgene dönüştürür. Böylece, büyük pozlar altında dahi, konuşma eylemi ile en ilgili piksel-olaylarını kullanılması sağlanır. Şekil 2'de örnek ağız bölgeleri gösterilmektedir.

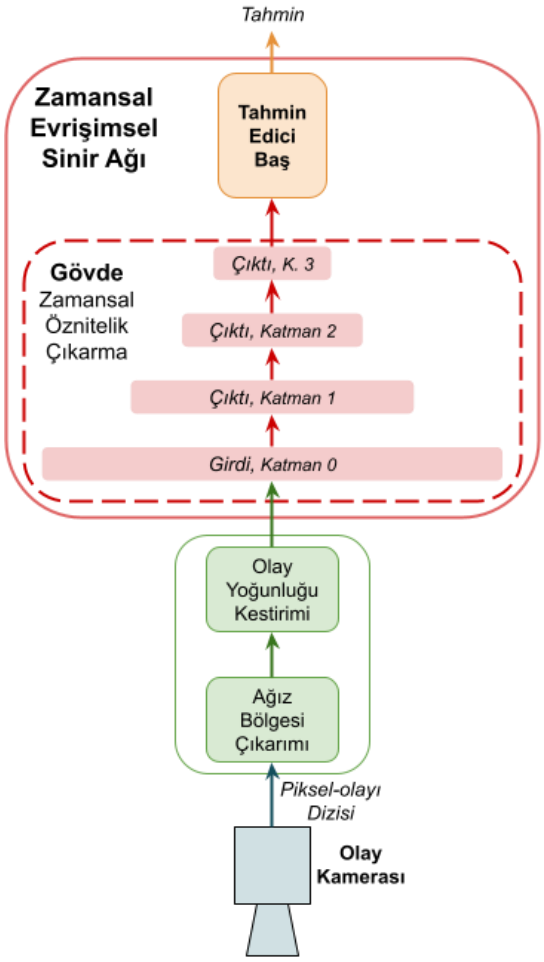
Piksel-olayları yığınlanarak istenen bir zaman çözünürlüğünde olay yoğunluğu örneklenir. Bunun için doğrusal interpolasyona dayanan ve olay dürtü-tensörü olarak da bilinen gösterimin (Gehrig vd., 2019) uzamsal olarak indirgenmiş özelleşmesi, zamanla değişen ağız alanı normalizasyonu yapılarak uygulanır. Ağız alanı normalizasyonu, görüntü düzlemindeki bölgenin poza bağlı büyüyüp küçülme farklılıklarına karşı değişmezlik kazandırmak için yapılır. Böylece, ağız üzerindeki toplam piksel-olayı sayısı yerine, ağız bölgesindeki olay yoğunluğuna göre analiz yapılır. Aşağıda formülü verilen bu gösterimde M ağız bölgesindeki olay indekslerinin kümesini, N toplam olay sayısını, T zamandaki örnekleme için (yani zamansal niceleme için) bir klipte kullanılan zaman selelerinin toplam sayısını, t sele indeksini, A_t de t anındaki ağız bölgesi alanını ifade eder. Bu sayede her klip, ifadesi

$$I \in R^{2\times T} \tag{1}$$

$$I_{p,t} = \frac{1}{A_t}\left(\sum_{i,p^i=p,i\in M} max\left(0, 1-\left|t-t_b^i\right|\right)\right) \tag{2}$$

$$t_b^i = (T-1)\cdot\frac{t^i-t^0}{t^N-t^0} \tag{3}$$

şeklinde olan bir olay yoğunluğu zaman dizisi **I** matrisi ile temsil edilir. Bu çalışmada, 10 ms genişliğinde zaman seleleri üzerinde 10 ms boyunda adımlarla hesaplama yapılarak olay yoğunluğu dizileri hesaplanır.



Şekil 1. Olay kamerası konuşma sesi tespitinin blok şeması (Event camera voice activity detection block diagram)

3.2. Alıcı Alan Genişletme Modelleri (Receptive Field Widening Models)

Evrişimsel sinir ağ tasarımlarında sık kullanılan alıcı alan (receptive field) terimi, sinir biliminden gelmektedir. Alıcı alan, kısaca, sinirsel bir yanıt üretilebilmesi için algı uzayında gereken bölgenin genişliğidir. Evrişimsel ağlarda ise, evrişim katmanlarındaki faydalı öznitelik çıktılarını oluşturabilen girdi sinyalindeki alan genişliği olarak tanımlanır. Alıcı alanın genişliği ve nasıl genişletildiği bir evrişimsel model mimarisinin başarımını belirleyen başlıca faktörlerdendir. Temel mimaride her ne kadar

ağın baş kısmında tüm girdi alanı kullanılabilir olsa da, gövdede oluşturulan zamansal öznitelikler kısıtlı bir alandaki örüntüleri yakalayabilmektedirler. Dolayısıyla, zamansal alıcı alanı etkili ve aynı zamanda verimli bir şekilde genişleterek kullanışlı öznitelikler çıkarabilmek için farklı yöntemlerin incelenmesi gerekir.

En basit olarak, evrişim çekirdeğinin genişliğini artırmak doğrudan alıcı alanı genişletir. Ancak çekirdek genişliğindeki artış alıcı alan genişliğindeki artışa eşit olacağından, bu yöntem son derece verimsizdir ve aşırı sayıda parametre (sinir bağlantı ağırlıkları) gerektirmesi öğrenme problemini çok zorlaştırır. Bundan dolayı bu yöntem hiçbir çalışmada kullanılmamıştır. Aksine modern mimariler en küçük çekirdek genişliği ile derinliği artırmayı ilke edinmişlerdir. Daha derin ağlar, daha çok aktivasyon fonksiyonunu da kullanmayı sağladığından, öğrenmenin başarısını artıran doğrusalsızlık derecesini de yükseltirler. *L* sayıda evrişim katmanına sahip bir mimaride, eğer her katmandaki çekirdek genişliği *k* ise, en son katmanın girdi katmanı üzerindeki alıcı alanının genişliği

$$RF = 1 + (k-1)L \tag{4}$$

formülü ile hesaplanır. Denklem 4'ten görüldüğü üzere katman sayısı ve alıcı alan arasında faktörü *k-1* olan doğrusal bir ilişki vardır. Daha hızlı, yani daha az sayıda katman ve parametre ile, benzer alıcı alan genişliklerine aşağı-örnekleme yoluyla ulaşılabilir. Evrişimsel ağlarda, maksimum-havuzlama ve adımlı-evrişim olmak üzere iki yaygın aşağı-örnekleme yöntemi kullanılır. Adımlı-evrişimde, evrişim çekirdeği kaydırılırken standart bir birimlik kaydırma yerine daha büyük bir tamsayı kaydırma miktarı kullanılır. Maksimum-havuzlamada ise, çok ufak bir alanda en büyük değeri seçen bir doğrusal-olmayan ara hesaplama katmanı eklenir ve havuzlama penceresi de birden büyük tamsayı bir adımla kaydırılır. Katmandan katmana değişen çekirdek ve adım genişlikleri ile genelleştirilmiş alıcı alanın kapalı çözümü

$$RF = 1 + \sum_{l=1}^{L}\left((k_l - 1)\prod_{i=1}^{l-1} s_i\right) \tag{5}$$

olduğu gösterilebilir (Araujo vd., 2019). Burada *s*, her iki yöntemdeki adım boyudur. Sabit *k* ve *s* için, birden büyük adımlar olduğunu varsayarsak, aşağıdaki formülü toplam serisi kuralına göre elde ederiz.

$$RF = 1 + (k-1)\frac{s^L - 1}{s-1} \tag{6}$$

Alıcı alanı hızla büyütmek için daha farklı bir yol ise genleştirilmiş evrişim uygulamaktır (Yu ve Koltun, 2016). Genleştirilmiş evrişimin standarttan farkı, bir çıktı noktasındaki evrişim yanıtını hesaplarken belli girdi noktalarının atlanarak çekirdek çarpımının yapılmasıdır. Genleşme faktörü *d* olsun. O zaman ardışık her girdi noktası çifti arasında *d-1* nokta atlanarak evrişim uygulanır. Genleştirilmiş evrişimde çekirdek genişlemez fakat girdi sinyali üzerinde kapsadığı sınırlar, bir nevi her araya *d-1* tane delik konularak suni olarak genişletilmiş olur. Dolayısıyla bir çıktı noktasının hemen altındaki girdi katmanında kapladığı alan, standart evrişimdeki *k* yerine,

$$k' = 1 + (k-1)d \tag{7}$$

olur. Dolayısıyla, eğer Denklem 4'te yerine koyarsak alıcı alanı

$$RF = 1 + (k-1)dL \tag{8}$$

şeklinde elde ederiz. Literatürde genleştirmeyi daha da hızlı yapabilmek için tercih edilen bir yöntem de genleştirmeyi üstel olarak yapmaktır (Rethage vd., 2018). Genleştirme faktör tabanına *m* dersek, katman endeksine göre genleştirme faktörü $d = m^{l-1}$ olur. Bu ifadeyi Denklem 7'de yerine koyup sonucunu da Denklem 5'te yerine koyarsak, adım genişliğini $s = 1$ olduğu takdirde, toplam serisi uygulamasıyla alıcı alan formülü

$$RF = 1 + (k-1)\frac{m^L - 1}{m-1} \tag{9}$$

olarak elde edilir. Ayrıca, birden farklı adım genişliği için alıcı alanın

$$RF = 1 + (k-1)\frac{(sm)^L - 1}{sm-1} \tag{10}$$

olduğu gösterilebilir. Bölüm 3.4.'te uygun tasarım parametreleri saptanarak, burada gösterilen çeşitli alıcı alan genişletme yöntemleri Bölüm 4.3.'te verimlilik ve başarım açılarından değerlendirilirmiştir.

3.3. Derinlemesine ve Gruplamalı Ayrılabilir Evrişim Modelleri (Depthwise and Groupwise Separable Convolution Models)

Derinlemesine ayrılabilir evrişim yöntemi, MobileNets (Howard vd., 2017) ile yaygın kullanımı ortaya çıkan, çok çeşitli görevlerde yüksek verimlilik kazandırdığı gösterilen bir yöntemdir. Burada kullanılan derinlemesine terimi kanal ekseni üzerinde yapılan anlamına gelir. Bu yöntem ile, çok kanallı evrişimler tek kanallı evrişimlere ayrılma yoluyla faktörize edilerek, kanal sayısıyla orantılı işlem tasarrufu sağlanır. İki aşamada gerçekleşir. İlk olarak, her bir girdi kanalına özgü ayrı bir filtre uygulanır. Bu ilk işlem derinlemesine evrişimdir ve bu aşamada girdi kanalları arasındaki

herhangi bir etkileşim örüntüsü öğrenilmez. İkici aşamada ise, tam tersi şekilde, kanallar arasındaki etkileşimi öğrenen fakat girdi alanı üzerindeki örüntüleri algılamayan noktalamasına evrişim (pointwise convolution) uygulanır. Noktalamasına evrişim, girdi alanında sadece bir birim kaplayan evrişimdir. Bu evrişim filtresi, sadece tek bir girdi noktasındaki kanal değerlerinden hedeflenen çıktı kanal sayısını oluşturacak şekilde ayarlanır. Böylece, normalde çok kanallı girdi ve çıktılar için uygulanan tek bir büyük filtre yerine, çok daha az parametre sayısına sahip olan kanala özgü filtreler ve ardından da, kanalların doğrusal birleşimini gerçekleştiren noktalamasına filtre uygulanmasıyla aynı görevin daha verimli bir şekilde yerine getirilmesi sağlanır.

Bir evrişim katmanında filtre genişliği *K*, girdi kanal sayısı *M* ve çıktı kanal sayısı *N* olsun. O zaman standart yöntemde toplam evrişim parametre sayısı $M \times K \times N$ olur. Eğer çıktı alan genişliği de T ise, evrişim hesaplama yükü de $M \times K \times N \times T$ olur. Derinlemesine evrişimde ise, $M = N$ olur ve her kanal için ayrı filtre uygulandığından $M \times K$ parametre gelir. Noktalamasına filtreden dolayı da $M \times N$ parametre olduğundan, sonuçta toplam parametre sayısı $M \times (K + N)$ olur. Dolayısıyla derinlemesine evrişimin toplam hesaplama yükü $M \times (K + N) \times T$ olarak bulunur. Böylece hesaplama azalma oranı

$$\frac{M \times (K+N) \times T}{M \times K \times N \times T} = \frac{1}{N} + \frac{1}{K} \tag{11}$$

şeklindedir. Filtre genişliği *K* genelde ufak sabit bir değerdir. Tipik olarak *K*=3 değeri için, iki üç kat arasında bir hesaplama kazancı elde edilir.

Kanallar üzerinde faktörizasyon yaparken diğer bir seçenek de her bir kanal yerine, kanalları gruplandırıp her bir kanal grubu için bir filtre kullanmaktır, yani gruplamalı evrişim uygulamaktır (Krizhevsky vd., 2012). Derinlemesine evrişimle aynı şekilde yine ikinci aşamada noktalamasına evrişim uygulanarak ayrılabilir gruplamalı evrişim gerçekleştirilir. Dolayısıyla, *G* tane grup kullanıldığında gruplamalı evrişim toplam parametre sayısı

$$G \times M/G \times M/G \times K = M^2 \times K/G \tag{12}$$

ve noktalamalı evrişim toplam parametre sayısı $M \times N$ olduğundan, toplam parametre sayısı $M \times \left(\frac{M \times K}{G + N}\right)$ olur. Böylece, Denklem 11'de yaptığımız gibi hesaplama azalma oranı

$$\frac{M \times (M \times K/G + N) \times T}{M \times K \times N \times T} \tag{13}$$

$$= \frac{M}{G \times N} + \frac{1}{K} \tag{14}$$

olarak bulunur. Burada grup sayısını artırarak hesaplama kazancını artırabildiğimiz ve kanal sayısını *M*'ye eşitlersek de en fazla kazanç olan Denklem 1'deki sonucu elde ettiğimiz görülür. Hesaplama yükünü azaltırken başarım düşebileceğinden, Bölüm 4.4'te farklı grup sayılarına bakılarak hesaplama ve başarım değerleri incelenmiştir.

3.4. Ağ Mimarileri (Network Architectures)

Şekil 1'de gösterilen gövde ve tahmin edici baş kısımları için farklı tasarımlar ele alınabilir. Tahmin edici baş kısmında, gövdede elde edilen zamansal öznitelikler üzerinden sınıflandırma gerçekleştirilir. Bu tür görevler için sıklıkla çok-katmanlı algılayıcı kullanılır. Ancak, çok-katmanlı algılayıcılara gerek olmadan yüksek başarımların elde edilebileceği gösterilmiş ve yaygınlaşmıştır (Szegedy vd., 2015). Bu tür modellerde önce, bütünsel ortalama havuzlama yoluyla çok kanallı olan bütün alan kanal sayısını değiştirmeden indirgenir ve sonra, doğrusal katman uygulanarak istenen hedefler tahmin edilir. Dolayısıyla çok katmanlı modellere göre oldukça basittir ve hiper-parametre gerektirmez. Bu çalışmada, bu en basit tahmin edici baş tasarımının yüksek KST başarımı elde ettiği görüldüğü için karmaşıklığı daha yüksek olan çok-katmanlı tasarımların kullanılmasına gerek duyulmamıştır.

Gövde tasarımında ilke olarak, her bir yukarı katmanda bir birimdeki öznitelik miktarını yani kanal sayısını artırmak ve böylece daha zengin bir temsil elde etmek hedeflenirken, aşağı-örnekleme ile de aktivasyon hacminin aşırı büyümesinin önlenmesi hedeflenmiştir; çünkü aktivasyon hacminin fazla büyük olması işlem yükünü ciddi oranda artırmaktadır. Bu uygulama, sayısal bilgisayar donanımı için elverişli olan ikili sisteme göre

$$c_l = 2^{1+l} \tag{15}$$

parametrizasyonu ile yapılır. Burada, *l* katman indisidir ve girdi katmanı da zaten iki kanallı olduğundan $c_0 = 2$'dir. Evrişim, genişliği üç olan çekirdekler kullanıldığında aktivasyon hacmini ikinin katları şeklinde olmasını sağlamak amacıyla, bir birim genişliğindeki sıfır-dolgulama ile yapılır. Her katmanda, evrişimden sonra ReLU aktivasyonu uygulanır.

Aktivasyon alanını indirgemek için, Bölüm 3.2.'de anlatılan farklı alıcı alan genişletme modelleri uygulanır. Bu çalışmada, SE olarak adlandırdığımız standart model, aşağı-örnekleme indirgeme faktörü, yani adım genişliği *s*=2, ile maksimum-havuzlama yapılan model anlamına gelmektedir. Maksimum-havuzlama yerine doğrudan adımlı evrişim ile aynı adım genişliği, literatürde tamamen evrişimsel, yani TE modeller ile yaygındır (Long vd., 2015). Bölüm 4.4'te bu modeller karşılaştırılır. Bu iki modelin algısal alan genişlikleri Denklem 6 ile hesaplandığında, *k*=3 olduğundan, katman sayısına bağlı olan formül

$$RF = 1 + (3-1)\frac{2^{L}-1}{2-1} = 2^{L+1} - 1 \quad (16)$$

olarak bulunur. Aşağı-örnekleme yapmayıp, sadece üstel genleştirme yaptığımızda (ÜGE), eğer Denklem 9'da m=2 olarak alırsak, yine Denklem 16'daki aynı alıcı alan ilişkisini elde ederiz.

Genleştirme yolu ile alıcı alanı genişletip ve aynı zamanda aşağı-örnekleme yapan iki farklı genleştirme modeli uygulanır. Bunlar sabit genleştirme ve üstel genleştirme modelleridir. Genleşmiş filtre genişliği Denklem 7 ile hesaplanır. Böylece, sabit genleştirme ve aşağı-örnekleme için (GSE), d=2, s=2 ve, k=3 olduğundan, Denklem 6 ile ifade edilen alıcı alan formülüne göre

$$k' = 1 + (3-1)2 = 5 \quad (17)$$

$$RF = 1 + (5-1)\frac{2^{L}-1}{2-1} \quad (18)$$

$$= 2^{L+2} - 3 \quad (19)$$

bulunur. Alıcı alanı artan katman sayısıyla çok daha hızlı büyüten üstel genleştirme ve aşağı-örnekleme için (ÜGSE) ise, m=2 için $d = 2^{l-1}$ olur ve Denklem 10'da m=2, s=2 ve k=3 için

$$RF = 1 + (k-1)\frac{(2\cdot 2)^{L}-1}{2\cdot 2-1} \quad (20)$$

$$= (1 + 2^{2L+1})/3 \quad . \quad (21)$$

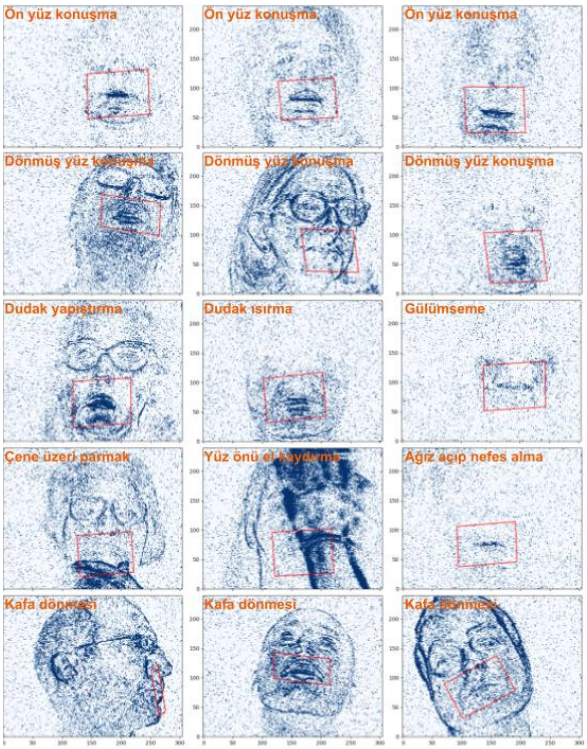
Tablo 1. Evrişimsel (E.) modellerin kısaltmaları (Abbreviations of the convolutional (E.) models)

Kısaltma	Açılım
SE	Standart Evrişimsel
TE	Tamamen Evrişimsel
ÜGE	Üstel Genleştirilmiş E.
ÜGSE	Üstel Genleştirilmiş Standart E.
GSE	Sabit Genleştirilmiş Standart E.
ASE	Ayrılabilir Standart E.
GASE	Sabit Gen. Ayrılabilir Standart E.

Diğer taraftan, verimliliği daha da artırabilmek için, Bölüm 3.3'te anlatılan derinlemesine ayrılabilir ve gruplamasına ayrılabilir evrişimsel tasarımları uygulanır. Gruplamasına evrişim uygularken, ikinin katları şeklinde üstel olarak farklı grup sayıları ile tasarımlar değerlendirilir. Ayrıca, en verimli modeli araştırırken, genleştirme ve ayırma teknikleri birleştirilerek, genleştirilmiş ayrılabilir evrişim en fazla verimlilik için incelenmiştir. Tablo 1'de, bu çalışmada değerlendirilen bütün tasarımlar listelenmektedir. Bölüm 4'te, burada anlatılan bütün tasarımlar ile karşılaştırmalı değerlendirmeler sunulur.

4. Deneyler (Experiments)

Bu bölümde önce deneylerde kullanılan veri kümesi Bölüm 4.1'de anlatılır ve eğitim parametreleri Bölüm 4.2'de verilir. Bölüm 4.3'te farklı alıcı alan modellerinin başarımları ve Bölüm 4.4'te farklı verimlilik artırma modellerinin başarımları karşılaştırılarak incelenir. Bölüm 4.4'te ise modellerin başarım karakteristikleri detaylı bir şekilde analiz edilir.



Şekil 2. Örnek anlık histogram görselleştirmeleri ve hesaplanan ağız bölgeleri (Example histogram visualization snapshots and computed mouth regions)

4.1. Veri kümesi (Dataset)

Deneylerde kullanılan veri kümesi MHz mertebelerindeki zaman çözünürlüğüne erişen ve 304 x 240 piksel çözünürlüğüne sahip bir olay kamerası ile oluşturulmuştur (Savran vd., 2018). Toplam 486 tane klip vardır. Kliplerin 324 tanesi konuşma sesi içermekte, geri kalan 162 tanesi ise içermemektedir. Sesli klipler konuşma işleme çalışmalarında sıklıkla kullanılan fonetik olarak zengin TIMIT (Wrench 2006) metinlerinden oluşmuştur. Konuşma kliplerinin 54 tanesinde kafa dönmesi hareketleri de vardır ve kişiler açılı pozda konuşmuşlardır; diğer konuşma klipleri ön yüz şeklindedir. Konuşmasız klipler, konuşmaya benzeyen ağız hareketlerini de içeren çeşitli eylemler içermektedirler. Bunlar dudak yapıştırma şeklinde açıp kapama, dudak ısırma, gülümseme, çene üstüne parmakla dokunma, eli yüz üzerinde gezdirme ve ağız açarak nefes alma gibi yüz kapatma hareketleri ve, üç eksende farklı hız ve tekrarlama sayılarıyla yapılan kafa

dönmesi hareketleridir. Bu eylemleri içeren örnekler Şekil 2'de gösterilmektedir. 18 kişiden toplanan veri kümesinde, klip sayı ve tipleri aynıdır ancak içerikler farklıdır. Konuşmalı kliplerde TIMIT cümleleri kişiden kişiye değişmektedir ve konuşmasız kliplerde kişiler belirtilen eylemleri istedikleri gibi gerçekleştirmişlerdir. Bütün kliplerde ses kanalı görsel kanal ile senkronize edilmiştir. Ayrıca, kliplerin göz ve ağız merkezleri işaretlenmiştir. Çeşitli pozitif ve negatif örnek klipler ses dalga biçimleri ve olay yoğunluk grafikleri ile Şekil 3'te gösterilmektedir. Deneylerde iki kişi geçerleme ve dört kişi test kümesi için ayrılmıştır.

4.2. Eğitim Parametreleri (Training Parameters)

Kullanılan öğrenme kayıp fonksiyonu ağırlıklı ikili-çapraz-entropi-logits fonksiyonudur. Burada ağırlıklandırma, pozitif ve negatif sınıf örnek miktarlarının öğrenmedeki etkilerini dengeleyebilmek için, kayıp fonksiyonunu hesaplarken sınıf örnek sayısına ters orantılı ağırlık çarpanı uygulanarak yapılır. Ayrıca, olay yoğunluğu girdi kanalları üzerinde standart normalizasyon yapılır. Eğitimler için, 32 kliplik yığınlar üzerinden ADAM en iyilemesi 10^{-3} sabit öğrenme oranı ile çalıştırılır. Farklı uzunlukta klipler ile bir yığın tensörü oluşturabilmek için sıfır-dolgulama yapılarak sabit 1024 örnekli klipler elde edilir. Her bir örnek 10 milisaniyeye denk düştüğünden, sabit klip uzunluğu yaklaşık 10 saniyedir. Eğitimler, ayrılabilir evrişim modelleri haricinde varsayılan olarak sabit 50 devirlik döngülerle yapılmıştır; ayrılabilir evrişim modellerinde ise, yakınsamanın daha uzun sürdüğü görüldüğünden 100 devirlik döngülerle yapılmıştır.

4.3. Alıcı Alan Genişletme Modellerinin Değerlendirilmesi (Evaluation of the Receptive Field Widening Models)

Tablo 2'de farklı evrişimsel modellerin katman sayılarına göre değişen alıcı alanları gösterilmektedir. SE ve ÜGE'nin alıcı alanlarının eşit olduğu görülmektedir. Yani sadece aşağı-örnekleme yapıldığında (SE) ve aşağı-örnekleme olmadan üstel genleştirme yapıldığında (ÜGE) alıcı alanlar eşit çıkmaktadır. Bunun nedeni, Bölüm 3.4'te anlatılan tasarımların Denklem 19'daki aynı ilişkiye varmasıdır. Diğer taraftan, sabit genleştirme ile beraber aşağı-örnekleme yapıldığında (GSE), Denklem 22'den dolayı yaklaşık iki kat daha fazla alan genişliği elde edilir. Üstel genleştirme ile beraber yapıldığında (ÜGSE) ise, Denklem 24'ten dolayı çok daha hızlı bir büyüme gerçekleşir.

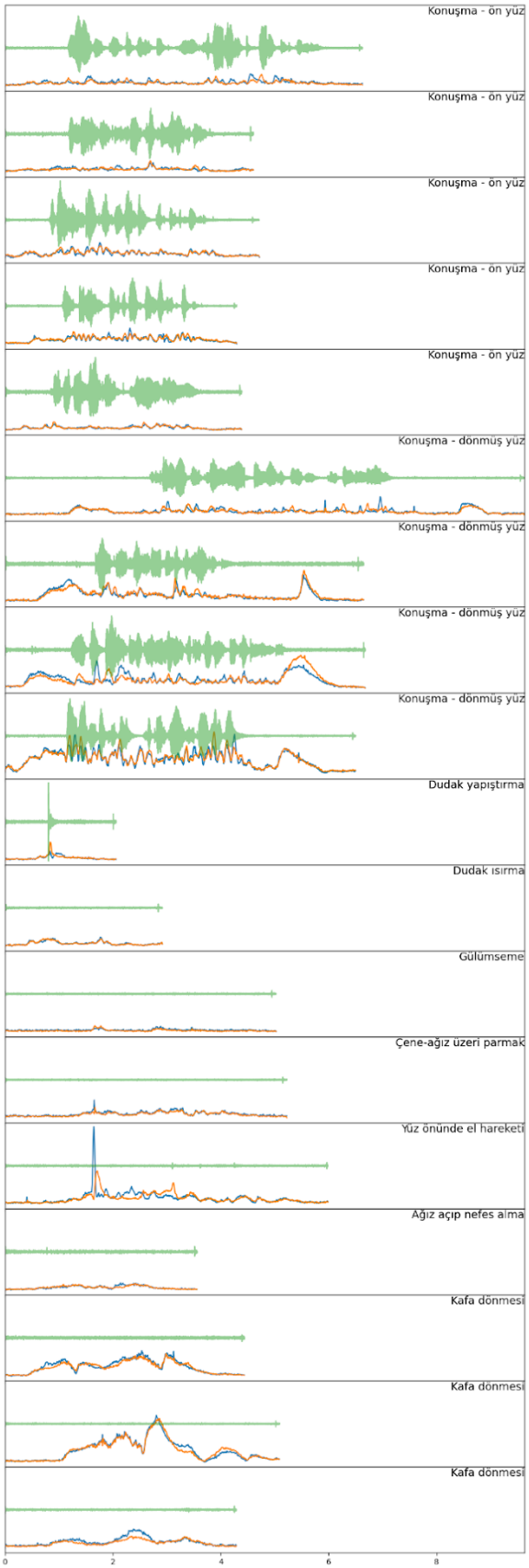
Tablo 2. Alıcı alan genişletme modellerinin alıcı alanları (Receptive fields of the receptive field widening models)

Kat	SE	ÜGE	ÜGSE	GSE
1	3	3	3	5
2	7	7	11	13
3	15	15	43	29
4	31	31	171	61
5	63	63	683	125
6	127	127	2731	253
7	255	255	10923	509

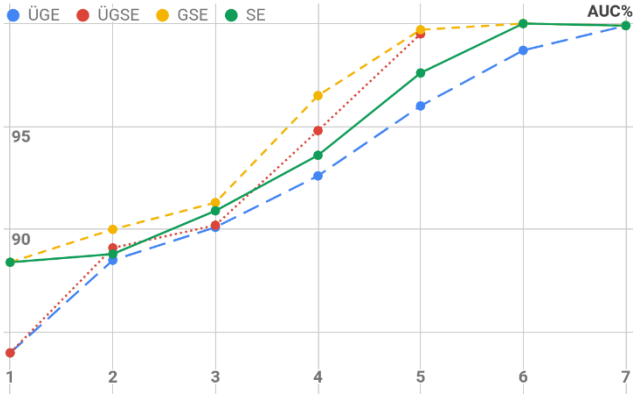
Tablo 3. Alıcı alan genişletme modellerinin MFLOPS çarpma-toplama yükleri (MFLOPS multiplication-addition loads of the receptive field widening models)

Kat	SE	ÜGE	ÜGSE	GSE
1	0.03	0.03	0.03	0.03
2	0.08	0.13	0.08	0.08
3	0.18	0.54	0.18	0.18
4	0.39	2.11	0.35	0.38
5	0.78	8.15	0.51	0.75
6	1.57	30.63	-----	1.44
7	3.15	107.9	-----	2.62

Tablo 3'te de, Tablo 2'deki alıcı alanları listelenen modellerin toplam çarpma-toplama kayan nokta hesaplama yükleri, milyon birimiyle, MFLOPS adı altında, gösterilir. ÜGE modelinde aşağı-örnekleme olmadığı için, Denklem 18'teki ilişkiye göre artan kanal sayısıyla beraber aktivasyon hacmi çok büyür; dolayısıyla da, işlem yükünün katman sayısıyla beraber çok hızlı bir şekilde arttığı görülmektedir. Diğer üç model ise benzer seviyelerde işlem yüküne sahiptir. Artan kanal sayısıyla SE, ÜGSE ve, GSE modellerinin işlem yüklerinde farklılıkların gözlemlenmesinin sebebi, genleştirilmiş modellerde alıcı alanların deneylerde kullanılan 1024 tensör uzunluklarına göre fazla büyümesi ve kırpılma olmasıdır. Bu büyüme ve kırpılmanın, ÜGSE ile altıncı ve yedinci katmanlarda aşırı olmasından dolayı, ÜGSE'nin o katman sayıları deneylere dahil edilememiştir; zaten o kadar büyük alıcı alan kullanmaya zorlamak fazla gecikmeye neden olacağından pratikte kullanışlı da değildir.



Şekil 3. Pozitif ve negatif örnek kliplerin ses dalga biçimleri ve olay yoğunlukları ile gösterimi (Example positive and negative clips shown by audio waveforms and event intensities)



Şekil 4. Alıcı alan genişletme modellerinin değişen katman sayılarına göre (yatay eksen) AUC başarım yüzdeleri (dikey eksen) (AUC percentage performance (vertical axis) of the receptive field widening models with varying layer counts (horizontal axis))

Modellerin başarımları, Şekil 4'te alıcı çalışma karakteristiği (ROC) eğrisi altında kalan alan (AUC) ölçümü ile karşılaştırılır. ROC eğrisi, her bir yanlış pozitif oranına karşı doğru pozitif oranını olası bütün karar eşiği değerleri için verdiğinden, AUC, eşik değeri seçiminden bağımsız olarak elde edilebilecek teorik doğru sınıflandırma başarımını belirtir. Şekil 4'te, GSE ve ÜGSE modellerinin, altı katmana kadar SE'den daha iyi başarım elde ettiği, ÜGE'nin de her katman sayısı için en kötü başarımı elde ettiği görülür. Bu sonuç, alıcı alanın genişliğinin büyük olmasının avantajlı olduğunu göstermektedir (Tablo 2'de alıcı alanlara bakınız). ÜGE'nin işlem yükü gereksinimi de diğer modellere göre çok daha yüksektir. Bu derece keskin başarım ve işlem yükü dezavantajları, aşağı-örnekleminin bu KST probleminde çok önemli olduğuna dair kuvvetli kanıt teşkil eder. Her katmanda GSE'nin başarımı ÜGSE'nin başarımından biraz daha yüksektir. Bu da, sinir ağı derinliği ile artan alıcı alanın gereğinden çok fazla artmasının öğrenmeyi olumsuz etkilediğine işaret eder.

En yüksek başarımlar altı katman ile elde edilmiştir ve, yedi katman kullanıldığında az bir miktar başarım düşüşü gözlemlenmiştir. Altı katman derinlikte, %100 AUC başarımına sahip SE ve GSE modelleri, sırasıyla 127 birimlik ve 253 birimlik alıcı alana sahip olurlar. Düşük alıcı alanla çalışmak daha düşük gecikme anlamına geldiğinden bir avantajdır. Diğer taraftan, beş katman derinliğinde, %99.7 AUC ile GSE açık olarak SE'den daha iyi başarım elde eder ve 125 birimlik bir alıcı alana ihtiyaç duyar. Daha düşük katman sayısı ile yüksek başarım elde ettiğinden GSE'nin işlem yükü de altı katmanlı SE'ye göre daha azdır. Altı katmanlı SE 1.57 MFLOPS ile çalışırken, beş katmanlı GSE 0.75 MFLOPS ile neredeyse aynı alıcı alan ihtiyacı ile çalışır. Bu nedenlerden ötürü, en yüksek başarım istenildiği görevlerde altı katmanlı SE modelinin, fakat verimliliğin kritik olduğu ve az bir miktar başarım

düşüşünün tolere edilebileceği koşullarda beş katmanlı GSE modelinin kullanılması uygun olacaktır.

4.4. Ayrılabilir Evrişim Yoluyla Verimlilik Artırma (Efficiency Improvement via Separable Convolution)

Bu bölümde, işlem yükünü hafifleterek verimliliği daha da yukarı seviyelere çıkarmak amacıyla, ayrılabilir evrişim tekniğinin kullanılması değerlendirilir. Hedefimiz düşük işlem yükü olduğu için, Bölüm 4.3'te yüksek başarım seviyesinde en iyi verimliliği sağladığı saptanan beş katmanlı GSE modelinin verimliliğinin daha da artırılması için deneysel inceleme yapılır. Tablo 4'te, derinlemesine ve gruplamasına ayrılabilir evrişim modellerinin MFLOPS işlem yükleri ve AUC başarım yüzdeleri sunulmaktadır. Genleştirilmiş ayrılabilir modeli ifade eden GASE-# biçiminde # sembolü ya gruplamalı yapıdaki grup sayısını ya da Der kısaltmasıyla derinlemesine ayrılabilir evrişimi gösterir. Grup sayısı ikinin katları şeklinde artırılmıştır. Tablo 4'te, grup sayısı artarken işlem yükünün azaldığı fakat azalma hızının grup sayısıyla beraber yavaşladığı görülür. Bunun sebebi alt katmanlardaki düşük kanal sayısı ama büyük aktivasyon alanıdır. Örneğin ilk katmanda iki kanal vardır, dolayısıyla en fazla ikiye bölünebilir. Yukarı katmanlardaki bölünmelerin işlem kazancı ise azalan aktivasyon alanından dolayı nispeten azdır. Diğer bir ifadeyle, alt katmanlarda kanalları ayırmak üst katmanlardakine kıyasla bize daha çok işlem tasarrufu sağlar; çünkü alt katmanlardaki aktivasyon alanları aşağı-örneklemeden dolayı çok daha büyüktür.

Tablo 4. Modellerin (GASE-#, #: grup sayısı veya derinlemesine ayrışım) işlem yükleri ve başarımları (Computation loads and performances of the models (GASE-#, #: group count or depthwise separation)

Model	MFLOPS	AUC %
GSE	0.75	99.7
GASE-2	0.46	99.7
GASE-4	0.37	99.3
GASE-8	0.33	99.1
GASE-16	0.31	98.6
GASE-32	0.30	99.8
GASE-Der	0.30	99.7

GSE'nin yüksek AUC başarım yüzdesi genelde oldukça korunduğu görülmektedir. En düşük işlem yükü derinlemesine ayrılabilir model (GASE-Der) ile elde edildiğiden ve GSE ile aynı başarımı yakaladığı gözlemlendiğinden, en verimli model olarak seçilir. 0.75 MFLOPS işlem yükü, iki kattan fazla düşerek 0.30 MFLOPS seviyesine inmiştir. Ayrılabilir evrişim tekniği SE ve diğer modellerde de verimlilik kazandırır. Ancak ayrılabilir SE'nin ulaştığı başarım, GSE'nin başarım değerlerine ve hatta daha düşük seviyelere düştüğü gözlenmiştir ve buna rağmen işlem yükü daha fazladır.

4.5. Başarım ve Verimlilik Karşılaştırmaları (Comparisons of Performance and Efficiency)

Bu bölümde, en yüksek başarım için seçilen altı katmanlı SE modeli ve en yüksek verimlilik için seçilen beş katmanlı GASE modeli, KST problemi için detaylı olarak karşılaştırılır. Ayrıca, SE modeliyle aynı nitelikte olan fakat tek farkı aşağı-örneklemeyi maksimum-havuzlama yerine doğrudan adımlı evrişim ile yapan TE modeli de incelemeye dahil edilir. Tablo 5'te bu üç modelin MFLOPS işlem yükleri ve AUC başarım yüzdeleri gösterilmektedir. TE modeli 0.79 MFLOPS ile SE'den iki kat daha verimli çalışmakta fakat %98.2 AUC ile en düşük başarımı elde etmektedir. İki kat verimlilik, evrişimin iki birimlik adımla gerçekleşmesi sayesindedir; oysa maksimum havuzlamadan önce yapılan evrişim tek adımlı olur. Tamamen evrişimli sinir ağının bu problemde geride kalmasının sebebi, maksimum-havuzlama ile daha kuvvetli öteleme değişmezliği elde edilmesi olabilir; çünkü, konuşma aralıkları son derece değişken olarak ortaya çıkar. Tablo 5'te en verimli GASE modelinin, sadece %0.3'lük bir başarım kaybına uğrayarak 0.30 MFLOPS ile işlem yükünü SE'nin beşte birinden daha azına indirdiği görülmektedir.

Tablo 5. Tamamen (TE), standart (SE) ve, sabit genleştirilmiş derinlemesine ayrılabilir evrişimsel (GASE) modellerinin işlem yükleri ve başarımları (Computation loads and performances of the fully (TE), standard (SE) and, dilated depthwise separable convolutional (GASE) models)

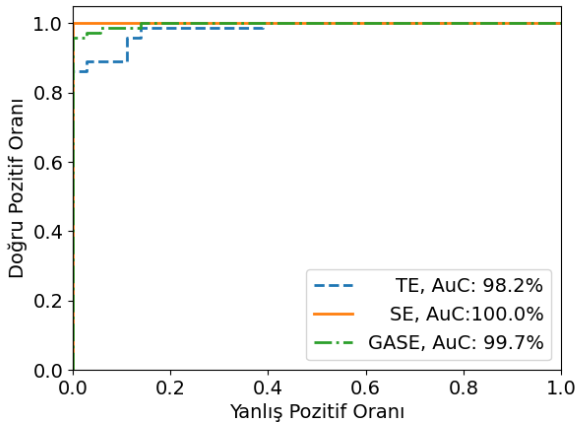
Model	MFLOPS	AUC %
TE	0.79	98.2
SE	1.57	100.0
GASE	0.30	99.7

Şekil 5'te bu üç modelin ROC eğrileri karılaştırılmaktadır. SE'nin ideal ROC eğrisine sahip olduğu, TE'nin sesi (pozitif) %100 doğru saptama başarısını ancak %50'ye yakın yanlış pozitif oranlarında elde ettiği görülür. GASE'nin doğru pozitif oranları ise SE'ye yakındır ve %100'e çabuk ulaşır.

Bu üç modelin farklı kategorideki kliplerdeki başarımları, karar eşik değerine bağımlı iki farklı çalışma koşulu altında detaylı olarak Tablo 6'da gösterilmektedir. Bu koşullar, tüm test kümesinde en az %80 doğru negatif sağlama ve en az %99 doğru pozitif sağlama koşullarıdır. İlk koşul bize, konuşma sesi olan klipleri saptama başarımını en fazla %20 yanlış alarm toleransı ile değerlendirme imkanı verirken, ikinci koşul

ise neredeyse kusursuz sesli klip saptama yaparken yanlış alarm vermeme başarımını, yani ses aktivitesi dışındaki yüz aktivitesi ve eylemlerine karşı gürbüz olma başarımını, daha iyi analiz etmemizi olanaklı kılar. Pozitif örnekler, ön yüzlü ve dönmüş yüz pozlu konuşma kategorilerine ayrılmıştır. Negatif örnekler ise, konuşma içermeyen fakat kafa dönmeleri, dudak açma-kapama ısırma, ağız açıp nefes alma, gülümseme, çene veya ağız üzerinde parmak tutma ve, yüz önünde eli hareket ettirme kategorileridir.

Tablo 6'da, SE'nin her kategoride kusursuz çalıştığı görülmektedir. GASE, en az %80 doğru negatif koşulu altında, %100 doğru pozitif ve %86.1 doğru negatif elde etmiştir. TE de aynı doğru negatif oranında kalmış fakat %98.6 doğru pozitif elde edebilmiştir. Ancak TE'nin ön yüzlü konuşmalarda kusursuz tanıma başarımı gösterdiği fakat dönmüş yüzlerde hata yaptığı görülür. En az %99 doğru pozitif oranı altında ise GASE yine %86.1 doğru negatif yüzdesini elde ederken, TE'de bu oran çok büyük bir düşüşle %61.1'e inmiştir. TE'nin açık olarak konuşma dışındaki aktivitelere genelde başarılı olmadığı görülmektedir, yani tasarımda maksimum havuzlama kullanmadan yüksek KST başarımı elde etmek mümkün olmamaktadır. Negatif örneklere bakıldığında, genel olarak verimliliği yüksek olan GASE modelinin en çok hata yaptığı kategorilerin dudak yapıştırma-ısırma, çene-ağız üzeri parmak ve, yüz önünde el kaydırarak kapatma kategorileri olduğu görülmektedir. Kafa dönmelerinde %90 üzeri başarı ile daha iyi ayırt etmekte, gülümseme ve ağız açarak nefes alma kategorilerinde ise hata yapmadığı gözlemlenmektedir.



Şekil 5. Tamamen (TE), standart (SE) ve, genleştirilmiş derinlemesine ayrılabilir evrişimsel (GASE) modellerinin ROC eğrileri (ROC curves of the fully (TE), standard (SE) and, dilated depthwise separable convolutional (GASE) models)

Tablo 6. En az %80 doğru negatif koşulu ve en az %99 doğru pozitif koşulları altında, tamamen (TE), standart (SE) ve, genleştirilmiş derinlemesine ayrılabilir evrişimsel (GASE) modellerinin kategoriler için doğru pozitif (P) ve doğru negatif oran yüzdeleri (N) (True positive (TPR %) and true negative (TNR %) rate percentages of the fully (TE), standard (SE) and, dilated depthwise separable (GASE) models, with the minimum 80% true negatives condition and the minimum 99% true positive condition, for the categories)

	En az %80 Doğru Negatif			En az %99 Doğru Pozitif		
Kategori	TE	SE	GASE	TE	SE	GASE
P: Tümü	98.6	100.0	100.0	100.0	100.0	100.0
P: Ön yüz konuşma	100.0	100.0	100.0	100.0	100.0	100.0
P: Dönmüş yüz konuşma	91.7	100.0	100.0	100.0	100.0	100.0
N: Tümü	86.1	100.0	86.1	61.1	100.0	86.1
N: Kafa dönmeleri	100.0	100.0	91.7	91.7	100.0	91.7
N: Dudak yapıştırma-ısırma	87.5	100.0	75.00	62.5	100.0	75.0
N: Ağız açıp nefes alma	75.0	100.0	100.0	25.0	100.0	100.0
N: Gülümseme	100.0	100.0	100.0	75.0	100.0	100.0
N: Çene-ağız üzeri parmak	50.0	100.0	75.0	00.0	100.0	75.0
N: Yüz önü el kaydırma	75.0	100.0	75.0	50.0	100.0	75.0

5. Sonuçlar (Conclusions)

Bu çalışma, görsel konuşma sesi saptama için literatürde yaygın olarak kullanılan video kameralara verimlilik açısından daha üstün bir alternatif olarak olay kamerasının kullanımını ele almıştır. KST birçok uygulamada ön işlem olarak çalışması gerektiğinden, enerji verimliliği ve işlem karmaşıklığı kritik öneme sahiptir. Bu ihtiyaç doğrultusunda, olay kamerası sayesinde yüksek zaman çözünürlüğünde elde edilen ağız bölgesi olay yoğunluğu dizisini, evrişimsel sinir ağları ile temel alan yeni bir yöntem önerilmiştir.

Uzamsal boyut tamamen indirgendiğinden, son derece düşük hesaplama gereksinimi olan hafif modeller oluşturulmuştur. Verimliliği üst seviyelere çıkarabilen teknikleri ve tasarımları bulabilmek için farklı mimari tasarımlar karşılaştırılmıştır. KST problemini ele alan önceki bütün video kamerası veya olay kamerası çalışmalarından farklı olarak, başarım ve verimlilik birlikte kapsamlı bir şekilde değerlendirilmiştir.

Konuşma artikülasyonundan farklı yüz eylemleri durumundaki başarımları ölçebilmek için, veri kümesinde çeşitli yüz aktivitesi klipleri de dahil edilmiştir. Böylece, her bir farklı kategorideki yüz eylemleri karşısında KST'nin dayanıklılığı kategori özelinde de incelenmiştir. Konuşma dışı yüz aktivitelerinin olduğu durumlarda dahi yüksek başarımların elde edilmiş olması, hiçbir uzamsal bilgi kullanılmadan, yani sadece ayırt edici dinamik yoğunluk örüntü modellerinin öğrenilmesiyle, mükemmel seviyede saptama yapılabildiğini göstermiştir. Bu sonuç, ele alınan problemde, ağız bölgesindeki uzamsal bilginin tanıma için aslında gerekli olmadığını destekleyen bir bulgudur. Uzamsal örüntülerin modellenmesi, önemli derecede karmaşıklık ve büyük işlem yükü maliyeti getireceğinden, bunların gerekmeyecek olması verimlilik açısından çok önemli bir sonuçtur.

Zamansal alıcı alanı, sinir ağı derinliğine göre farklı hızlarda artıran aşağı-örnekleme, sabit evrişim genleştirme ve, üstel evrişim genleştirme teknikleri incelenmiştir. Deneysel sonuçlar, öncelikle, aşağı-örneklemenin hem yüksek başarım hem de yüksek verimlilik için gerekli olduğunu göstermiştir. Ancak, aşağı-örneklemeyi gerçekleştiren maksimum-havuzlama uygulamasının (SE), daha az işlem yükü ile çalışan adımlı evrişim aşağı-örneklemesinden (TE) önemli ölçüde daha yüksek başarım elde ettiği bulunmuştur. Bunun sebebi, maksimum-havuzlama ile daha kuvvetli öteleme değişmezliği elde edilmesi olabilir; çünkü, konuşma aralıkları son derece değişken olarak ortaya çıkar. Ayrıca, derinliğin çok fazla arttırılmasının ise (altı katmandan fazla) başarım ve verimliliği olumsuz etkilediği görülmüştür. Başarım-verimlilik dengesi açısından bakıldığında, aşağı-örnekleme ve maksimum havuzlama yapan altı katmanlı modelin (SE) mükemmel başarım sağladığı ancak, beş katmanlı fakat sabit oranla genleştirilmiş evrişim uygulayarak aynı zamansal alıcı alana erişen modelin (GSE), sadece %0.3'lük ufak bir başarım düşüşüyle, verimliliği iki kattan fazla artırdığı görülmüştür (1.57'den 0.75 MFLOPS'a). Verimliliği daha da artırabilmek amacıyla, evrişim çekirdeklerini kanal ekseni üzerinde ayıran derinlemesine ve farklı gruplamalı ayırma tasarımları karşılaştırılmıştır. Deneyler, verimliliği en yüksek seviyede artıran derinlemesine ayırmanın (GASE-Der), GSE'nin başarımını koruyarak işlem yükünü 0.30 MFLOPS'a indirdiğini, yani SE'nin beşte birinden daha az işlem gücüyle de KST yapılabileceğini göstermiştir. Dolaysıyla, verimliliğin kritik olduğu uygulamalarda GASE-Der modeli, SE'ye göre %0.3'lük ufak bir başarım düşüşüyle, tercih edilebilecek bir modeldir.

Bu çalışmada elde edilen sonuçlar, olay kamerası ile KST çözümünün çok verimli bir alternatif olabileceğini göstermiş olup bu alandaki başka araştırmaları motive edicidir. İleride ele alınabilecek bir konu, gerçek zamanlı uygulamalar için önemli olan gecikme süresi olabilir. Sabit bir zaman penceresi kullanılarak tespit yapıldığında, çıktı zamanı pencere merkezi olarak varsayılır, yani aynı sayıda geçmiş ve gelecek girdi örneği kullanılır. Bu durumda gecikme süresi zamansal alıcı alanla orantılıdır. Öte yandan, nedensel evrişim filtrelemesi yapılırsa, yani bütün girdiler geçmiş zaman örnekleri olursa, sıfır gecikme süresinde KST yapılabilir. Ancak bunun başarımı olumsuz etkilemesi beklenir; çünkü, konuşmanın belli bir bağlam süresi vardır. Dengeli bir çözüm, farklı oranda asimetrik zaman pencereli evrişim yapmak olabilir; yani fazla sayıda geçmiş örnek ama az sayıda gelecek örnekle çalışan evrişim modelleri tasarlanabilir. Dolayısıyla, sonraki bir çalışmada, gecikme süresini bu şekilde en aza indiren tasarımlar hedeflemek önemli bir araştırma konusu olacaktır.

Teşekkür (Acknowledgment)

Bu çalışma, Yaşar Üniversitesi Proje Değerlendirme Komisyonu (PDK) tarafından kabul edilen BAP112 no.lu ve "Nöromorfik Kamera ile Dinamik Yüz Analizi" başlıklı proje kapsamında desteklenmiştir.

Kaynaklar (References)

Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., Nayak, T., Andreopoulos, A., Garreau, G., Mendoza, M., Kusnitz, J., Debole, M., Esser, S., Delbruck, T., Flickner, M., Modha, D., 2017. A Low Power, Fully Event-Based Gesture Recognition System. CVPR2017, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA.

Araujo, A., Norris, W., Sim, J., 2019. Computing Receptive Fields of Convolutional Neural Networks. Distill, https://distill.pub/2019/computing-receptive-fields.

Ariav, I., Dov, D., Cohen, I., 2018. A deep architecture for audio-visual voice activity detection in the presence of transients. Signal Processing 142, 69–74.

Arriandiaga, A., Morrone, G., Pasa, L., Badino, L., Bartolozzi, C., 2021. Audio-Visual Target Speaker Enhancement on Multi-Talker Environment Using Event-Driven Cameras. ISCAS 2021, IEEE International Symposium on Circuits and Systems, Daegu, South Korea, May 22-28, 2021.

Bai, S., Kolter, J.Z., Koltun, V., 2018. Convolutional Sequence Modeling Revisited. ICLRW2018, 6th International Conference on Learning Representations - Workshop Track Proceedings, April 30 - May 3, 2018, Vancouver, BC, Canada.

Barua, S., Miyatani, Y., Veeraraghavan, A., 2016. Direct face detection and video reconstruction from event cameras. WACV2016, Winter Conference on Applications of Computer Vision, March 7-10, 2016, Lake Placid, NY, USA.

Berlincioni, L., Cultrera, L., Albisani, C., Cresti, L., Leonardo, A., Picchioni, S., Becattini, F., Del Bimbo, A., 2023. Neuromorphic Event-based Facial Expression Recognition. CVPRW2017, The IEEE/CVF Conference on Computer Vision and Pattern Recognition - Workshop Track., June, 2023, Vancouver, Canada, pp. 4108–4118.

Çubukçu, A., Kuncan, M., Kaplan, K., Ertunç, H.M., 2015. Development of a voice-controlled home automation using Zigbee module. In: 23nd Signal Processing and Communications Applications Conference (SIU). pp. 1801–1804.

Deng, Y., Chen, H., Liu, H., Li, Y., 2022. A Voxel Graph CNN for Object Classification With Event Cameras. CVPR2022, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, June 2022.

Gallego, G., Lund, J.E.A., Mueggler, E., Rebecq, H., Delbrück, T., Scaramuzza, D., 2018. Event-Based, 6-DOF Camera Tracking from Photometric Depth Maps. IEEE Trans. Pattern Anal. Mach. Intell. 40, 2402–2412.

Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D., 2022. Event-Based Vision: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 154–180.

Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D., 2019. End-to-End Learning of Representations for Asynchronous Event-Based Data, ICCV2019, The IEEE International Conference on Computer Vision, October 2019.

Ghaemmaghami, H., Dean, D., Kalantari, S., Sridharan, S., Fookes, C., 2015. Complete-linkage clustering for voice activity detection in audio and visual speech. Interspeech, Dresden, Germany, 2015.

Guy, S., Lathuilière, S., Mesejo, P., Horaud, R., 2020. Learning Visual Voice Activity Detection with an Automatically Annotated Dataset. ICPR2020, 25th International Conference on Pattern Recognition, January 10-15, 2020, Milan, Italy.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arxiv:1704.04861.

Kim, J., Hwang, I., Kim, Y.M., 2022. Ev-TTA: Test-Time Adaptation for Event-Based Object Recognition. CVPR2022, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, June 2022.

Korkmaz, Y., Boyacı, A., 2023. Hybrid voice activity detection system based on LSTM and auditory speech features. Biomedical Signal Processing and Control 80, 104408.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. NIPS2012, Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2012, December 3-8, 2012, Lake Tahoe, Nevada, USA.

Lenz, G., Ieng, S.H., Benosman, R.B., 2020. Event-based Face Detection and Tracking using the Dynamics of Eye Blinks. Frontiers in Neuroscience 14, 587.

Li, J., Li, J., Zhu, L., Xiang, X., Huang, T., Tian, Y., 2022. Asynchronous Spatio-Temporal Memory Network for Continuous Event-Based Object Detection. IEEE Transactions on Image Processing 31, 2975–2987.

Li, X., Neil, D., Delbruck, T., Liu, S., 2019. Lip Reading Deep Network Exploiting Multi-Modal Spiking Visual and Auditory Sensors. ISCAS 2019, IEEE International Symposium on Circuits and Systems, May, 2019.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. CVPR2015, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2015, Boston, USA.

Maqueda, A.I., Loquercio, A., Gallego, G., García, N., Scaramuzza, D., 2018. Event-Based Vision Meets Deep Learning on Steering Prediction for Self-Driving Cars. CVPR2018, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, June 2018.

Moreira, G., Graça, A., Silva, B., Martins, P., Batista, J.P., 2022. Neuromorphic Event-based Face Identity Recognition. ICPR2022, 26th International Conference on Pattern Recognition, Montreal, August 21-25, 2022, QC, Canada, pp. 922–929.

Neil, D., Pfeiffer, M., Liu, S.-C., 2016. Phased LSTM: Accelerating Recurrent Network Training for Long or Event-based Sequences. NIPS2016, Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, pp. 3889–3897.

Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y., 2019. Bringing a Blurry Frame Alive at High Frame-Rate With an Event Camera. CVPR2019, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 2019.

Paredes-Valles, F., de Croon, G.C.H.E., 2021. Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy. CVPR2021, The IEEE/CVF Conference on Conference on Computer Vision and Pattern Recognition, June 2021.

Patrona, F., Iosifidis, A., Tefas, A., Nikolaidis, N., Pitas, I., 2016. Visual Voice Activity Detection in the Wild. IEEE Transactions on Multimedia 18, 967–977.

Perot, E., de Tournemire, P., Nitti, D., Masci, J., Sironi, A., 2020. Learning to Detect Objects with a 1 Megapixel Event Camera. NIPS2020, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, December 6-12, 2020.

Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D., 2019. Events-To-Video: Bringing Modern Computer Vision to

Event Cameras. CVPR2019, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 2019.

Rethage, D., Pons, J., Serra, X., 2018. A Wavenet for Speech Denoising. ICASSP2018, IEEE International Conference on Acoustics, Speech and Signal Processing, April 15–20, 2018 Calgary, Alberta, Canada, pp. 5069–5073.

Ryan, C., O'Sullivan, B., Elrasad, A., Cahill, A., Lemley, J., Kielty, P., Posch, C., Perot, E., 2021. Real-time face & eye tracking and blink detection using event cameras. Neural Networks 141, 87–97.

Savran, A., Tavarone, R., Higy, B., Badino, L., Bartolozzi, C., 2018. Energy and Computation Efficient Audio-Visual Voice Activity Detection Driven by Event-Cameras. FG2018, 13th IEEE International Conference on Automatic Face & Gesture Recognition, May 15-19 2018, Xi'an, China.

Savran, A., Bartolozzi, C., 2020. Face Pose Alignment with Event Cameras. Special Issue: Sensor Systems for Gesture Recognition, Vol. 20, Issue 24, Article 7079.

Savran, A., 2023. Multi-timescale boosting for efficient and improved event camera face pose alignment. Computer Vision and Image Understanding, Vol. 236, 103817.

Savran, A., 2023a. Fully Convolutional Event-camera Voice Activity Detection Based on Event Intensity. ASYU2023, IEEE Innovations in Intelligent Systems and Applications Conference, October, 2023, Sivas, Türkiye.

Savran, A., 2023b. Comparison of Timing Strategies for Face Pose Alignment with Event Camera. In: 8th International Conference on Computer Science and Engineering (UBMK). pp. 97–101.

Schaefer, S., Gehrig, D., Scaramuzza, D., 2022. AEGNN: Asynchronous Event-Based Graph Neural Networks. CVPR2022, The IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, June 2022.

Shahid, M., Beyan, C., Murino, V., 2021. S-VVAD: Visual Voice Activity Detection by Motion Segmentation. WACV2021, Winter Conference on Applications of Computer Vision, January 3-8, 2021, Waikoloa, HI, USA, pp. 2331-2340

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. CVPR2015, The IEEE/CVF Conference on Conference on Computer Vision and Pattern Recognition, June 2015, Boston, USA.

Sharma, R., Somandepalli, K., Narayanan, S.S., 2019. Toward Visual Voice Activity Detection for Unconstrained Videos. ICIP2019, International Conference on Image Processing, September 22-25, 2019, Taipei, Taiwan.

Tan, G., Wang, Y., Han, H., Cao, Y., Wu, F., Zha, Z.-J., 2022. Multi-Grained Spatio-Temporal Features Perceived Network for Event-Based Lip-Reading. CVPR2022, The IEEE/CVF Conference on Conference on Computer Vision and Pattern Recognition, New Orleans, USA, June 2022.

Tulyakov, S., Bochicchio, A., Gehrig, D., Georgoulis, S., Li, Y., Scaramuzza, D., 2022. Time Lens++: Event-Based Frame Interpolation With Parametric Non-Linear Flow and Multi-Scale Fusion. CVPR2022, The IEEE Conference on Conference on Computer Vision and Pattern Recognition, New Orleans, USA, June 2022.

Wang, D., Xiao, X., Kanda, N., Yoshioka, T., Wu, J., 2023. Target Speaker Voice Activity Detection with Transformers and Its Integration with End-To-End Neural Diarization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Wang, Y., Du, B., Shen, Y., Wu, K., Zhao, G., Sun, J., Wen, H., 2019. EV-Gait: Event-Based Robust Gait Recognition Using Dynamic Vision Sensors. The IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, June 2019.

Wang, Y., Zhang, X., Shen, Y., Du, B., Zhao, G., Cui, L., Wen, H., 2022. Event-Stream Representation for Human Gaits Identification Using Deep Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 3436–3449.

Wrench, A., 2006. MOCHA-TIMIT, www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

Yu, F., Koltun, V., 2016. Multi-Scale Context Aggregation by Dilated Convolutions. 4th International Conference on Learning Representations, ICLR, San Juan, Puerto Rico, May 2016.

Zhang, X.-L., Wang, D., 2016. Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24, 252–264.

Zhang, J., Dong, B., Zhang, H., Ding, J., Heide, F., Yin, B., Yang, X., 2022. Spiking Transformers for Event-Based Single Object Tracking. CVPR2022, The IEEE Conference on Conference on Computer Vision and Pattern Recognition, New Orleans, USA, June 2022.

Zhu, L., Wang, X., Chang, Y., Li, J., Huang, T., Tian, Y., 2022. Event-Based Video Reconstruction via Potential-Assisted Spiking Neural Network. CVPR2022, The IEEE Conference on Conference on Computer Vision and Pattern Recognition, New Orleans, USA, June 2022.