MULTIMODAL REPRESENTATION LEARNING FOR SYNCHRONIZED SPEECH AND VIDEOS

by

Öykü Deniz Köse

B.S., Electrical and Electronics Engineering, Boğaziçi University, 2017

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University

2020

ACKNOWLEDGEMENTS

I would like to dedicate this thesis to both my grandfathers, Pirağa Aslan and Hüseyin Köse, with all my gratitude and respect. Their dedication to education inspired me through my academic life.

I can't express how grateful I am for having my family, especially my parents and sister. My parents and their loving hearts are the greatest treasure I have in this world, and I feel blessed to carry their legacy of living peacefully and with dignity. I am blessed with the support of my sister. She made me feel loved in the most difficult of times. I feel stronger for the upcoming obstacles, as I know she will be there for me and be my best friend throughout. Last but not least, I want to thank Pamuk Köse for the joy and love they brought to our lives.

I owe special thanks and regards for Prof. Murat Saraçlar; I could not create this thesis without his supervision. He was always patient with my failures and supportive through challenging times, which fired me up to continue. His dedication to his job made me see the delightful parts of academia, and shaped what I want to do in my life. I am also thankful to Prof. Levent Arslan and Prof. Engin Erzin for their valuable contributions and support.

I cannot express how lucky and thankful I am for having Can Gürsoy in my life. His respect and passion for his research inspired me to find my way, and I could not be where I am without his invaluable help. His fascinating personality and warm heart make me want to be a better person. I cannot wait for the pleasant and challenging times that I will share with you.

I would like to thank all my friends from within and without the department. I feel very lucky to meet all the members of BUSIM: You made my M.Sc. period was a pleasure. I am specifically thankful to Bolaji Yusuf, Alican Gök, and Dr. Batuhan Gündoğdu for their endless help. Especially, Bolaji Yusuf, it was a joyful experience

to share a cubicle with you. Ayça Gökçe, Berrak Aktürk, and Yılmaz Genç, your friendship makes this journey of life enjoyable and fulfilling. I am very grateful for each time, you turned all the frowns into smiles.

The work in this thesis was partially supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under the national scholarship programme for MSc students (2210-A).

ABSTRACT

MULTIMODAL REPRESENTATION LEARNING FOR SYNCHRONIZED SPEECH AND VIDEOS

The amount of multimedia data has been increased rapidly in recent years. While this data growth enables multimodal neural network based studies, it has also resulted in a need for efficient storage and retrieval systems for multimodal data. In this thesis, different data fusion schemes are examined to see the benefits of the use of different data sources. Proposed fusion schemes differ in their stages in which the data fusion is performed. Additionally, several representation learning methods are investigated for efficient data storage and retrieval systems. Representations are generated in such a way that they reflect the distance between the represented data segments according to a certain distance metric. A joint representation and distance metric learning scheme is also considered for a performance gain. Several deep neural network models are designed for representation learning and data fusion, and their performances are evaluated with the same-different word-discrimination and phone classification tasks, respectively. Experiments are performed on two different multimodal data sets; USC-TIMIT rtMRI and Signed Turkish broadcast news. Outcomes of the experiments show that the data fusion indeed brings a performance improvement over unimodal approaches, and performing fusion in earlier stages yields better results than fusing the data in later stages. Additionally, the proposed methods for the representation learning outperform the corresponding baseline systems in the same-different word-discrimination task. Therefore, generated representations of video and audio segments can be considered as an important step towards a fast cross-modal query-by-sign search system.

ÖZET

EŞZAMANLI KONUŞMA VE VİDEO İÇİN ÇOK KİPLİ GÖSTERİM ÖĞRENİMİ

Son yıllarda gelişen teknoloji ile birlikte çoğulortam veri miktarı hızlı bir şekilde artmaktadır. Bu artış çoklu veri kipleri için sinir ağları ile yapılan çalışmaları olanaklı kılsa da, verimli bir şekilde veri saklanımını ve geri getirimini önemli bir ihtiyaç haline getirmektedir. Bu tezde, çoklu veri kiplerinden faydalanmanın sağladığı yararları araştırmak için çeşitli veri kaynaştırma yöntemleri sunulmaktadır. Sunulan veri kaynaştırma yöntemleri sistemde kaynaşımın yapıldığı aşamaya göre farklılık göstermektedir. Ayrıca, veri saklanımı ve geri getirimi amacıyla veri altbirimleri için gösterim öğrenim sistemleri incelenmiştir. Bu gösterimler temsil ettikleri veri altbirimleri arasındaki belirlenmiş bir uzaklık metriğine göre hesaplanan uzaklığı yansıtacak şekilde oluşturulmuştur. Olası bir başarım artışı için bileşik gösterim ve uzaklık metriği öğrenimi problemi de irdelenmiştir. Gösterim öğrenimi ve veri kaynaştırma amacıyla derin sinir ağlarından faydalanılmıştır ve bu problemlerin başarım ölçümleri sırasıyla aynı-farklı kelime ayrımsaması ve sesbirim sınıflandırılmasıyla yapılmıştır. Deneylerde iki farklı veri kümesinden faydalanışmıştır: USC-TIMIT rtMRI ve Türk işaret dili haber bültenleri. Elde edilen deney sonuçları veri kaynaştırmanın faydalarını ortaya koyarken, aynı zamanda bu kaynaşımın erken safhalarda yapılmasının daha da iyi sonuçlar getirdiğini göstermektedir. Aynı zamanda, gösterim öğrenimi için sunulan yöntemler, aynı problem için daha önce kullanılmış temel sistemlerden fark edilebilir derecede daha iyi başarımlar sağlamıştır. Bu nedenle, video ve konuşma işaretlerinin altbirimleri için öğrenilmiş bu gösterimler çapraz kipli verilerin geri getirimi için önemli bir adım olarak kabul edilebilir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . v
ÖZET . . . vi
LIST OF FIGURES . . . ix
LIST OF TABLES . . . xi
LIST OF SYMBOLS . . . xii
LIST OF ACRONYMS/ABBREVIATIONS . . . xiv
1. INTRODUCTION . . . 1
2. RELATED WORK . . . 8
3. METHODOLOGY . . . 13
3.1. Preprocessing . . . 13
3.2. Hidden Representations for Phones . . . 14
3.2.1. Unimodal Approach for Audio . . . 14
3.2.2. Unimodal Approach for Video . . . 15
3.2.3. Multimodal Approach with Data Fusion at Output Stage . . . 16
3.2.4. Multimodal Approach with Cross-modal Deep Neural Network . 17
3.3. Representation Learning for Words . . . 20
3.3.1. Hidden Representation Learning . . . 22
3.3.2. Embedding Generation . . . 24
3.3.3. Joint Representation and Distance Metric Learning . . . 29
4. EXPERIMENTS . . . 33
4.1. Data Sets . . . 33
4.2. Experimental Setup . . . 34
4.2.1. Hidden Representations for Phones . . . 34
4.2.2. Hidden Representations for Words . . . 34
4.2.3. Word Embedding Generation . . . 35
4.2.4. Joint Representation and Distance Metric Learning . . . 35
5. RESULTS . . . 37
5.0.1. Hidden Representations for Phones . . . 37

5.0.2. Hidden Representations for Words 41
5.0.3. Word Embedding Generation 43
5.0.4. Joint Representation and Distance Metric Learning 49
6. CONCLUSION AND FUTURE WORK 51
REFERENCES 54

LIST OF FIGURES

3.1 An example of a non-clipped video frame before (left) and after (after) preprocessing. 13

3.2 Deep neural network used for phone classification using audio-only data. 15

3.3 Deep network used for phone classification using video-only data. . 16

3.4 Proposed multimodal deep network used for phone classification, which performs the data fusion at the output stage. 17

3.5 Proposed cross-modal deep network used for phone classification, which performs the data fusion through cross-connections. 19

3.6 First proposed end-to-end model for the cross-view same-different word discrimination task. 22

3.7 Cross-connection-aided, second proposed end-to-end model for the cross-view same-different word discrimination task. 24

3.8 Encoder network used to generate smaller dimensional rtMRI feature vectors. 25

3.9 The base network for embedding creation. 26

3.10 The presented network for joint representation and distance metric learning. 31

5.1 Saliency maps for different phones generated for the network described by Fig. 3.3. The upper left image corresponds to the phone 'V', the upper right one is for the phone 'Z', the lower left one is for the phone 'SH', and the last image corresponds to the phone 'K'. 40

5.2 Histogram of misclassified phone pairs using the system described by Fig. 3.3. 41

5.3 Saliency maps for video frames corresponding to the most frequently misclassified phone pair using the system described by Fig. 3.3. The left image represents the saliency map for the video frame corresponding to the phone 'Z', whilst the right image represents the saliency map for the video frame corresponding to the phone 'S'. 41

5.4 Precision-recall curves for different speaker and word length cases when the objective function is $\mathcal{L}_0$ for cross-view same-different word discrimination task. 47

LIST OF TABLES

3.1 Processing of the video sequence to enable information exchange with the audio sequence. 20

3.2 Processing of the audio sequences to enable information exchange with the video sequences. 21

5.1 Results for the broad phone classification task. 38

5.2 Phone classification accuracy results for the several models presented for this task. 39

5.3 Overall accuracies and error distributions on false positives and false negatives when the word pairs are uttered by the same subject or by different subjects for different proposed networks. 42

5.4 Words and word pairs that are frequently misclassified in the cross-view same-different word task. 44

5.5 AP for the unimodal same-different word discrimination task for different objective functions and baseline systems. 45

5.6 AP for cross-view same-different word discrimination task for different objective functions and DTW system. 46

5.7 The percentage of the errors when the speakers are the same for the best performing unimodal and multimodal systems. 48

5.8 AP values of all the proposed schemes for the creation of multitask embeddings. 48

5.9 AP values for the same-different sign discrimination task. 50

LIST OF SYMBOLS

a	Audio data sequences corresponding to words
a^+	Audio data sequence that has the same word label as anchor sample
a^-	Audio data sequence that has a different word label from anchor sample
$d(\cdot, \cdot)$	A specified distance metric
$\mathcal{D}(\cdot)$	Diagonal matching algorithm cost function for the subsequence Y_{sub}
$\mathcal{D}_{DTW}(\cdot, \cdot)$	Dynamic time warping alignment cost for the given input sequences
$f(\cdot)$	Neural network that creates embeddings for given audio inputs
$f(\cdot, \cdot)$	Overall output of the neural network designed for the joint representation and distance metric learning
$g(\cdot)$	Neural network that creates embeddings for given video inputs
$\mathcal{L}(\cdot, \cdot, \cdot)$	Triplet loss defined to train Siamese networks
$\mathcal{L}_0$	Triplet loss in which the anchor example is a video sequence and the corresponding positive and negative examples are audio sequences
$\mathcal{L}_1$	Triplet loss in which the anchor and negative examples are audio sequences and the corresponding positive example is a video sequence
$\mathcal{L}_2$	Triplet loss in which the anchor example is an audio sequence and the corresponding positive and negative examples are video sequences
$\mathcal{L}_3$	Triplet loss in which the anchor and negative examples are video sequences and the corresponding positive example is an audio sequence
$\mathcal{L}_{m0}$	Objective function that is the sum of both unimodal triplet losses and $\mathcal{L}_2$

$\mathcal{L}_{m1}$	Objective function that is the sum of unimodal triplet loss defined for the video and $\mathcal{L}_2$
$\mathcal{L}_{m2}$	Objective function that is the sum of unimodal triplet loss defined for the audio and $\mathcal{L}_2$
m	Margin value in the utilized triplet loss
P_i	The class probablity of the ith phoneme
$\mathcal{U}(\cdot,\cdot)$	Learned distance metric in the joint representation and distance metric problem
v	MRI video segments corresponding to words
v^+	MRI video segment that has the same word label as anchor sample
v^-	MRI video segment that has a different word label from anchor sample
X	OpenPose feature sequence of length M
Y	OpenPose feature sequence of length N
Y_{sub}	OpenPose feature subsequence of length M found applying diagonal matching algorithm on Y

LIST OF ACRONYMS/ABBREVIATIONS

AP	Average Precision
ASR	Automatic Speech Recognition
CCA	Canonical Correlation Analysis
CNN	Convolutional Neural Network
DTW	Dynamic Time Warping
EMA	Electromagnetic Articulography
HMM	Hidden Markov Model
LRCN	Long-Term Recurrent Convolutional Network
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficient
MRI	Magnetic Resonance Imaging
PReLU	Parametric Rectified Linear Unit
ReLU	Rectified Linear Unit
RGB	Red-Green-Blue
rtMRI	Real-Time Magnetic Resonance Imaging
TD	Time Distributed
TRT	Turkish Radio and Television Corporation
TSL	Turkish Sign Language
VCV	Vowel-Consonant-Vowel

1. INTRODUCTION

Data in several modalities has accumulated recently due to the increased social media use and technological advancements. People are posting images, videos and comments on several online platforms, which boosts the amount of available data in several modalities. Additionally, technological advancements make possible to record data in new modalities. For example, improvements in medical imaging lead to new valuable magnetic resonance imaging (MRI) [1], ultrasound imaging [2], electromagnetic articulography (EMA) [3], and X-rays [4] collections.

Data collections can provide different information for a common source. For instance, EMA and MRI recordings can both provide information for speech articulation. While all different data sources can be exploited individually, it is intuitive to claim that fusing all information that come from different sources directly improves system performance. This intuition stems from the notion that different sources can carry complementary information which can better characterize the system and indeed reflects the reality for certain cases. For instance, [5] shows that using multi-modal information coming from face image, fingerprint, and hand geometry results in a better performance for the person identification task, comparative to unimodal approaches. Another example may be given from the speech processing domain, where the reliability of a speech recognition system can be enhanced using the information that comes from lip movements [6].

Despite the idea of data fusion being intuitive, there are critical issues to be handled when making use of the concept. In the case of heterogeneous data fusion, one goal is to handle the process such that the drawbacks coming from one modality are inhibited by the other [7]. Furthermore, in order to maximize the performance gain, one also naturally seeks to maximally utilize the information coming from different modalities. These design criteria make the data fusion task non-trivial as much as it is potentially beneficial, which led several works in the literature to tackle potential issues on it [8–10].

While data collections in several modalities can help data fusion studies, multimodal data accumulation on online platforms has also put importance on efficient data storage and fast data retrieval systems. First step to create such systems is the creation of representative vectors for the original data subunits. Those representations should be generated for several data modalities, and they should reflect the distance between original represented subunits according to a certain distance measure.

Representation creation for the subunits of speech signal such as phonemes, words, sentences is an important research area in speech, natural language processing, and vision communities [11–14]. Embeddings are the projections of arbitrary or fixed dimensional data segments into a continuous space preserving the similarity between the original data segments. They are fixed and generally smaller dimensional vectors to allow efficient storage. In the literature, it is demonstrated that embedding-based studies perform better than sequence based methods (like dynamic time warping, DTW) with lower computational complexity in tasks involving the comparison of sequences [15–17]. Even though embeddings are utilized in several unimodal studies [16, 17], they can also provide a similarity measure between the subunits of different data modalities with a simple distance metric [18, 19].

This thesis is built around the representation learning problem which is studied on two different multimodal data sets; USC-TIMIT rtMRI data set [18] and Turkish Radio and Television Corporation (TRT) Signed Turkish broadcast news. USC-TIMIT rtMRI data set is one of the most popular and publicly available MRI collections. It consists of the rtMRI videos of the upper airway in the mid-sagittal plane and corresponding speech waveforms of five female and five male subjects for the sentences used in an earlier database, known as the MOCHA-TIMIT database [20]. This data set is used for the hidden representation learning experiments and analyses. In these experiments, the effects of different data fusion schemes on the created representations are examined. In addition to the hidden representation learning, this data set is also utilized in the experiments of word embedding generation. In addition to the representation learning studies on the USC-TIMIT data set, this problem is also analysed together with a distance metric learning with the use of TRT Signed Turkish broadcast news.

To investigate the benefits of data fusion on the hidden representations, unimodal and multimodal neural network models are presented in this thesis for phone classification task. For the unimodal phone classification task with 39 different phonetic classes, we independently train two different deep neural network models. One of the networks acquires data solely from the rtMRI videos whilst the other one considers the audio data as the only inputs. We obtain the classification accuracy results for these unimodal systems in order to demonstrate the improvements that data fusion can bring. In multimodal networks, both data modalities are utilized through the data fusion. We employ the data fusion concept on rtMRI and audio, through two different models we develop. The methods differ in the stage where the fusion is executed. In the first system, the fusion takes place after the independent generation of the video and audio features. On the contrary, the second system performs the fusion in an earlier stage, by generating the features of a data modality using the information coming from the other data source.

Other than the proposed approaches on the phone classification, we define a cross-view same-different word discrimination task on the USC-TIMIT rtMRI data set. In the defined task, audio and video-based data sequences describing words are taken as inputs, and the system outputs a decision whether or not those two sequences represent the same word in an end-to-end fashion. With this task, we investigate the possible interchangeable use of the audio and rtMRI video with each other. Results of this task demonstrate whether the smaller dimensional representations for different data streams can carry essential information for their sources, which corresponds to the words they represent in this task.

For the cross-view same-different word discrimination task, we propose two different neural network-based schemes to measure the performances with the use of hidden representations. In the first approach, representations of the words are independently created for different modalities by employing a long short-term memory (LSTM) based deep neural network. Afterwards, obtained representations are concatenated and fed to fully connected layers for the output. Similar to the use of earlier fusion on the phone classification task, we propose another scheme where the representations for a

data source are generated by considering the other one. With this task, we investigate the possible interchangeable use of the audio and rtMRI video with each other.

In the next step, word embeddings are generated for audio and rtMRI video units independently in such a way that the cosine distance between them reflects the similarity between original data sequences. To generate more compact feature vectors for the rtMRI video sequences that are inputs in this task, a correspondence autoencoder is designed. Since using the original frames' pixel values as feature vectors requires an important amount of memory and training time, new compact features allow faster training with decreased memory requirements. The correspondence autoencoder is trained with randomly chosen data sequence pairs that have the same or different word labels. After the generation of video feature vectors, a recurrent Siamese neural network is employed to create embeddings with the use of a triplet loss. The performance of the embeddings are measured with the unimodal and cross-view same-different word discrimination task.

For the cross-view same-different word discrimination task, embeddings are created jointly for both MRI videos and speech signal with the use of a Siamese neural network. Cosine distance between the generated embeddings reflects the similarity between original video and speech segments. Therefore, it can be said that sequences from different modalities are projected into the same space while preserving the original similarity. For these cross-view embeddings, several triplet loss functions are examined and their performances are compared.

After the word embedding generation for the unimodal and cross-view same-different word discrimination tasks independently, word embeddings that can be used in both discrimination tasks are created, and they are named as multitask embeddings. A single Siamese network is trained for the multitask embeddings, and multitask learning methods are utilized in the training process. These embeddings are created to allow a common representation reflecting the similarity between original data sequences independent of their modalities.

After proposed models for the representation learning problem on the USC-TIMIT data set, joint representation and distance metric learning is investigated by using Signed Turkish broadcast news. This data set is employed to work with different modalities that are not synchronous exactly. While the signs and speech have the same word order in the videos, their temporal information is not exactly the same. Speech signal found in the news is used to generate a transcription with the use of an automatic speech recognition (ASR) tool. Since signs and the words in the speech are not exactly synchronous, diagonal matching algorithm is applied to extract sign examples to be used in the training stage. Therefore, this study is important, because it does not require any manual annotations.

Extracted sign units through the diagonal matching algorithms are used to train the proposed neural network that learns the embeddings for the signs and a distance metric to be used with those embeddings in dynamic time warping (DTW) algorithm. The performance of the learned distance metric is evaluated with the unimodal same-different sign discrimination task, and it is compared to the performances of common distance metrics used in the DTW algorithm, such as Euclidean distance, or cosine distance.

The study on the broadcast news cannot be regarded as cross-modal, since performance evaluation is made by only employing the unimodal same-different word-discrimination task. However, speech signal found in the broadcast news is used to generate an alignment for the signs by utilizing ASR tools. Therefore, the benefits of the use of different data modalities are again obvious.

Overall, the main contributions of this thesis can be listed as follows:

- We find that the representation creation using information from the cross-modal source yields better results in both phone classification and same-different word discrimination tasks than performing the fusion after independent feature creations. Obtained results show that representations generated with the information coming from other modalities are more powerful.

- In the proposed early fusion schemes generated for both tasks, the residual connections are set between the data sequences with using LSTM units. In the introducing papers of these cross-modal connections [21, 22], these connections are only used for the information exchange between single data frames even when the inputs are time sequences. Therefore, to the best of our knowledge, our work is the first study proposing an enabling scheme for the cross-modal information exchange between data sequences. Proposed recurrent connections can be beneficial when working with asynchronous multimodal data.
- We created word embeddings for the speech and rtMRI videos by adopting unimodal and multi-view losses presented for text and speech pair. Word embeddings are generated in such a way that the cosine distance between them reflect the similarity between original data segments, which can enable faster unimodal and cross-view retrieval schemes. Additionally, we generated word embeddings that can be used in both unimodal and cross-modal applications, and these multi-task embeddings can provide more efficient storage schemes. Obtained results show the success of proposed methods, since our methods overperform DTW-based baseline systems. All these endeavours on the word embeddings resulted in a published conference proceedings paper in ICASSP 2020 [23].
- We presented a method that allows joint representation and distance metric learning for the sign units found in Signed Turkish broadcast news without any manual annotation. Our proposed approach provides also an alignment method for asynchronous data. Our representation and distance metric learning system is designed by taking inspiration from a work presented for keyword spotting [24]. Our experiments show that the distance metric learned with the proposed method overperform common distance metrics like cosine distance. This work on the Signed Turkish broadcast news led to an accepted conference proceedings paper in SIU 2020.

The rest of this thesis manuscript is organized as follows: Chapter 2 discusses the related work in the fields of representation generation and data fusion. In this chapter, other studies presented in the utilized data sets are also explained to clarify the differences of our study with them. Proposed unimodal and cross-view network

schemes for the data fusion and representation generation problems are presented in Chapter 3. Utilized data sets and the experimental setups are explained in Chapter 4. Chapter 5 presents the results of all considered tasks and the related discussions for acquired results. Finally, Chapter 6 concludes this thesis and presents possible future works.

2. RELATED WORK

Since USC-TIMIT rtMRI data set is one of the biggest collections of valuable MRI videos, there are many studies on it for different tasks and with varying goals [18]. Works including [25, 26] define acoustic-to-articulatory inversion mappings to generate articulatory videos, given audio signals are provided. Note that, these studies do not seek to use different modalities of the data together, but rather provide an inversion between the two modalities with a goal of reducing the efforts needed to produce articulatory videos. Furthermore, other studies, including [27–29], exploit the data coming solely from the rtMRI videos, and explore air-tissue boundary segmentation methods to provide a better examination for the rtMRI videos. In [30], a deep learning based approach is employed for the estimation of vocal tract area function utilizing the rtMRI videos. Although it does not use the USC-TIMIT dataset, [31] is also worth mentioning due to its utilization of a similar but smaller rtMRI dataset called as USC Speech and Vocal Tract Morphology MRI Database [32]. Said database consists of rtMRI videos and their corresponding audio data for 54 different vowel-consonant-vowel (VCV) sequences, and is made use by [31] for VCV classification through the use of a Long-term Recurrent Convolutional Network (LRCN). [31] can also be classified as a unimodal work, since it only utilizes the data coming from the rtMRI videos.

There are also several works in the literature that utilize both modalities of the data in the USC-TIMIT database [33,34]. One of such works, [33], tackles the phonetic recognition task by generating a transformation matrix between the video and audio using canonical correlation analysis (CCA). In the study, rtMRI videos are assumed to be available only at the training stage and the test results are presented using only the audio data. From the multi-modal perspective, the findings of [33] suggest that by utilizing the information coming from videos in the training stage, phone error rate (PER) for the phone recognition task can be reduced. In [34], broad phonetic class recognition task is considered on a total of five classes (vowels, stops, fricatives, nasals, and silence). The study uses both rtMRI videos and speech waveforms and finds a satisfactory representation of the data with an information theoretical approach.

However, the approach presented in [34] requires manual annotations for each rtMRI videos and the performed classification is only a broad one.

Several studies take a similar approach to the proposed early data fusion scheme in this thesis on different data sets and for different applications. Early data fusion with the use of residual connections is proposed initially for the inputs from the same source (images) [21]. The study shows the advantages of the information exchange through cross connections between RGB channels of the images. This early fusion idea is also presented with the cross connections set between different data modalities (video and audio) [22], and the resulting network is called as a cross-modal deep neural network. However, these connections are utilized for the information exchange between single data frames even when the inputs to the networks are time sequences in [22].

Representation generation for the word units is a hot topic in speech processing and natural language processing fields. In one of the earlier works on acoustic word embedding generation, DTW distances to several word templates are used to represent the acoustic segments [35]. Recurrent neural networks are also utilized to represent acoustic data sequences corresponding to words for a simple query-by-example search [36]. For the purpose of acoustic word embedding learning, a Siamese convolutional neural networks (CNN) is also trained with a designed classification cost [16]. The Siamese network is trained in such a way that if the the speech segments have the same label, the embeddings created for them are close to each other for a distance metric. To generate embeddings that reflect the similarities between the original data sequences, a hinge loss is employed with a margin parameter in the study. Acoustic word embeddings are also used for the speech recognition task [37] by using a neural network that is trained with a triplet loss.

Apart from the unimodal studies on representation learning for acoustic word units, cross-view approaches have also been utilized for the same problem. In the literature, most of the proposed schemes for cross-view representation learning are based on the use of audio and text modalities [18, 38]. Since textual data also provides the semantic information, it can help acoustic embeddings to capture the semantic sim-

ilarities between the words. Furthermore, as textual data is a completely noiseless information source, its use generally improves the representation power of the acoustic word embeddings. Apart from multimodal studies that are based on audio-text, audiovisual data is also used to learn word embeddings with the employment of deep neural networks [39–41].

In recent studies, representation generation for the audio and video segments are also investigated for cross-modal applications [19, 42–46]. In [19], a CCA based cost function is designed to learn representations for text, video and audio sequences with the use of deep neural networks. Representation generation for the fixed-length video and audio segments is also presented as a binary classification problem [43]. In [43], temporal labeling is utilized, and a network is trained to discriminate whether the input sequences are from the same time window or not. In this study, audio-visual temporal synchronization task is utilized to measure the overall performance. Since cross-entropy loss results in a failure in the training of the network, the study presents a designed contrastive loss. Additionally, [43] demonstrates the advantages of the employment of curriculum learning [47] in the training stage, while choosing negative examples. While [43] presents a method to generate representations for the synchronized audio and video segments, there are other studies working on the unsynchronized data for the same purpose [45]. [45] utilizes from the generated representations for both audio and video segments for two different purposes: to identify the visual object that distinguishes the event, and to find the temporal information for a certain event. The defined tasks can be interpreted as multiple instance learning problems, and a deep neural network is utilized to present a method for the defined tasks in this study. In [46], embeddings for the audio and video frames are generated such that if the input pairs are from the same video, the cosine distance between them is smaller compared to the case where input audio and image are sourced from different videos. In this study, a combination of a contrastive and cross-entropy loss is presented to tackle with the representation generation problem. In [42], embeddings for the audio and video units are created for the cross-view retrieval and audio-visual correspondence tasks. Embeddings are generated with a deep neural network designed for binary classification, and the inputs are fixed-length audio and corresponding video frames. In this study, tem-

poral alignment of the audio and video is used for self-supervision. While this study has similarities with our proposed systems for the embedding learning for speech and MRI video units, it generates representations for fixed-length data segments, not for words or any other speech units carrying semantic information. Additionally, audio data used in that study consists of general sounds not only speech.

Besides our study on a Signed Turkish data set, numerous works have been presented for different applications on several sign languages because of their importance for the deaf and mute communities. Several labeled data sets are generated to help sign language recognition studies, even though their creation are expansive due to required manual annotations [48, 49]. Multimodal data sets like broadcast news [50], television shows [51], and movies [52] are also utilized for the sign language studies. In the mentioned data sets, other data modalities like text or speech can also convey the same information that signs carry.

There are many studies presented on the multimodal sign language data collections. In [53], two independent Hidden Markov Models (HMMs) are trained with the information of hand gestures and other information sources for the sign language like facial expressions, head movements, and body postures. Trained HMMs are utilized in the sign recognition task. Another study [54], utilizes from speech signal and sign language videos to present a system that enables the translation between the spoken and sign languages. The study utilizes deep neural networks and attention mechanism to tackle the translation task. In another study [52], subtitles of the movies are utilized to provide an alignment for the corresponding sign sequences to automatize the annotation of the sign language videos. In another study [55], for the same purpose, multiple instance learning methods are applied on television shows consisting of sign language and the corresponding subtitles. For the sign classification task, multitask learning methods are utilized again with the use of subtitles of sign language videos [51, 56]. In another study, texts and speech signal found in the broadcast news created for the deaf people are utilized to generate a query-by-sign search system [57]. In [58], texts that describe the signs are used to train a deep neural network for the classification of sign examples that are not found in the training set.

Apart from universal sign language studies, many studies are also presented on Turkish Sign Language (TSL). In one of the studies, HMMs are used for a sign retrieval system [57]. For the purpose of sign retrieval, sequence alignment methods are also utilized [59]. Several studies use DTW algorithm to generate a query-by-sign retrieval system for TSL [50, 60]. Additionally, there are several dictionaries and educational tools to improve TSL [61, 62].

3. METHODOLOGY

For the proposed methods in this thesis, data in several modalities are used in the experiments. Utilized MRI videos, sign language videos, and speech signal require specific preprocessing operations, and these operations for different modalities are explained in Section 3.1. In the next sections, designed deep neural networks and utilized algorithms for the data fusion and representation generation problems are explained in detail to clarify the contributions of this thesis.

3.1. Preprocessing

In order to make it possible to work with the rtMRI videos, a preprocessing step is essential due to the noisiness of the video frames. For this purpose, the steps proposed in [63] are followed, and the noise found in the video frames is reduced noticeably. In addition, video frames are cropped before they are fed to the network, since some pixels in the frames are consistently black and do not carry any information of the vocal tract. Specifically, left-most four columns and upper-most four rows of each video frame are clipped out, resulting in 64×64 pixel frames. In order to demonstrate the effect of the preproccesing step, an original video frame and the corresponding processed one (before the clipping stage) can be seen in the Fig. 3.1.

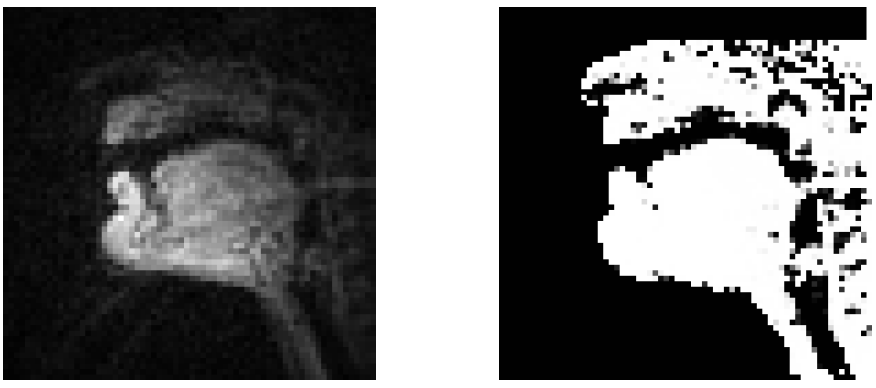


Figure 3.1: An example of a non-clipped video frame before (left) and after (after) preprocessing.

For the speech signal found in the rtMRI videos, features representing the data are extracted by shifting a window over the speech signal. The window length and shift are specified as 25 ms and 10.8 ms, respectively. Each frame is parametrized with 13 dimensional standard Mel-frequency cepstral coefficients (MFCCs). Shift value is chosen as 10.8 ms, since it makes four audio frames span a time interval of 43.2 ms, which is the time interval of a single rtMRI video frame. Therefore, one frame of rtMRI videos corresponds to four frames of audio data.

For our study on Signed Turkish broadcast news, skeleton features are used to represent the signs in the videos. OpenPose [64] is utilized to extract those features for upper body and both hands. OpenPose is a tool that is based on deep neural networks, and provides real-time pose estimations in two dimensional space. In this study, $2x50$ dimensional feature vectors (21 joints for each hand and 8 joints for upper body and all considered joints' x and y axis positions) are generated for each sign frame. Those vectors are reshaped into 100 dimensional feature vectors to be used in the models presented for the processing of Signed Turkish videos.

3.2. Hidden Representations for Phones

In this section, unimodal and multimodal models are presented for the phone classification task defined on USC-TIMIT data set. Unimodal systems use the information coming from only audio or MRI videos for this task. In multimodal schemes, both data modalities are utilized. Those systems are presented to investigate the performance gain that data fusion can bring on the phone classification task.

3.2.1. Unimodal Approach for Audio

For the task of phone classification on rtMRI-TIMIT dataset, one way is to solely utilize recorded speech as the data source. In order to extract the information coming from audio, a deep neural network is employed. The input sequences for each phone are extracted using the transcripts provided with the data set. Each audio frame is represented with a 13-dimensional MFCC feature vector. After the preprocessing

procedure that is mentioned in Section 3.1 is applied on the raw data, the input is fed to the recurrent neural network. The network outputs the class probabilities for each phone.

The considered recurrent neural network consists of two LSTM units followed by five fully-connected layers with 1024, 512, 256, 128, and 39 numbers of nodes, respectively. The subnetwork composed of fully-connected layers are named DenseNet, and referred to with this name throughout this thesis. This scheme is preferred, since a similar network model is used in all models presented for the phone classification task after representative features of the raw inputs are created. Number of layers and hidden layer node numbers are determined after several experiments, in which employing smaller networks results in lower classification accuracies and bigger networks do not bring significant performance gain and add computational burden. After the input layer, a batch normalization is performed in order to increase the generalization power of the network. After each hidden layer, a dropout with $p = 0.1$ is applied. The details of the overall network is presented in Fig. 3.2.

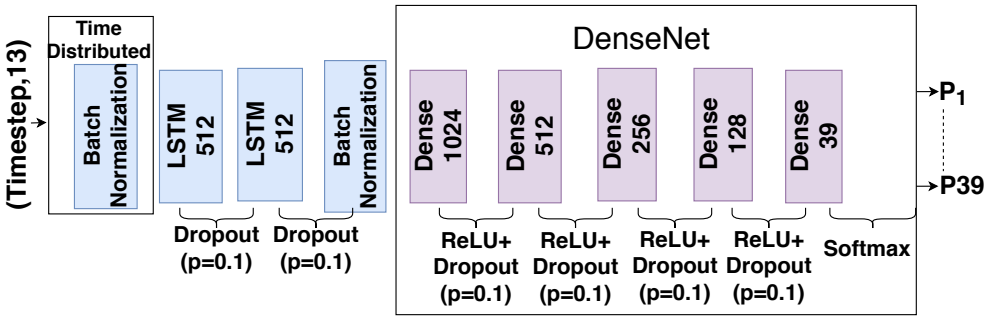


Figure 3.2: Deep neural network used for phone classification using audio-only data.

3.2.2. Unimodal Approach for Video

In order to investigate the information that video frames carry for phonetic discrimination, a unimodal system that inputs video frames for the phone classification task is also designed in this study. To accomplish this task, a deep convolutional network followed by LSTM units and fully connected layers is trained. All operations that are applied before feeding the video frames into the network are as described in Section 3.1. For each phonetic class, video frame sequences are extracted according to

the time boundaries provided by the transcription of the speech. As the output, the network predicts the probability for each phonetic class.

The deep network presented here is designed by taking inspiration from the deep convolutional networks employed in [27] and [28]. In our system, the first part of the network is composed of the convolutional layers which are named ConvNet1 and ConvNet2. ConvNet1 and ConvNet2 are jointly referred to as ConvNet throughout the thesis. The output of ConvNet is fed into the LSTM units to utilize sequential information. After LSTM units, DenseNet is employed for the classification task. The overall network topology is presented in Fig. 3.3.

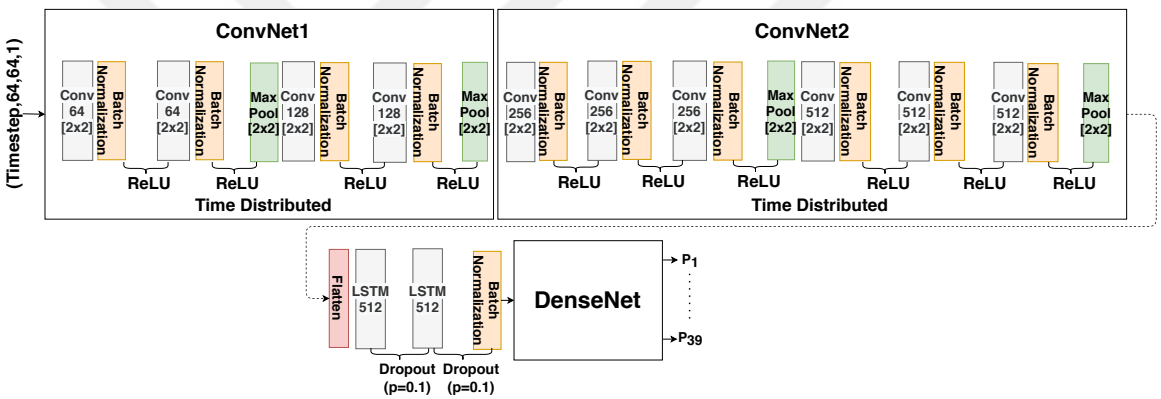


Figure 3.3: Deep network used for phone classification using video-only data.

3.2.3. Multimodal Approach with Data Fusion at Output Stage

After the description of the designs of the employed unimodal systems, we move onto the description of the multi-modal approaches that are proposed for the phone classification task. Firstly, noting the main objective of exploiting the improvements audio-video data fusion can bring, a deep neural network that unifies the models presented in the Sections 3.2.1 and 3.2.2 is designed. In the model, the representations of audio and video data are first generated through separate neural networks. Afterwards, the obtained representations are concatenated to utilize the information different modalities bring in an end-to-end system. It can be inferred that for the considered system, the data fusion is performed at the output stage of the model. The proposed system takes two sets of inputs corresponding to each modality, in their respective forms described in Figs. 3.2 and 3.3. Overall, the proposed model is shown in Fig. 3.4.

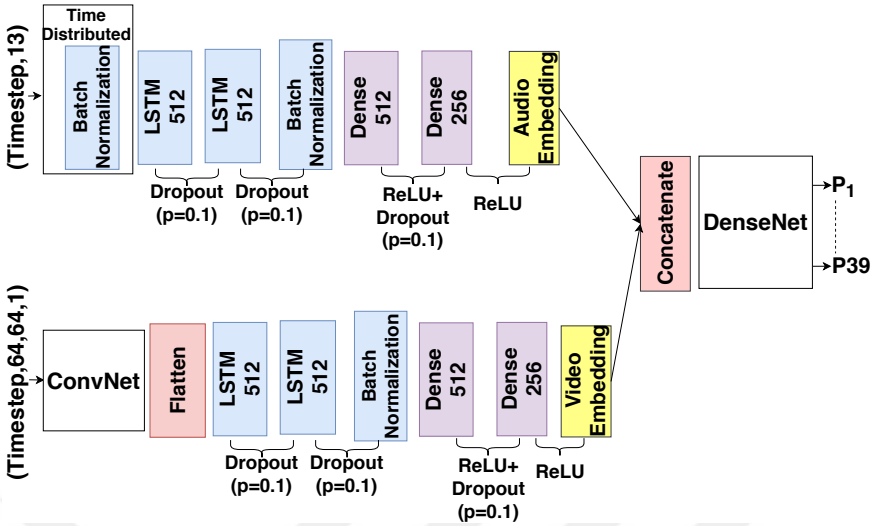


Figure 3.4: Proposed multimodal deep network used for phone classification, which performs the data fusion at the output stage.

3.2.4. Multimodal Approach with Cross-modal Deep Neural Network

In the network presented in Fig. 3.4, information coming from video and audio are fused after generating embeddings. In this context, embeddings refer to the smaller dimensional data representations carrying necessary information coming from each data source, for the given task. Recalling that embeddings are generated independently for each data source, it can be emphasized that for the system described by Fig. 3.4, fusion is executed after the feature extraction process is over. However, these smaller dimensional representations can also be acquired by utilizing the information coming from other data modalities. In other words, data fusion can be handled in an earlier stage of the process, resulting in a better utilization of the multi-modality concept. Several studies take a similar approach on different datasets and for different applications [21, 22]. However, these connections are utilized for the information exchange between single data frames even when the inputs to the networks are time sequences in other studies. Since the defined inputs have multiple frames corresponding to phones or words in this study, we extend the use of cross connections to sequential data. To our knowledge, proposed sequential residual connection blocks are used for the first time between cross-modal data sequences.

In our work, a deep neural network is designed to observe the effects of early fusion through cross-connections for the phone classification task. One of the difficulties of employing such a system is the mismatch between the timesteps of audio and video input vectors. Since some of the connections are set between input frames, those connected frames should contain the information coming from the same time interval. To solve this issue, each phone is represented with 12 audio frames and 3 video frames. These frames are selected by centering the frame found in the middle of the time interval provided by the transcript. Then, the audio input is reshaped in the form of 3 frames, each having 52 dimensional features, solving the mismatch between the timestep values of audio and video.

In order to set the connections between different data modalities (hence between data with different dimensionalities) by allowing information exchange between different streams, several transformations are performed on the input sides of the connections. Transformations applied at connection branches are based on the 2D-to-1D, and 1D-to-2D transformations proposed in [22]. Those transformations are modified in this study to allow information exchange between audio and video sequences with the aim of utilization of temporal relationship. Overall, the network scheme, alongside the mentioned transformations, can be seen in Fig. 3.5. The number of parameters of the cross-modal deep neural network (approximately 11 million) is similar, even lower, to the number of parameters of the network presented in Fig. 3.4 (approximately 12 million) to compare their performances fairly. Additionally, it can be noticed that the LSTM units found in the network presented in Fig. 3.4 are changed to fully-connected layers in the cross-modal deep neural network. While LSTM units lead to a better performance for the network presented for the fusion at the output stage, fully-connected layers work better with the cross-modal deep neural network. One thing to note in Fig. 3.5 is that the information exchange is provided from the raw input data of the cross-modality. This decision is inspired by the findings of [22], where the experimental results support the superiority of connecting the cross-modal information in its raw form, rather than its intermediate representations, with the other data source.

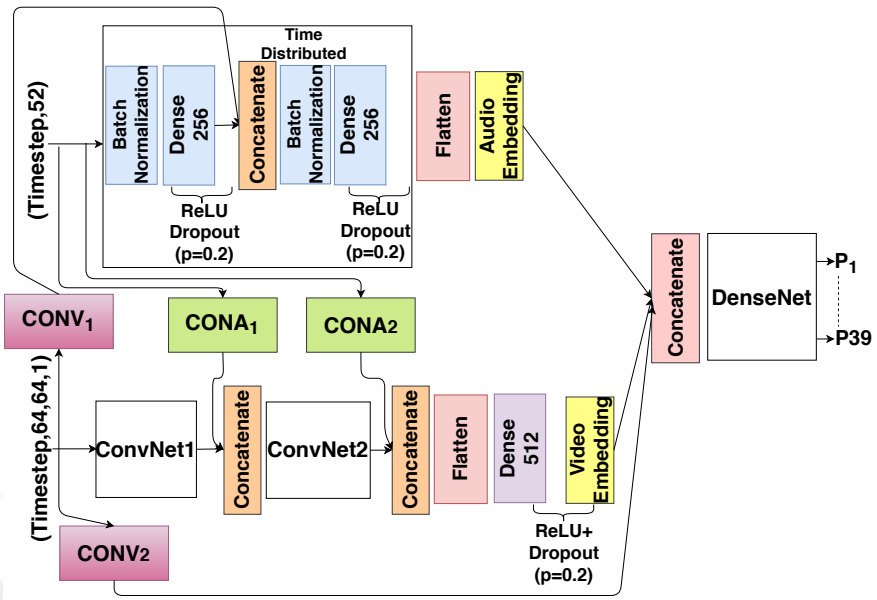


Figure 3.5: Proposed cross-modal deep network used for phone classification, which performs the data fusion through cross-connections.

From the representation in Fig. 3.5, four types of cross-connections can be observed. Among the implemented connections, two are from video-to-audio (called CON_{V1} and CON_{V2}), whilst two are from audio-to-video (called CON_{A1} and CON_{A2}). The video-to-audio connections are presented in Table 3.1. "TD" is used as an abbreviation for the time distributed layer in Table 3.1, and is employed in the remaining tables throughout the thesis, when necessary. Additionally, "SEQ" abbreviation in the LSTM units is used to imply that all hidden states corresponding to all time steps in the input are used in the next layer. If there is no "SEQ" abbreviation, only the last hidden state is transferred to the next layer. It can be seen from the steps described by Table 3.1 that after the applications of batch normalization and max-pooling operations on each time step, CON_{V1} and CON_{V2} differ in their latter operations. The main reason behind this design is that there is a difference in the stage where the information coming from video sequence is provided to the network. In both connections, LSTM units are used to determine the dimensionality, and also to utilize the temporal information found in the input sequences. Upon completion of these steps, the 2-D data streams at hand are successfully converted into 1-D using the aforementioned operations. Lastly, parametric rectified linear unit (PReLU) activation function [65] in conjunction to a dropout of with $p = 0.4$ is utilized. The choice of PReLU over

rectified linear unit (ReLU) is due to ReLU's nature of outputting zero for negative values and not permitting the network to utilize all coming information. As also aptly described by [22], PReLU yields a better cross information utilization in this regard.

Table 3.1: Processing of the video sequence to enable information exchange with the audio sequence.

$\mathbf{CON}_{V1}$	$\mathbf{CON}_{V2}$
TD(Batch Normalization())	TD(Batch Normalization())
TD(MaxPool $[2 \times 2]$)	TD(MaxPool $[2 \times 2]$)
TD(Flatten())	TD(Flatten())
LSTM(256,SEQ)	LSTM(256)
TD(PReLU())	PReLU()
TD(Dropout($p = 0.4$))	Dropout($p = 0.4$)

For the connections corresponding to audio-to-video transformations (CON_{A1} and CON_{A2}), the applied processes are presented in Table 3.2. Similar to the presented approach in video-to-audio, it can be observed from Table 3.2 that after the batch normalization layer, LSTM units are used to transform data in desirable dimensions and again to use the temporal relation found in the input sequences. Afterwards, 1-D audio data is transformed into the same form of the 2-D video data with the help of a deconvolution operation. Lastly, activation and dropout layers are selected the same as those used in video-to-audio transformations.

3.3. Representation Learning for Words

Representation learning is an important research problem, since it can enable efficient data storage and fast data retrieval systems. In this thesis, several unimodal and multimodal systems are presented to deal with this problem. These models are evaluated on unimodal and cross-view same-different word discrimination tasks. In this task, the sequences representing the words in same or different modalities are discriminated on whether they correspond to the same word label or not. The task is mainly utilized to examine how much information can be extracted from the repre-

Table 3.2: Processing of the audio sequences to enable information exchange with the video sequences.

$\mathbf{CON}_{A1}$	$\mathbf{CON}_{A2}$
TD(Batch Normalization())	TD(Batch Normalization())
LSTM(49,SEQ)	LSTM(1,SEQ)
Reshape$((3, 7, 7, 1))$	Reshape$((3, 1, 1, 1))$
TD(Deconv$([8 \times 8, 128])$)	TD(Deconv$([1 \times 1, 512])$)
TD(Batch Normalization())	TD(Batch Normalization())
TD(PReLU())	TD(PReLU())
TD(Dropout$(p = 0.4)$)	TD(Dropout$(p = 0.4)$)

sentations corresponding to different data sequences for their common sources, which are the labels of the words they represent in this case. The performance of the same-different word discrimination task can be a sign for many other applications, such as query-by-example search in unimodal or cross-view settings.

For the representation learning problem, two different deep neural networks are proposed as a first step, with their main difference lying in their utilization of sequential cross connections. In these models, inputs are the data segments from different modalities (rtMRI and speech) and the output is the decision for the cross-view same-different word discrimination. Representation generation and word discrimination decision are handled in an end-to-end system in these models. In these methods, generated representations are the hidden layer outputs of the designed networks.

In the next step, embeddings for MRI videos and speech segments are generated such that the cosine distance between the embeddings reflects the similarity between the original data sequences. In this part, embeddings are created for both unimodal and cross-view word-discrimination tasks. Additionally, mutli-task embeddings to be used in both unimodal and cross-view settings are generated.

In the final step, a deep neural network is designed for the joint representation and distance metric learning. Unimodal same-different sign discrimination task is utilized to evaluate the proposed model. In this task, the video segments corresponding to signs are discriminated on the labels of the signs. Results are acquired on the Signed Turkish broadcast news.

3.3.1. Hidden Representation Learning

In the first proposed system for hidden representation learning, a network similar to the one presented in Fig. 3.4 is exploited to extract the features representing the words from audio and video separately. The obtained features are then concatenated and connected to a series of fully-connected layers that in the end, output whether or not the input sequences belong to the same word label. All said operations are handled in an end-to-end fashion. LSTM units are utilized in this proposed network, which are employed to catch the temporal information coming from the input sequences. The overall topology of the network is presented in the Fig. 3.6.

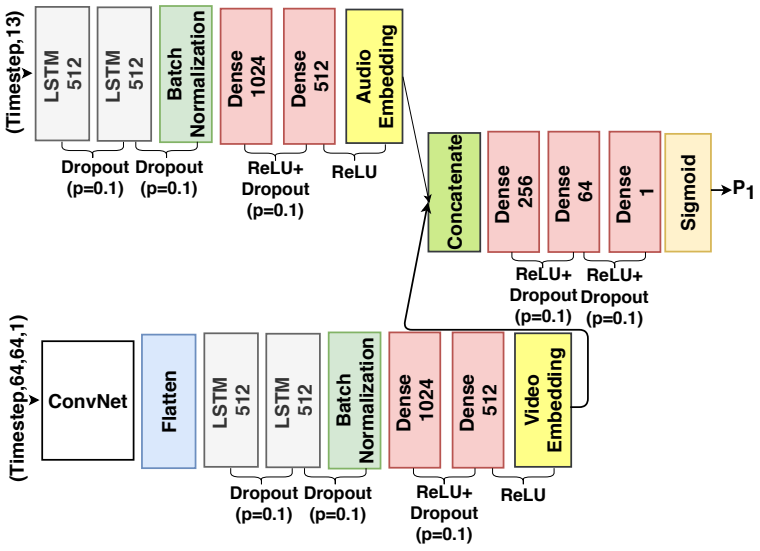


Figure 3.6: First proposed end-to-end model for the cross-view same-different word discrimination task.

The better performance obtained by the system in Fig. 3.5 over the one in Fig. 3.4 on the phone classification task signs that the representations of the video and audio data carry more information when cross-connections are implemented between the

networks that extract those representations. If these embeddings represent the data better, then it can be inferred that they can also provide a better word discrimination performance. Motivated by this implication, the concept of cross-connections is exploited once again here to propose another system for the defined cross-view same-different word discrimination task.

The first challenge before employing such a system that utilizes cross-connections is the difference between the timesteps of audio and video input vectors. As also mentioned in Section 3.1, each video input vector spans a time interval that is covered by four audio input vectors. Furthermore, the networks taking the audio and video inputs need the inputs to have same dimensionality in terms of the time axis in each batch, whereas the input sequences corresponding to different words do not have the same length. For these two reasons, video input vectors are zero padded in order to ensure the same time-dimensionality, which is the maximum timestep length in each batch. A zero padding process is also applied to the audio inputs in order to make their timestep length the quadruple of the timestep length of the corresponding video input vectors. In the next step, the audio input vectors are reshaped as vector sequences with the same timestep length as video inputs, and each audio vector has the dimension of 52.

Overall, the proposed cross-connection-based deep neural network is presented in Fig. 3.7. The structure of the network can be thought of as the unification of the networks presented in Figs. 3.5 and 3.6, where the connections presented in Fig. 3.5 are utilized in a similar manner and the general network topology is adopted from the system shown in Fig. 3.6. The transformation operations within the connection blocks are the same with the ones explained in Subsection 3.2.4. Again, it should be noted that the networks presented in Fig. 3.7, and Fig. 3.6 have similar number of parameters (approximately 13 million) for a fair performance comparison.

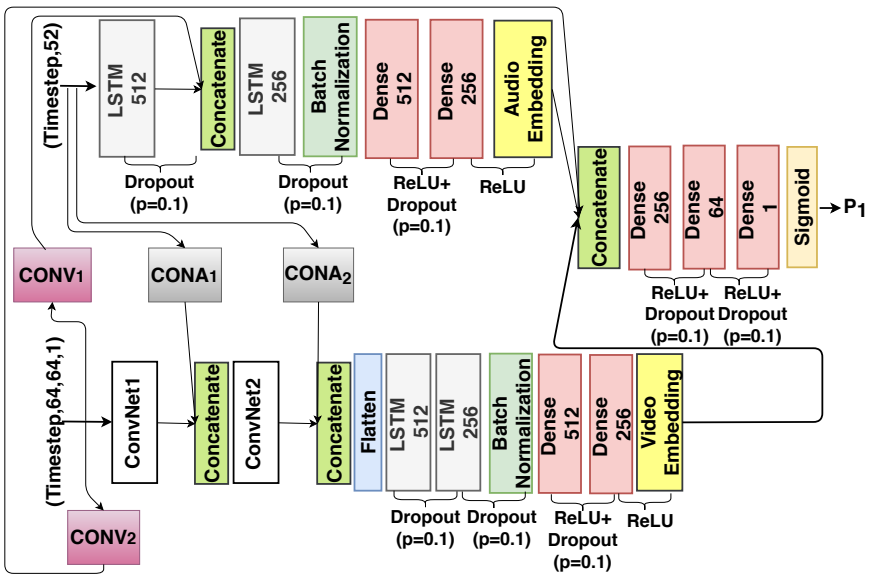


Figure 3.7: Cross-connection-aided, second proposed end-to-end model for the cross-view same-different word discrimination task.

3.3.2. Embedding Generation

After hidden representation generation, embeddings for the MRI videos and speech segments are generated in such a way that the cosine distance between the embeddings reflects the similarity between original data sequences. The main purpose of the generation of these embeddings is to show that even considerably noisy MRI videos and corresponding speech units can be represented and retrieved with a simple distance metric. Presented methods can be utilized for speech applications like query-by-example search or spoken term discovery in both unimodal and cross-view settings.

In this section, rtMRI videos are represented with features generated through a correspondence convolutional autoencoder to limit memory use and training time requirements. The autoencoder is inputted with the pixel values of MRI video segments corresponding to words and encodes them to denser representations. The convolutional autoencoder is trained to generate the pixel values of the video frame sequences having the same word label as corresponding input sequences have. Video segment pairs with the same word label are chosen randomly. The proposed model in [27] is utilized to design the autoencoder presented in this study, since both networks are trained on the

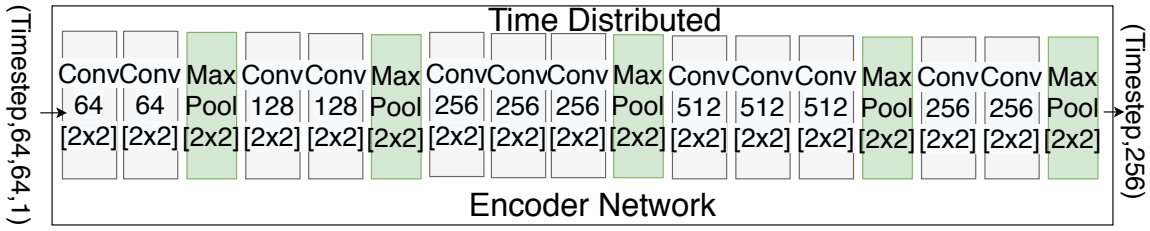


Figure 3.8: Encoder network used to generate smaller dimensional rtMRI feature vectors.

USC-TIMIT data set. Utilized encoder model can be seen in Fig. 3.8. After every convolutional layer presented in Fig. 3.8, batch normalization and ReLU activation are applied. Decoder part of the correspondence autoencoder is the exact symmetric version of the presented encoder model. It should be noted that the use of a regular autoencoder that is trained to generate the input vectors it takes is also investigated. However, feature vectors obtained with the correspondence autoencoder perform better than the representations acquired with the regular autoencoder in the defined tasks.

In the unimodal settings, word embeddings are generated for speech or rtMRI videos independently. The general topology of the network for unimodal word embeddings is designed by taking inspiration from the method utilized in [17]. A Siamese network is utilized to learn unimodal word embeddings. In the proposed method, the Siamese network inputs three different data sequences, where two of them share the same word label (a, a^+) or (v, v^+), and the other one has another word label (a^-) or (v^-) for audio and rtMRI videos, respectively. The network is trained to jointly learn the embeddings for all of the three inputs. Audio inputs a are composed of 13 dimensional MFCC feature sequences, and video inputs v consist of 256 dimensional feature vector sequences that are generated by the correspondence autoencoder. The Siamese network is trained in such a way that the cosine distance between the embeddings corresponding to (a, a^+) or (v, v^+) is smaller than the cosine distance between the embeddings of (a, a^-) or (v, v^-) by a margin m. The utilized objective function in

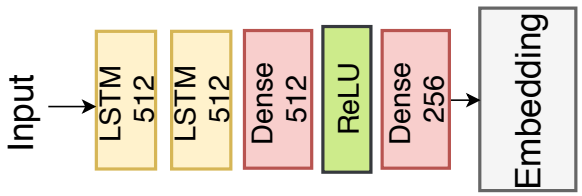


Figure 3.9: The base network for embedding creation.

the training stage for audio is as follows:

$$\mathcal{L}(f(a), f(a^+), f(a^-)) = \frac{1}{N}\sum_{n=1}^{N} \max(0, m + d(f(a_n), f(a_n^+)) - d(f(a_n), f(a_n^-))), \tag{3.1}$$

where f is the base network used to generate all embeddings jointly, m is the margin, and $d(.,.)$ is a distance metric which is selected as cosine distance in this thesis. a_n is used to represent the feature sequence corresponding to the nth audio word sample, and a_n^+, and a_n^- are for the nth positive, and negative acoustic word samples, respectively. For the MRI videos, $\mathcal{L}(g(v), g(v^+), g(v^-))$ can be defined as the objective function used in the training stage, where g represents the designed network for embedding creation.

The general model for the base networks represented by f and g can be seen in Fig. 3.9. This network is chosen, after several experiments. In the experiments, it is observed that increasing LSTM units from one to two boosts performance considerably, however additional units do not result in significant performance gain, while they add computational burden. The dimension of the word embeddings is specified with the node number of the last fully-connected layer, which is chosen as 256. Batch normalization (after input layer and fully-connected layers), and dropout (after LSTM units) with a rate 0.2 are utilized in the network to increase its generalization power on the unseen data.

The presented models for the unimodal approach enable embedding generation for the speech and MRI videos independently, and the models are trained to discriminate the word examples from the same modality. However, embeddings that provide a distance measure between the subunits of different data modalities are required for cross-modal applications. For this reason, recurrent neural networks with multi-view

losses [66] are designed to jointly create audio and MRI video word embeddings for the cross-view same-different word discrimination task. The presented network is trained in such a way that the cosine distance between the cross-modal embeddings with the same word label is smaller than the cosine distance between the embeddings having different word labels by a margin m. To investigate the information that different data modalities carry, several multi-view losses are examined in the training stage. Considered objective functions can be defined as:

$$\mathcal{L}_0 = \mathcal{L}(g(v), f(a^+), f(a^-)), \tag{3.2}$$

$$\mathcal{L}_1 = \mathcal{L}(f(a), g(v^+), f(a^-)), \tag{3.3}$$

$$\mathcal{L}_2 = \mathcal{L}(f(a), g(v^+), g(v^-)), \tag{3.4}$$

$$\mathcal{L}_3 = \mathcal{L}(g(v), f(a^+), g(v^-)), \tag{3.5}$$

where f and g are the presented networks for the embedding learning for audio and video modalities, respectively. They have the same topology with the network presented in Fig. 3.9.

In the presented multi-view losses, input triplets can be sampled in a variety of ways. In this study, three different schemes are considered: Random, online semi-hard, and hard mining sampling methods. In online semi-hard mining scheme [67], the network is trained utilizing every available positive samples in a mini batch, and the corresponding negative examples are the most difficult samples satisfying:

$$d(g(v_n), f(a_n^+)) < d(g(v_n), f(a_n^-))). \tag{3.6}$$

The inequality 3.6 is defined for $\mathcal{L}_0$, however the definition can be extended for the other objective functions. In the hard mining sampling, every possible positive pairs in a mini batch is utilized again for the training, but the most difficult negative example for each positive pair is sampled in this method without any constraint. We observed that the random sampling works the best among all examined methods for our data set, and all input triplets in the considered objective functions are created with this method.

In this thesis, different network models are presented so far to create embeddings for unimodal and cross-view settings. However, a unified model for the creation of embeddings that can be utilized in both unimodal and cross-modal applications without a severe performance degradation can lead to more efficient storage systems. While [18] shows that the embeddings generated with the multi-view losses perform also better in unimodal tasks, text which is a completely noiseless information source is utilized in that work as a secondary information source. Since MRI videos are used with the speech signal in this thesis, both information sources are noisy in our system. Therefore, it cannot be expected that the findings of [18] are directly applicable to our study.

Embeddings generated to be used in both unimodal and cross-modal applications are called multitask embeddings, and they are created with the use of a single recurrent neural network. This network is trained with the mixture of unimodal and multi-view losses, and several mixture options are examined. The considered objective functions are:

$$\mathcal{L}_{m0} = (\mathcal{L}(f(a), f(a^+), f(a^-)) + \mathcal{L}(g(v), g(v^+), g(v^-)) + \mathcal{L}_0), \tag{3.7}$$

$$\mathcal{L}_{m1} = (\mathcal{L}(f(v), g(a^+), g(a^-)) + \mathcal{L}(g(v), g(v^+), g(v^-))), \tag{3.8}$$

$$\mathcal{L}_{m2} = (\mathcal{L}(f(a), g(a^+), g(a^-)) + \mathcal{L}(f(a), f(a^+), f(a^-))). \tag{3.9}$$

Another method is also investigated to create multitask embeddings, in which a neural network is trained with $\mathcal{L}_0$ objective function by initializing the f and g base networks with the weights of the best models obtained for unimodal tasks. In this approach, pretrained models for unimodal tasks are tuned with a multi-view loss with the aim of multitask embedding generation. The presented network for this method is named as MultiNet.

3.3.3. Joint Representation and Distance Metric Learning

In Section 3.3.2, several methods are presented to generate embeddings that can be used with cosine distance. In literature, the advantages of learning a distance metric jointly with the data representations are shown for keyword spotting application [24]. This approach has also inspired us to design a system in which the data representations and a distance metric are learnt together.

In this section, all methods for the joint representation and distance metric learning are presented on TRT Signed Turkish broadcast news. While the presented approaches for the representation generation on USC-TIMIT rtMRI data set are applicable to cross-modal applications, embeddings are created only for the video frames corresponding to signs in the proposed method on Signed Turkish broadcast news. However, the second information source in the videos, the speech signal, is also utilized in this study to extract sign units found in the videos.

Most of the supervised studies on sign language videos are dependent on the existence of labeled data sets. However, the manual annotation of the sign language videos is an expensive process. In our study on the Signed Turkish broadcast news, we use the information coming from speech signal to provide temporal boundaries for the signs found in the videos. For this purpose, we utilize an ASR tool, Google Cloud Speech-to-Text [68], to acquire the transcription of the speech signal. This tool can provide transcriptions for speech signals in 120 different languages (including Turkish) with the use of deep neural networks.

After the transcriptions for the Signed Turkish broadcast news are acquired, temporal boundaries for all word examples are extracted. In the utilized data set, the signs are the corresponding units of the words in TSL, however the signs and the words have the same grammatical order differing from TSL. That is why the transcriptions of the speech roughly give the temporal information for the signs. However, the time boundaries of the words are not directly used for the corresponding signs in this work.

To acquire a better alignment for the sign examples, the average duration of each word is found. In the next step, the word example that has the closest duration to the average duration is used with adding an 80 ms window to its end to align other examples having the same word label for each word sample. For each word example, the time period beginning 240 ms before the word's start and ending 480 ms after its end is considered to find an approximate alignment. Diagonal matching algorithm is used on the sequence found in the considered time window for the reference example and the other examples in the same word class to extract sign samples. In the diagonal matching algorithm, if the reference example $X = (x_1, \ldots, x_N)$ is a sequence of length N consisting of 100 dimensional OpenPose features, and the sequence $Y = (y_1, \ldots, y_M)$ of length M is found in the considered time period for another sample; a subsequence $Y_{sub} = (y_{\hat{i}}, \ldots, y_{\hat{i}+N})$ of length N is extracted from the sequence Y. The variable $\hat{i}$ in Y_{sub} can be found through the equations 3.10 and 3.11.

$$\mathcal{D}(i) = \sum_{j=1}^{N} d(x_j, y_{i+j}), \tag{3.10}$$

$$\hat{i} = \underset{i}{\operatorname{argmin}}\, \mathcal{D}(i). \tag{3.11}$$

In Equation 3.10, $d(.,.)$ is used to represent Euclidean distance between two vectors. The extracted video frame sequences with the diagonal matching algorithm and the reference video frame sequences are used as the sign examples for the corresponding word labels.

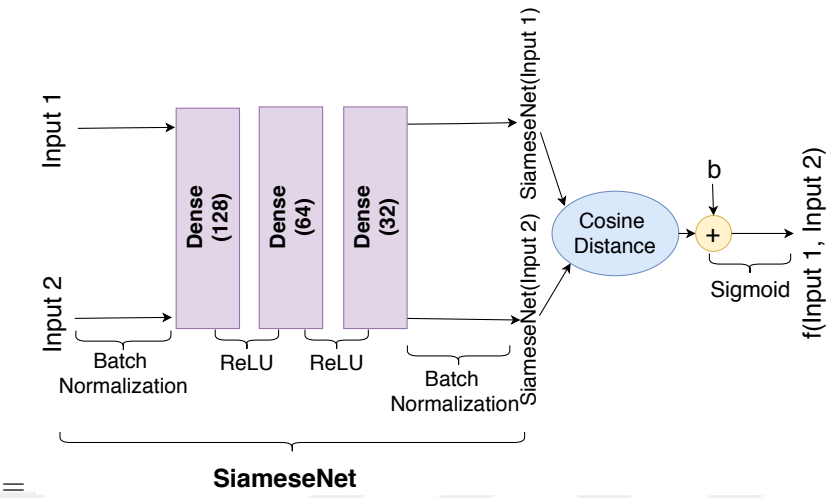


Figure 3.10: The presented network for joint representation and distance metric learning.

After the sign samples are extracted, those examples are utilized to train a network that learns representations for the video frames and a distance metric to be used in DTW algorithm with the generated representations. The proposed method for this purpose is adapted from a study presented for keyword spotting [24]. In the designed network, inputs are the frames of the video segments having the same length, and those frames are represented with 100 dimensional OpenPose features. The network is trained such that if the input frames are coming from the video segments of the same signs, the desired output is 1, otherwise the desired output is 0. The overall network topology can be seen in Fig. 3.10.

In the proposed method, 32 dimensional representations are generated for each 100 dimensional input feature vectors through a deep neural network composed of three fully-connected layers. Therefore, video segments can be represented with the created embeddings for each frame. The distance metric for the created representations is defined as it is explained in [24]. If OpenPose feature vectors for two video frames are represented with x and y, then the distance metric for the generated representations can be defined as;

$$\mathcal{U}(x, y) = 1 - f(x, y). \tag{3.12}$$

In Equation 3.12, $\mathcal{U}(.,.)$ represents the distance metric which is learnt through the presented neural network. In the same equation, $f(x,y)$ is the output of the network shown in Fig. 3.10 for the inputs x and y. According to the defined distance metric, if the video frames are sourced from the video segments of the same signs, generated representations for them should be close to each other. If the frames belong to the segments corresponding to different signs, then the created representations for them will be distant to each other according to the defined distance metric in Equation 3.12. Therefore, the proposed method provides representations for the video frames and a distance metric that can be used with these representations.

The performance evaluation of the proposed method is done with the use of same-different sign discrimination task. In this task, two video segments are taken as inputs and a decision whether or not those two segments represent the same sign is made. For this task, to evaluate the distance between two sequences that consist of generated representations DTW algorithm is employed. DTW algorithm is utilized frequently in ASR applications [69] to evaluate the similarity between data sequences. In this algorithm, if the warping path between the sequence $X = (x_1, \ldots, x_N)$ of length N and another sequence $Y = (y_1, \ldots, y_M)$ of length M is $\phi = ((x_1, y_1), \ldots, (x_{i_k}, y_{j_k}), \ldots, (x_N, y_M))$ of length T, then the distance between these sequences can be defined as;

$$\mathcal{D}_{DTW}(X,Y) = \frac{1}{T}\sum_{k=1}^{T} d(x_{i_k}, y_{j_k}). \tag{3.13}$$

In equation 3.13, $d(.,.)$ is a distance metric defined for vectors. Generally, this distance metric is chosen as cosine distance or Euclidean distance, however $\mathcal{U}(.,.)$ function which is defined in Equation 3.12 is utilized as the distance metric in this study. Therefore, a distance measure can be evaluated between the data sequences composed of generated representations by using the learned distance metric in the DTW algorithm.

4. EXPERIMENTS

It is essential to clarify all the utilized model parameters in the experiments for the reproducibility of the results. Therefore, after explaining the features of utilized data sets, all experimental setup is explained in this chapter.

4.1. Data Sets

For all the experiments performed on the phone classification task, the USC-TIMIT data set [18] is used. While the USC-TIMIT data set is used in the problems of hidden representation and embedding generation, TRT Signed Turkish broadcast news are utilized in the experiments of joint representation and distance metric learning problem.

USC-TIMIT data set is composed of rtMRI videos of the upper airway in the mid-sagittal plane and corresponding speech waveforms of five female and five male subjects. The sentences in the database are the 460 sentences that are also used in the MOCHA-TIMIT [20] corpus. The videos in the data set are recorded with a frame rate of 23.18 frames/sec and a spatial resolution of 68×68 pixels for each frame. The frequency of the audio files is 20KHz and each file has an average duration of 25 seconds. In our experiments, the data corresponding to all ten subjects are utilized. In the performed experiments, phonetic and word level alignments provided with the USC-TIMIT dataset are utilized.

There are a total of 90 videos in the utilized TRT Signed Turkish broadcast news. All videos have a duration of approximately 12 hours. The frame rate of the videos is 25 frames/sec, and the videos are recorded with a spatial resolution of 352×288 pixels for each frame. The signs in the videos are the corresponding TSL translations of the words found in the speech. The signs and the words found in the broadcast news have the same grammatical order in contradistinction to TSL. All broadcast news are presented by the same person.

4.2. Experimental Setup

4.2.1. Hidden Representations for Phones

In total, there are 920 video and 920 audio files corresponding to the ten people in the USC-TIMIT data set. Among them, 736 files from each data stream are utilized in the training set, whilst the remaining 184 videos and audio files are divided equally among the validation set and the test set. Five-fold cross validation method is applied to get the final results for each experiment.For the phone classification task, there are totals of 99670 and 15137 phone samples in the training and validation sets, respectively. Mini-batch training method is utilized, and the batch size is chosen as 32 in all experiments for this first task. The networks are trained for 50 epochs. Adam optimizer [70] with a learning rate of 0.001 is utilized, due to the empirical experience of observing that bigger learning rates result in a failure in the training process.

4.2.2. Hidden Representations for Words

To tackle the cross-modal same-different word discrimination task, 23725 word instances that come from video and audio files are used for the training. The validation set consists of 3218 word examples from both data modalities. In order to evaluate the performance of the same-different word discrimination methods, a word coming from one data source is compared with a word example from the other data source, corresponding to the same or a different word label. If the word labels on a trial are the same, a class label of 1 is assigned, whereas the class label is specified as 0 otherwise. This process is applied for each word example acquired from audio and video. In the performed experiments, positive and negative examples (words having the same label and different label respectively) are chosen randomly. For the positive examples, pools are generated from each instance (from all people) of each word label. In each epoch, a random selection is done among all candidate words having the same word label.

In order to provide a fair comparison between the systems presented in Figs. 3.6 and 3.7, the randomness in the test sets should be prevented. To satisfy this need,

the exact same generated test sets are used for the performance evaluations of both systems. In the word discrimination task, a batch size of 16 is used to decrease the total memory consumption and training process is performed for 10 epochs. Adam optimizer with the learning rate 0.001 is used in the training.

4.2.3. Word Embedding Generation

In the experiments of embedding generation problem, again USC-TIMIT data set is utilized. Therefore, the sizes of training, validation and test sets are equal to the ones used in the experiments of hidden representation learning problem. For the training of the proposed models, mini-batch training algorithm with batch size 128 is used in all experiments. Each model is trained for 50 epochs with a learning rate 0.001, and bigger learning rates result in a failure in the training process. The margin value presented in all the defined objective functions in Section 3.3.2 is specified as 0.3.

Experiments for embedding generation are defined on unimodal and cross-view same-different word discrimination tasks. In these tasks, data sequences corresponding to words are discriminated on whether they have the same word label or not. In the unimodal setting, input sequences have the same modality (audio-audio or video-video), while sequences come from different data sources (audio-video) in the cross-view task. Test set for the experiments are created randomly sampling positive and negative examples for each word instance. Thus, the number of positive and negative samples are equal in the test set, which means that an average precision (AP) of 50% can be thought as the naive baseline for the defined tasks.

4.2.4. Joint Representation and Distance Metric Learning

In the experiments of joint representation and distance metric learning, 70 videos are utilized in the training stage. Remaining 20 videos are divided equally to be used in the validation and test sets. There are 465294 and 49031 video frames corresponding to 5781 and 1071 words in the training and validation sets, respectively. Mini-batch algorithm with a batch size 256 is utilized in the training of the network. The proposed

neural network is trained for 3 epochs with a learning rate 0.001.

In the experiments, performance measurement is made by using the same-different sign discrimination task. In the defined task, video segments are taken as inputs and a decision whether the signs in the videos are the same or different is outputted. Created test set for this task consists of 1168 words and randomly sampled positive and negative examples for each anchor word. Therefore, there are equal number of positive and negative samples in the test set, and it can be said that %50 AP can be used as a naive baseline for the defined task.

5. RESULTS

5.0.1. Hidden Representations for Phones

To the best of our knowledge, our work is the first study working with the USC-TIMIT dataset on the phone classification task. As a close work in the literature, [34] can be given due to the work's endeavors on the broad phone classification task. Even though the defined classes in the phone classification task presented in this thesis do not include silence as a class label, results for the *broad* classification task are also acquired for the networks proposed in this work, in order to be able to compare our results with another work in the literature. In the broad phone classification task considered by [34], there are a total of five classes consisting of vowels, stops, fricatives, nasals, and silence, which is also considered as the broad phone classes in this paper. In [34], outcomes are obtained using the information extracted from both rtMRI videos and audio, therefore the study is also a multimodal one. Several results are provided for different features generated to represent data, and we compare our results with the highest average accuracy obtained by [34]. Comparative results on the broad phone classification task are presented in Table 5.1.

The results of Table 5.1 show that the proposed multimodal systems significantly outperform the best model presented in [34]. In addition, the proposed unimodal deep neural network utilizing the audio is also surprisingly found to yield a better accuracy than the multimodal method proposed in [34], while the video-based system has a comparable accuracy with it. One thing to note is that among the unimodal approaches presented in this paper, the network that uses audio-only data yields a better accuracy than the video-only network for this task. This phenomenon implies the significance of the audio data over video for this task. As it is expected, proposed multimodal systems perform better than the unimodal networks for broad phone classification task. The results also show the performance gain that can be achieved through the use of sequential cross-connections in the multimodal systems.

Table 5.1: Results for the broad phone classification task.

	Accuracy
Best score of [34]	61.50%
Unimodal Approach, Audio (Fig. 3.3)	69.74%
Unimodal Approach, Video (Fig. 3.3)	61.60%
Data Fusion at Output Stage (Fig. 3.4)	70.60%
Cross-modal Deep Neural Network (Fig. 3.5)	73.42%

The phone classification accuracies of the four methods presented in this thesis for the phone classification task are presented in Table 5.2. In the speech processing literature, presented results for the phoneme recognition accuracy task on TIMIT speech corpus are noticeably better than the accuracy provided by the audio-only approach in this study. For example, in [71], a study that utilizes deep recurrent neural networks, a phone recognition accuracy of 82.3% is reported. However, the audio files in USC-TIMIT dataset are noisy as a result of simultaneous recording of the audio and MRI videos. This situation degrades performance as observed in [33], where a phone error rate of 63.0% is attained on the audio-based speech recognition task.

Overall, the results show that both rtMRI videos and audio files are valuable information sources for the speech and audiovisual applications. The main outcome of the experiments that yield Table 5.2 is the performance gain that multimodality can bring, noting the models using information that come from both data streams overperform the unimodal approaches. The lowest accuracy in phone classification is obtained when the input is solely video data, which may sourced to the noisiness in the rtMRI video frames.

The results of Table 5.2 also signify the importance of the stage where the data fusion takes place. As hypothesized, the results confirm that performing the data fusion at earlier stages results in a better utilization of different data streams. Therefore, while extracting smaller dimensional data representations through a deep neural network, providing the raw input data information from other modality to that network may

Table 5.2: Phone classification accuracy results for the several models presented for this task.

	Accuracy
Unimodal Approach, Audio (Fig. 3.2)	41.86%
Unimodal Approach, Video (Fig. 3.3)	35.68%
Data Fusion at Output Stage (Fig. 3.4)	43.95%
Cross-modal Deep Neural Network (Fig. 3.5)	47.86%

bring better representations. Heuristically, this achievement may be due to giving more freedom to the network in the way that it extracts information coming from different data streams.

To analyze the classification mechanism of the network presented in Fig. 3.3, saliency maps for different phonemes are generated. Saliency maps signify the regions in the input image that have the greatest effect in the predicted class [72], and they are used to visualize the results of convolutional neural networks on a similar data set consisting of MRI videos [73]. They are created by slightly changing the input image, and computing its effect in the output. Generated saliency maps for the video frames corresponding to the midtime of the phones v, z, sh, and k can be seen in Fig. 5.1.

In the figures, brighter parts show the image regions affecting the output greatly. For the phone v, a labiodental phone according to the place of articulation, it can be seen that regions around the lips and teeth have more importance for the classification decision. For the phone z, an alveolar phoneme for which the place of articulation near the alveolar ridge, saliency map signifies the importance of the region around alveolar for the classification output. In the saliency map of a palato-alveolar phoneme sh, the region found in the backwards of the mouth is found to be important for the output. Lastly for a velar phone k, for which the place of articulation is the back point of palate, saliency map also points out the place of articulation as the most influential region on the classification decision. This analysis can imply that the deep neural networks' classification mechanism is related to the place of articulations of the

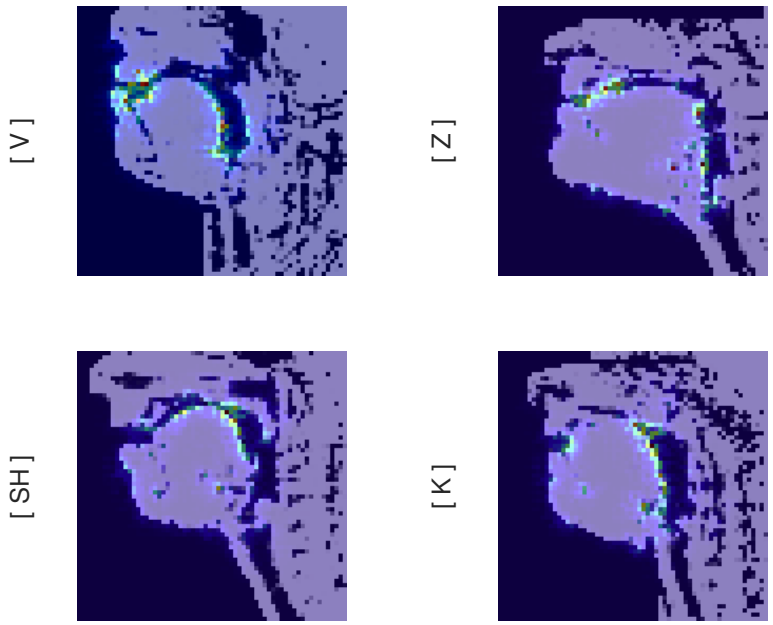


Figure 5.1: Saliency maps for different phones generated for the network described by Fig. 3.3. The upper left image corresponds to the phone 'V', the upper right one is for the phone 'Z', the lower left one is for the phone 'SH', and the last image corresponds to the phone 'K'.

phonemes. Since rtMRI videos provide this essential information, they can improve the speech recognition applications.

To analyze the reasons behind phone misclassifications when the rtMRI videos are the only input source to the network, a histogram of the phone pairs that is mixed up with each other most frequently is generated. The histogram can be found in Fig. 5.2. It can be deduced that the majority of the errors result from the phones for which the vocal tract has similar shapes in their generation. To demonstrate this similarity, the saliency maps for video frames corresponding to the midtime of the mostly mixed up phones 'Z' and 'S' are presented in the Fig. 5.3. Fig. 5.3 shows that the area around vocal folds are more active when the phone z is uttered, since it is a voiced consonant while s is a voiceless one. Despite this difference, the alikeness of the saliency maps and the original image frames are obvious, and some of the misclassifications can be

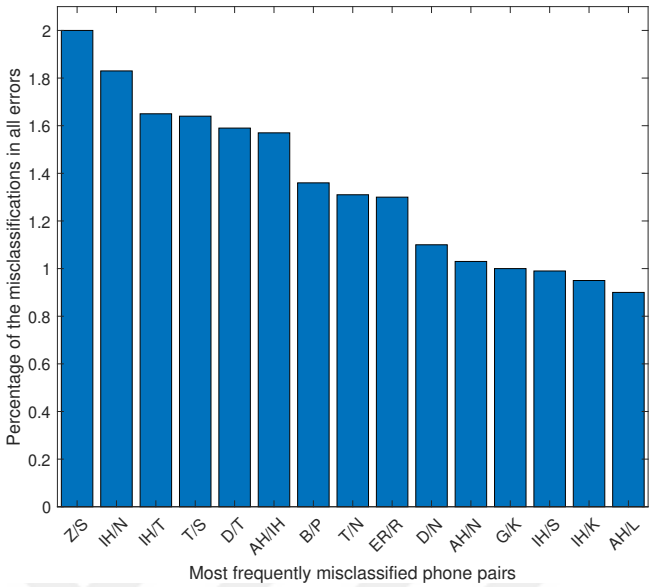
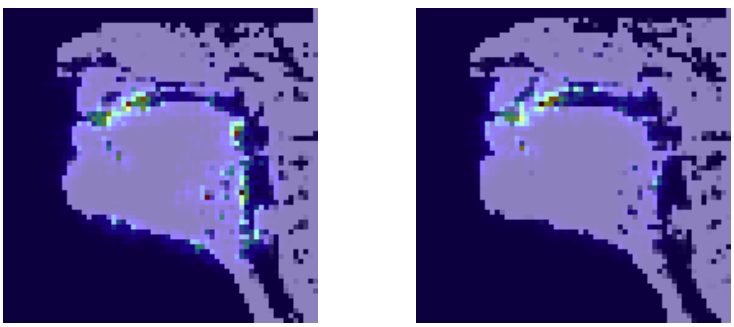


Figure 5.2: Histogram of misclassified phone pairs using the system described by Fig. 3.3.

justified with this similarity.



[Z] [S]

Figure 5.3: Saliency maps for video frames corresponding to the most frequently misclassified phone pair using the system described by Fig. 3.3. The left image represents the saliency map for the video frame corresponding to the phone 'Z', whilst the right image represents the saliency map for the video frame corresponding to the phone 'S'.

5.0.2. Hidden Representations for Words

The results of the cross-view same-different word discrimination task for the two networks proposed in Figs. 3.6 and 3.7 are presented in Table 5.3. The naive baseline of this task can be thought as a coin-flip random choice among the *same* and *different*

options, which would result in an accuracy of 0.5, given the availability of enough test samples.

The accuracy results of the word discrimination task in Table 5.3 show that embeddings extracted from rtMRI videos and audio carry the mutual information about their sources, since the accuracies that are obtained with both deep neural networks are promising. Noting that the use of sequential cross-connections in the deep neural network provides an accuracy advantage of roughly 4% over the network presented in Fig. 3.6, it can also be concluded that if data representations can be acquired with the knowledge of the other data stream, the discriminability of the representations increases. Overall, similar to the phone classification task, performing the fusion at an earlier stage with the help of cross connections improves the performance for the same-different word discrimination task, as well.

Table 5.3: Overall accuracies and error distributions on false positives and false negatives when the word pairs are uttered by the same subject or by different subjects for different proposed networks.

Network in Fig. 3.6	Accuracy	False Positive (% of Errors)	False Negative (% of Errors)
Total	71.16%	74.91%	25.09%
Same Speaker	75.73%	88.22%	11.78%
Different Speakers	70.57%	73.52%	26.48%
Network in Fig. 3.7	Accuracy	False Positive (% of Errors)	False Negative (% of Errors)
Total	75.13%	68.12%	31.88%
Same Speaker	81.02%	87.21%	12.79%
Different Speakers	74.46%	66.29%	33.71%

Table 5.3 also presents accuracy, and the distributions of false positives and false negatives over all errors when the subjects are the same or different for both networks presented in Section 3.3.1. For the network proposed in Fig. 3.6, false positives are 74.91% of the total errors and for the network presented in Fig. 3.7, this ratio is 68.12%.

Therefore, most of the errors are composed of false positives. For both networks, false positives have higher ratios in same speaker errors compared to false positive ratios in overall errors. This outcome indicates that when the word pairs are uttered by the same subject, the probability of a false positive increases. Additionally, false negatives have a higher fraction in the errors of different speakers compared to the ratio of false negatives in all errors for both networks. Therefore, the probability of false negatives increases when the word pairs are uttered by different speakers. Accuracy outcomes demonstrate that word pairs produced by different speakers are more prone to be misclassified for both network schemes. To conclude, speaker variability is one of the sources of errors made for the word discrimination task, and the results can be improved with speaker adaptation methods.

In order to investigate the sources of errors in the cross-view same-different word discrimination task, most frequently misclassified words and word pairs are obtained and presented in Table 5.4. In Table 5.4, false negatives and false positives are demonstrated in the left column and right column respectively. The table is arranged in a descending order in frequency of error occurrences which is normalized with the rate of words' appearances in the test sets. From the results of Table 5.4, it can be deduced that the main sources of errors are the words with short lengths. This implication holds for both false positives and false negatives. This phenomenon may signify that the embeddings created for the words with shorter lengths carry less information about their sources, which makes harder to discriminate them.

5.0.3. Word Embedding Generation

Results of the experiments for the embedding generation problem are acquired by averaging the outcomes of 5-fold cross-validation. In the embedding generation problem, AP (the total area under precision-recall curve) is utilized as a performance measure. For this problem, results are obtained for the same-different word discrimination task defined in both unimodal and cross-view settings. The overall performances of the audio and MRI video embeddings for the unimodal same-different word discrimination task can be seen in Table 5.5. In the table, presented results for DTW are

Table 5.4: Words and word pairs that are frequently misclassified in the cross-view same-different word task.

Common False Negatives	**Common False Positives**
ahead	the - for
all	and - the
cliff's	my - the
huge	measured - be
each	an- the
by	masquerade - classrooms
in	iguanas - gas
beggar	can - good
my	we - the
money	to - is

collected by using the DTW alignment costs between the MFCC feature vectors or video features acquired with the correspondence autoencoder as the distance measure in the unimodal same-different word discrimination task. Additionally, the presented result for DTW_{raw} in Table 5.5 is calculated by employing the DTW alignment cost between the pixel values of video frame sequences as the distance measure in the word-discrimination task.

The presented results in Table 5.5 demonstrate that using the rtMRI videos while creating the embeddings of speech units does not bring any performance gain, in contrary it degrades the overall performance considerably in unimodal same-different word discrimination task. This outcome contradicts with the findings of [18], since using the second information source (text data) while creating audio embeddings boosts the performance in the unimodal word discrimination task in [18]. However, textual data is a completely noiseless information source, while rtMRI videos are inherently noisy because of their generation process. Therefore, different outcomes can be expected. The table also shows that utilizing the information coming from speech signal in the creation of MRI video embeddings results in a similar performance with the unimodal

Table 5.5: AP for the unimodal same-different word discrimination task for different objective functions and baseline systems.

	Video	**Audio**
Unimodal Objective	$77.96\% \pm 0.65$	$86.30\% \pm 3.34$
$\mathcal{L}_0$	$77.22\% \pm 1.13$	$76.84\% \pm 1.84$
$\mathcal{L}_1$	$77.04\% \pm 0.51$	$78.57\% \pm 1.12$
$\mathcal{L}_2$	$77.50\% \pm 1.16$	$77.37\% \pm 1.12$
$\mathcal{L}_3$	$77.37\% \pm 1.14$	$77.76\% \pm 0.95$
$\mathcal{L}_0 + \mathcal{L}_2$	$77.25\% \pm 1.63$	$78.00\% \pm 1.52$
DTW	$68.34\% \pm 0.50$	$72.92\% \pm 0.69$
DTW_{raw}	$68.13\% \pm 0.82$	

loss in the unimodal same-different word discrimination task.

Another important implication of the results presented in Table 5.5 is that the performance of the features acquired with the correspondence autoencoder is slightly better than the raw pixel values of the video frames. Additionally, the features coming from the autoencoder are denser compared to pixel values of the images, and they require less memory and time in the training process. Therefore, it can be said that the features acquired with the autoencoder represent the MRI videos as good as the pixel values with denser vectors. Overall, the table illustrates that the proposed methods in this thesis for the embedding generation clearly overperform the DTW-based opponents in the unimodal word discrimination task.

The calculated AP values for the proposed methods in this thesis for the cross-view same-different word discrimination task can be seen in Table 5.6. In the table, presented AP values for DTW is acquired by zero padding 13 dimensional MFCC features into 256 dimensional vectors and using the DTW alignment cost between the padded vectors and video features outputted from the autoencoder as the distance measure in the word discrimination task. However, the performance acquired with this method is approximately the outcome one would get with random decision, since

the data modalities are different. Therefore, the presented AP value for DTW can be thought as a naive baseline for the proposed methods in this study for the same task.

In Table 5.6, performance comparisons for several multi-view losses are presented. The outcomes demonstrate that the best performance can be achieved with the use of $\mathcal{L}_0$ objective function in the training stage. While $\mathcal{L}_0 + \mathcal{L}_2$ function results in a closer performance, trained networks with objective functions $\mathcal{L}_1$ and $\mathcal{L}_3$ perform considerably worse than the model trained with $\mathcal{L}_0$ objective. Additionally, the performance of the mixture of the objective functions $\mathcal{L}_0$ and $\mathcal{L}_2$ is also considered, since its use brings performance gain in [18] and it is a mixture of the two best-performing losses. While the use of mixture function leads to results closer to the best ones for all the considered tasks, it does not provide the best performance for any of the tasks.

Table 5.6: AP for cross-view same-different word discrimination task for different objective functions and DTW system.

	Video-Audio
$\mathcal{L}_0$	$68.16\% \pm 2.49$
$\mathcal{L}_1$	$64.12\% \pm 1.19$
$\mathcal{L}_2$	$65.04\% \pm 2.06$
$\mathcal{L}_3$	$61.66\% \pm 3.59$
$\mathcal{L}_0 + \mathcal{L}_2$	$67.40\% \pm 1.18$
DTW	$50.45\% \pm 0.36$

In Fig. 5.4, precision-recall curves acquired with the objective function $\mathcal{L}_0$, the best model, for overall word pairs in the test set, and for the ones uttered by the same or different speakers can be found. An analysis on the effects of speaker variability for the word discrimination task can be done by examining these curves. The figure shows that the AP value is considerably higher when the words are uttered by the same person. Therefore, it can be said that speaker variability is one of the sources of misclassifications, and the overall performances can be improved by utilizing speaker

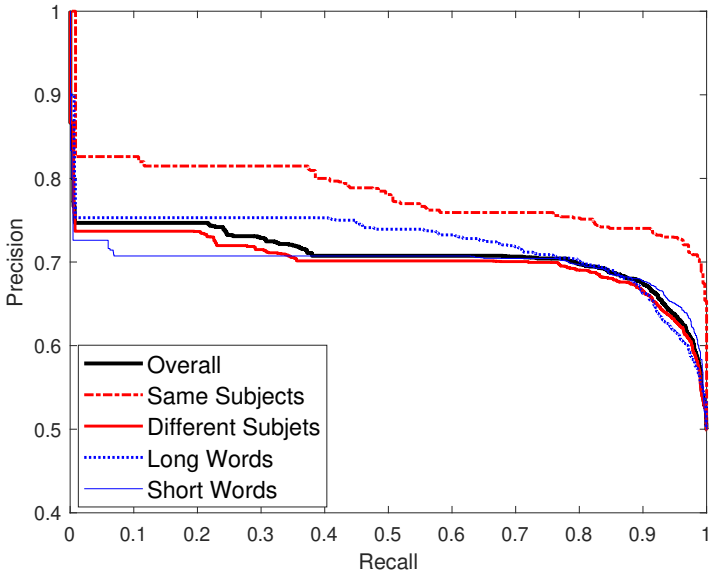


Figure 5.4: Precision-recall curves for different speaker and word length cases when the objective function is $\mathcal{L}_0$ for cross-view same-different word discrimination task.

adaptation methods.

To observe the effect of word lengths on the performance, precision-recall curves of the word pairs whose anchor word is longer or shorter than the average word length are demonstrated in Fig. 5.4. The curves illustrate that there is a higher chance for a short word to be misclassified. Additionally, the analysis for the effects of the frequency of the words on the performance demonstrates that less frequent words have a slightly better AP compared to frequent words. This may seem as a contradiction to the expectation that the performance of the frequent words should be better, since the network can be trained better for them. However, it is observed that most of the frequent words are significantly shorter compared to the less frequent words. Therefore, the effect of the word length might dominate the effect of the frequency of the words.

The distribution of the total errors, false positives, and false negatives for the word pairs that are uttered by the same person can be found in Table 5.7 for all the best performing methods in the embedding generation. If it is assumed that speaker identity does not bring any bias for the outcomes, there is an approximately 10% chance (there are 10 speakers) that the erroneously classified word pairs are uttered by the

same subject. However, approximately 11% of the false positives take place for the word pairs uttered by the same person. Since the percentage 11% is higher than the expected 10%, it can be said that false positives are more probable for the word pairs of the same speakers. Additionally, Table 5.7 shows that all false negatives occur when the word pairs are uttered by different speakers, which implies that word pairs with the same label are more prone to be misclassified if they are uttered by different speakers.

Table 5.7: The percentage of the errors when the speakers are the same for the best performing unimodal and multimodal systems.

	Video	**Audio**	**Video-Audio**
Total Errors	5.53%	5.52%	5.55%
False Positives	11.05%	11.08%	11.10%
False Negatives	0%	0%	0%

Table 5.8: AP values of all the proposed schemes for the creation of multitask embeddings.

	Audio	**Video**	**Cross-view**
$\mathcal{L}_{m0}$	$79.93\% \pm 2.63$	$77.51\% \pm 1.55$	$61.00\% \pm 3.36$
$\mathcal{L}_{m1}$	$75.03\% \pm 2.92$	$77.41\% \pm 1.51$	$64.48\% \pm 3.68$
$\mathcal{L}_{m2}$	$80.05\% \pm 1.66$	$77.60\% \pm 0.49$	$65.87\% \pm 1.25$
MultiNet	$77.56\% \pm 1.48$	$77.15\% \pm 0.76$	$67.41\% \pm 2.36$

The results of all the proposed methods for the generation of multitask embeddings can be seen in Table 5.8. In the table, performances of different methods are presented on both unimodal and cross-view same-different word discrimination tasks. The main purpose of the proposed models is to create embeddings that can be used in both unimodal and cross-modal settings without a severe performance degradation. Although [18] shows that embeddings generated with multi-view losses work also better in unimodal taks, the utilization of the rtMRI videos while generating audio embeddings causes performance loss in the unimodal settings in this study. Thus, the approaches in Table 5.8 are presented for the multitask embeddings.

The results presented in Table 5.8 demonstrate that the utilization of the objective $\mathcal{L}_{m2}$ and network Multinet can lead to similar performances to the best results presented in Tables 5.5, and 5.6 in the unimodal (defined for the video modality), and cross-view same-different word discrimination tasks. In the unimodal word discrimination task (defined for the audio), proposed systems trained with objectives $\mathcal{L}_{m0}$ and $\mathcal{L}_{m2}$ overperform the networks trained with multi-view objective functions. Since the utilization of $\mathcal{L}_{m2}$ results in better performances than the models trained with $\mathcal{L}_{m0}$ and $\mathcal{L}_{m1}$ objectives, it can be said that the use of $\mathcal{L}(g(v), g(v^+), g(v^-))$ function in the objective degrades the performance of the multitask embeddings. This performance loss can be justified with the noise found in the rtMRI videos, since noisy videos can provide misleading information to the system. Overall, the network trained with the objective function $\mathcal{L}_{m2}$ can generate embeddings that can be used in both unimodal and cross-modal tasks without any severe performance loss.

5.0.4. Joint Representation and Distance Metric Learning

For the joint representation and distance metric learning problem, same-different sign discrimination task is utilized to measure the performance of the proposed method. AP is used as the performance measure for the defined task. The results are acquired by calculating distances between the sign language video segments. The main purpose is getting lower distances for the video segment pairs in which there are the same signs compared to the distances between the video segments of different signs. Overall results for this problem can be found in Table 5.9, and the presented results are obtained by employing several distance metrics in the DTW algorithm.

In the table, the result presented in the first row is acquired by using cosine distance in the DTW algorithm to measure the distances between the video segments. The AP value shown in the second row is obtained by employing Euclidean distance again in the DTW algorithm. In the systems whose outcomes are presented in the first two rows, the time windows of the words are used to extract sign examples without a change. Therefore, diagonal matching algorithm is not used in the example extraction stages of those methods. In the results presented in third and forth rows, sign examples

Table 5.9: AP values for the same-different sign discrimination task.

Distance	**Diagonal Matching**	**AP**
Cosine Distance	✗	64.20%
Euclidean Distance	✗	66.12%
Cosine Distance	✓	70.42%
Euclidean Distance	✓	73.97%
Proposed Distance Metric	✓	80.88%

are generated by applying the diagonal matching algorithm on the time windows of the corresponding words. Those results are acquired again by employing the DTW algorithm to measure the distances between video segments with the distance metrics; cosine and Euclidean distances. The result in the last row is obtained by employing the distance metric which is learned through the method explained in 3.3.3 in the DTW algorithm. Since the proposed method requires input pairs which have the same length, the result for this method can only be presented for the sign examples acquired with the diagonal matching algorithm.

Table 5.9 clearly demonstrates that directly using the temporal information of the speech signal for the Signed Turkish videos is not an optimal approach for the extraction of the sign examples. The diagonal matching algorithm can provide a better alignment for the sign language videos. Additionally, Euclidean and cosine distances are the prevalent metrics that are used in the DTW algorithm to measure the similarity between sequences. Therefore, they can be thought as baseline systems. The table illustrates that the proposed joint representation and distance metric learning method for Signed Turkish videos considerably overperforms the corresponding baseline systems.

6. CONCLUSION AND FUTURE WORK

In this thesis, several methods for the data fusion and representation learning on multimodal data sets are investigated. Main purposes of the proposed methods are to exploit the advantages that multimodality can bring, and to meet the requirements that have been arisen due to the accumulation of data in different modalities. For these purposes, two different data sets are utilized in the experiments: USC-TIMIT rtMRI data set and Signed Turkish broadcast news.

In this thesis, the concept of data fusion is investigated on rtMRI videos and corresponding audio data, through the proposal of several deep neural network topologies. We developed various methods for the phonetic classification task, which perform data fusion among different data sources at different stages. The results of the phone classification task illustrate that the utilization of the data fusion does indeed increase classification accuracy on our experiments done on the USC-TIMIT dataset. Furthermore, our results show that the system is better in terms of classification accuracy when forming sequential cross-connections in an earlier stage (i.e. from the raw data), rather than fusing the representations that belong to different sources later on. Additionally, it has been shown that the proposed methods in this thesis work better than the systems presented before on the same data set for broad classification task.

On the same data set, in order to investigate the power of hidden representations for words in different data modalities (rtMRI videos and speech), we introduced a cross-modal same-different word discrimination task. We developed two different deep neural networks for this task, in one of which again employing sequential cross-connections. We found that the proposed schemes both perform considerably over a random same-different selection of accuracy $\sim 50\%$, while also demonstrating that a system with cross connections results in more discriminable representations, yielding a better performance for the proposed word discrimination task.

In the next step, embeddings for the words in the MRI videos and speech were generated such that the cosine distance between the embeddings reflects the similarity between original data sequences. We utilized recurrent neural networks to generate these representations for unimodal and cross-view same-different word discrimination tasks on the USC-TIMIT dataset. For the unimodal setting, proposed methods perform considerably better than the baseline systems that are based on the DTW algorithm. In the cross-view approach, several objective functions were examined, and promising results were acquired for further multimodal studies. Additionally, several system models were proposed to generate embeddings to be used in both unimodal and cross-view tasks.

As a final step to representation generation problem, a joint representation and distance metric learning method was proposed on the Signed Turkish broadcast news. In the broadcast news, the signs and the words have the same order, therefore transcriptions obtained from the speech signal can be used to align signs in the videos. In the alignment of sign examples, a diagonal matching algorithm was presented to finalize the time windows acquired from the transcriptions. With the extracted sign samples, a Siamese neural network was trained to jointly learn the representations for the video frames and a distance metric to be used with the learned representations in the DTW algorithm. The performance of the proposed method was evaluated with the same-different sign discrimination task, and the evaluated results show the performance superiority of the presented method over the use of common distance metrics (cosine and Euclidean distances) in the DTW algorithm.

The presented results for the hidden representation generation and embedding learning problems on the USC-TIMIT data set demonstrate the deteriorating effects of speaker variability on the same-different word discrimination task. Therefore, an important future work for this thesis can be the employment of speaker adaptation methods, while designing corresponding networks. Several image processing techniques can be utilized to normalize speaker variances in the rtMRI videos, and speech processing methods can be used to decrease the negative impacts of different speakers in the speech signal.

Query-by-example search systems are necessary to retrieve necessary information from accumulated data. Therefore, another important future work of this thesis is the design of networks for the query-by-example search based on the generated representations. While learned representations can be utilized in unimodal retrieval systems, they can also be useful in cross-view query-by-example search in which the query and searched data have different modalities. Additionally, query-by-sign search can be another application that representation learning can help. For example, with generated representation in the same space for speech and sign language units, the design of a cross-view query-by-sign search can be possible. Therefore, the proposed methods in this thesis can be used for unimodal and cross-view query-by-example search systems as a future work.

REFERENCES

1. Lingala, S. G., B. P. Sutton, M. E. Miquel and K. S. Nayak, "Recommendations for real-time speech MRI", *Journal of Magnetic Resonance Imaging*, Vol. 43, No. 1, pp. 28–44, 2016.

2. Akgul, Y. S., C. Kambhamettu and M. Stone, "Extraction and tracking of the tongue surface from ultrasound image sequences", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 298–303, IEEE, 1998.

3. Perkell, J. S., M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements", *The Journal of the Acoustical Society of America*, Vol. 92, No. 6, pp. 3078–3096, 1992.

4. Kiritani, S., "X-ray microbeam method for measurement of articulatory dynamics-techniques and results", *Speech Communication*, Vol. 5, No. 2, pp. 119–140, 1986.

5. Ross, A. and A. K. Jain, "Multimodal biometrics: An overview", *2004 12th European Signal Processing Conference*, pp. 1221–1224, Sep. 2004.

6. Chen, T., "Audiovisual speech processing", *IEEE Signal Processing Magazine*, Vol. 18, No. 1, pp. 9–21, Jan. 2001.

7. Lahat, D., T. Adali and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects", *Proceedings of the IEEE*, Vol. 103, No. 9, pp. 1449–1477, Sep. 2015.

8. Khaleghi, B., A. Khamis, F. O. Karray and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art", *Information fusion*, Vol. 14, No. 1, pp. 28–44, 2013.

9. Shivappa, S. T., M. M. Trivedi and B. D. Rao, "Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey", *Proceedings of the IEEE*, Vol. 98, No. 10, pp. 1692–1715, 2010.

10. Sorber, L., M. Van Barel and L. De Lathauwer, "Structured data fusion", *IEEE Journal of Selected Topics in Signal Proctimatesing*, Vol. 9, No. 4, pp. 586–600, 2015.

11. Hu, B., Z. Lu, H. Li and Q. Chen, "Convolutional neural network architectures for matching natural language sentences", *Proc. Neural Information Processing Systems*, pp. 2042–2050, 2014.

12. Hadsell, R., S. Chopra and Y. LeCun, "Dimensionality reduction by learning an invariant mapping", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2, pp. 1735–1742, 2006.

13. Dhillon, P., D. P. Foster and L. H. Ungar, "Multi-view learning of word embeddings via cca", *Proc. Neural Information Processing Systems*, pp. 199–207, 2011.

14. Zhu, P., L. Xie and Y. Chen, "Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings", *Proc. Interspeech*, pp. 2192–2196, 2015.

15. Levin, K., K. Henry, A. Jansen and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings", *Proc. Automatic Speech Recognition and Understanding*, pp. 410–415, 2013.

16. Kamper, H., W. Wang and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4950–4954, 2016.

17. Settle, S. and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches", *Proc. Spoken Language Technology (SLT)*, pp.

503–510, 2016.

18. Narayanan, S., A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)", *The Journal of the Acoustical Society of America*, Vol. 136, No. 3, pp. 1307–1311, 2014.

19. Holzenberger, N., S. Palaskar, P. Madhyastha, F. Metze and R. Arora, "Learning from Multiview Correlations in Open-domain Videos", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8628–8632, 2019.

20. Wrench, A., "The MOCHA-TIMIT articulatory database", 1999.

21. Velickovic, P., Duo Wang, N. D. Lane and P. Lio, "X-CNN: Cross-modal convolutional neural networks for sparse datasets", *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8, Dec. 2016.

22. Cangea, C., P. Veličković and P. Liò, "XFlow: Cross-modal Deep Neural Networks for Audiovisual Classification", *IEEE transactions on neural networks and learning systems*, Vol. 1, pp. 1–10, 2019.

23. Köse, Ö. D. and M. Saraçlar, "Recurrent Neural Audiovisual Word Embeddings for Synchronized Speech and Real-Time Mri", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6424–6428, IEEE, 2020.

24. Gündoğdu, B., B. Yusuf and M. Saraçlar, "Joint learning of distance metric and query model for posteriorgram-based keyword search", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 11, No. 8, pp. 1318–1328, 2017.

25. Li, H., J. Tao, M. Yang and B. Liu, "Estimate articulatory MRI series from acoustic signal using deep architecture", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4854–4858, April 2015.

26. Desai, U., C. Yarra and P. K. Ghosh, "Concatenative Articulatory Video Synthesis Using Real-Time MRI Data for Spoken Language Training", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4999–5003, April 2018.

27. Mannem, R. and P. K. Ghosh, "Air-tissue Boundary Segmentation in Real Time Magnetic Resonance Imaging Video Using a Convolutional Encoder-decoder Network", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5941–5945, May 2019.

28. Valliappan, C., A. Kumar, R. Mannem, G. Karthik and P. K. Ghosh, "An Improved Air Tissue Boundary Segmentation Technique for Real Time Magnetic Resonance Imaging Video Using Segnet", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5921–5925, May 2019.

29. Asadiabadi, S. and E. Erzin, "Vocal Tract Airway Tissue Boundary Tracking for rtMRI Using Shape and Appearance Priors.", *Interspeech*, pp. 636–640, 2017.

30. Asadiabadi, S. and E. Erzin, "A Deep Learning Approach for Data Driven Vocal Tract Area Function Estimation", *IEEE Spoken Language Technology Workshop (SLT)*, pp. 167–173, 2018.

31. Saha, P., P. Srungarapu and S. Fels, "Towards automatic speech identification from vocal tract shape dynamics in real-time MRI", *INTERSPEECH*, pp. 1249–1253, 2018.

32. Sorensen, T., Z. I. Skordilis, A. Toutios, Y.-C. Kim, Y. Zhu, J. Kim, A. C. Lammert, V. Ramanarayanan, L. Goldstein, D. Byrd *et al.*, "Database of Volumetric and Real-Time Vocal Tract MRI for Speech Science.", *INTERSPEECH*, pp. 645–649, 2017.

33. Douros, I. K., A. Katsamanis and P. Maragos, "Multi-View Audio-Articulatory Features for Phonetic Recognition on RTMRI-TIMIT Database", *IEEE Interna-*

tional Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5514–5518, April 2018.

34. Prasad, A. and P. K. Ghosh, "Information theoretic optimal vocal tract region selection from real time magnetic resonance images for broad phonetic class recognition", *Computer Speech & Language*, Vol. 39, pp. 108–128, 2016.

35. Levin, K., K. Henry, A. Jansen and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings", *Proc. Automatic Speech Recognition and Understanding*, pp. 410–415, 2013.

36. Chen, G., C. Parada and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5236–5240, 2015.

37. Bengio, S. and G. Heigold, "Word Embeddings for Speech Recognition", *Proceedings of the 15th Conference of the International Speech Communication Association, Interspeech*, pp. 1053–1057, 2014.

38. Chung, Y.-A., C.-C. Wu, C.-H. Shen, H.-Y. Lee and L. Lee, "Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks", *Proc. Interspeech*, pp. 765–769, 2016.

39. Harwath, D. and J. Glass, "Deep multimodal semantic embeddings for speech and images", *Proc. Automatic Speech Recognition and Understanding*, pp. 237–244, 2015.

40. Harwath, D. and J. R. Glass, "Learning word-like units from joint audio-visual analysis", *Association for Computational Linguistics (ACL)*, pp. 506–517, 2017.

41. Harwath, D., A. Torralba and J. Glass, "Unsupervised learning of spoken language with visual context", *Proc. Neural Information Processing Systems*, pp. 1858–1866, 2016.

42. Arandjelovic, R. and A. Zisserman, "Objects that sound", *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 435–451, 2018.

43. Korbar, B., D. Tran and L. Torresani, "Co-Training of Audio and Video Representations from Self-Supervised Temporal Synchronization", *ArXiv*, Vol. abs/1807.00230, 2018.

44. Korbar, B., D. Tran and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization", *Advances in Neural Information Processing Systems*, pp. 7763–7774, 2018.

45. Parekh, S., S. Essid, A. Ozerov, N. Q. Duong, P. Pérez and G. Richard, "Weakly Supervised Representation Learning for Unsynchronized Audio-Visual Events.", *Computer Vision and Pattern Recognition Workshops*, pp. 2518–2519, 2018.

46. Surís, D., A. Duarte, A. Salvador, J. Torres and X. Giró-i Nieto, "Cross-modal embeddings for video and audio retrieval", *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 711–716, 2018.

47. Bengio, Y., J. Louradour, R. Collobert and J. Weston, "Curriculum learning", *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

48. Camgöz, N. C., A. A. Kındıroğlu, S. Karabüklü, M. Kelepir, A. S. Özsoy and L. Akarun, "BosphorusSign: a Turkish sign language recognition corpus in health and finance domains", *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1383–1388, 2016.

49. Forster, J., C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. H. Piater and H. Ney, "RWTH-phoenix-weather: A Large Vocabulary Sign Language Recognition and Translation Corpus.", *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pp. 3785–3789, 2012.

50. Alaydin, J. G. and M. Saraçlar, "Query-by-sign system for Turkish sign language broadcasts", *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2018.

51. Pfister, T., J. Charles and A. Zisserman, "Large-scale Learning of Sign Language by Watching TV (Using Co-occurrences).", *British Machine Vision Conference (BMVC)*, 2013.

52. Farhadi, A. and D. Forsyth, "Aligning ASL for statistical translation using a discriminative word model", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2, pp. 1471–1476, IEEE, 2006.

53. Kelly, D., J. Reilly Delannoy, J. Mc Donald and C. Markham, "A framework for continuous multimodal sign language recognition", *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pp. 351–358, 2009.

54. Duarte, A. C., "Cross-modal Neural Sign Language Translation", *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1650–1654, 2019.

55. Buehler, P., A. Zisserman and M. Everingham, "Learning sign language by watching TV (using weakly aligned subtitles)", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2961–2968, IEEE, 2009.

56. Kelly, D., J. Mc Donald and C. Markham, "Weakly supervised training of a sign language recognition system using multiple instance learning density matrices", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 41, No. 2, pp. 526–541, 2010.

57. Aran, O., I. Ari, L. Akarun, E. Dikici, S. Parlak, M. Saraclar, P. Campr and M. Hruz, "Speech and sliding text aided sign retrieval from hearing impaired sign news videos", *Journal on Multimodal User Interfaces*, Vol. 2, No. 2, p. 117, 2008.

58. Bilge, Y. C., N. Ikizler-Cinbis and R. G. Cinbis, "Zero-Shot Sign Language Recog-

nition: Can Textual Data Uncover Sign Languages?", *British Machine Vision Conference (BMVC)*, 2019.

59. Santemiz, P., O. Aran, M. Saraclar and L. Akarun, "Alignment based extraction of isolated signs from sign language videos", *2009 17th Signal Processing and Communications Applications Conference (SIU)*, pp. 532–535, IEEE, 2009.

60. Tamer, N. C., O. Özdemir, M. Saraçlar and L. Akarun, "Dynamic Time Warping Based Sign Retrieval", *IEEE 27th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, 2019.

61. Ertuğrul, B. S., C. Gurpinar, H. Kivrak and H. Kose, "Gesture recognition for humanoid assisted interactive sign language tutoring", *IEEE 21st Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, 2013.

62. Aran, O. and L. Akarun, "Sign Language Processing and Interactive Tools for Sign Language Education", *IEEE 15th Signal Processing and Communications Applications (SIU)*, pp. 1–4, 2007.

63. Kim, J., N. Kumar, S. Lee and S. Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data", *Proceedings of the International Seminar on Speech Production (ISSP)*, pp. 222–225, 2014.

64. Cao, Z., T. Simon, S.-E. Wei and Y. Sheikh, "Realtime multi-person 2-D pose estimation using part affinity fields", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291–7299, 2017.

65. He, K., X. Zhang, S. Ren and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

66. Hermann, K. M. and P. Blunsom, "Multilingual distributed representations without word alignment", *Proc. International Conference on Learning Representations*

(ICLR), pp. 1–9, 2014.

67. Schroff, F., D. Kalenichenko and J. Philbin, "Facenet: A unified embedding for face recognition and clustering", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

68. "Google Cloud Speech-to-Text", `https://cloud.google.com/speech-to-text`, accessed on 21.06.2020.

69. Gündoğdu, B., L. Sarı, G. Çetinkaya and M. Saraçlar, "Template-based keyword search with pseudo posteriorgrams", *IEEE 24th Signal Processing and Communication Application Conference (SIU)*, pp. 973–976, 2016.

70. Kingma, D. P. and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

71. Graves, A., A. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649, May 2013.

72. Kotikalapudi, R., "keras-vis", `https://github.com/raghakot/keras-vis`, 2017, accessed on 21.06.2020.

73. van Leeuwen, K., P. Bos, S. Trebeschi, M. van Alphen, L. Voskuilen, L. Smeele, F. van der Heijden and R. van Son, "CNN-based phoneme classifier from vocal tract MRI learns embedding consistent with articulatory topology", *Proc. Interspeech 2019*, pp. 909–913, 2019.