Image processing techniques for speech signal processing

Leow, Su Jun

2018

Leow, S. J. (2018). Image processing techniques for speech signal processing. Master's thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/73231>

<https://doi.org/10.32657/10356/73231>

IMAGE PROCESSING TECHNIQUES FOR SPEECH SIGNAL PROCESSING

LEOW SU JUN

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University in
partial fulfilment of the requirement for the degree of Master of
Engineering

2018

Acknowledgements

I would like to thank Professor Lee Chin-hui from Georgia Institute of Technology for his kind guidance and encouragement for this research work. I would also like to thank my supervisor, Dr Chng Eng Siong, for his advice and support for this work.

Contents

1	Introduction	15
1.1	Motivation	15
1.2	Contribution	18
1.3	Organization of Report	19
2	Converting Speech Problems to Image Problems	21
2.1	Parallel Between Speech and Image Problems	21
2.2	Image Spectrogram	24
2.2.1	Spectrogram Production	24
2.2.2	Speech and its Appearance on the Spectrogram	26
2.2.3	Variations a Spectrogram Undergoes	29
2.2.4	Features on the Spectrogram Image	32
3	Past Works relating to Image Processing Techniques for Speech Processing Tasks	37
3.1	Speech Processing using Spectrogram Reading	37
3.2	Speech Processing using the Image Spectrogram	40
3.2.1	Image Processing for Speech Signal Processing	40
3.2.2	Image Processing for Acoustic Signal Processing	42
3.3	Summary	43
4	Unsupervised Speech Segmentation	45
4.1	Introduction	45
4.2	Related Works	46
4.3	Signal Analysis	49
4.3.1	Spectral versus visual features for boundary detection	49
4.3.2	Sub-band Division to Enhance Boundary Detection	53
4.4	Segmentation Algorithm	59
4.4.1	Spectrogram Generation	59
4.4.2	Boundary Detection	60
4.4.3	Boundary Selection	61
4.5	Experimental Results and Discussions	62
4.5.1	Image-based Phoneme Segmentation	62

CONTENTS

4.5.2	Fusion with HMM-based Segmentation	64
4.6	Conclusion and Future Work	68
5	Detection of Unit Selection Based Synthesized Speech for Anti-spoofing	71
5.1	Introduction	71
5.2	Related Works	72
5.3	Observations	74
5.4	Algorithm	74
5.4.1	Spectrogram Generation	75
5.4.2	Lines Detection	75
5.4.3	Detections Verification	76
5.4.4	Finding Best Region for Verification	77
5.5	Experiments and Discussions	78
5.5.1	Corpus	78
5.5.2	Experimental Results	79
5.5.3	Discussions	82
5.6	Conclusion and Future Work	83
6	Conclusion and Way Ahead	85
	Publications	89
	References	90

List of Figures

1.1	An example of a spectrogram with corresponding phonetic transcription.	16
1.2	Fricative in waveform (top) and in spectrogram (bottom).	17
2.1	An illustration of common problems in speech and image research communities.	22
2.2	Manifestation of speech problems as image problems on the spectrogram.	24
2.3	Log-spectrogram, with window size = 128 samples (8 ms for sampling frequency of 16000Hz). Each pixel $20\log(X(k, n_s))$ represents the power at time n_s and frequency index k , with darker intensities corresponding to higher power.	25
2.4	(Top) Narrow-band and (bottom) wide-band spectrogram of same signal in Figure 2.3 with window of 15 milliseconds and 3 milliseconds respectively	26
2.5	Voice bar seen together with voiced stop ‘d’.	27
2.6	TIMIT sequence “She had your dark suit in greasy wash water all year”. The abbreviation is as follows: vow - vowels, nas - nasals, app - approximants, stop - stop, fri -fricatives.	29
2.7	Spectrograms of two different speakers, one male and one female, saying the same utterance “She had your dark suit”.	30
2.8	Spectrograms of the same utterance under (top) clean conditions, with (middle) white noise of amplitude 0.03 (maximum amplitude 1.0) added to the utterance and (bottom) DTMF tones of amplitude 0.2 added to the utterance.	31
2.9	Spectrograms of the same utterance with reverberation added to the middle and bottom utterances. (Middle) Reverberation delay of 0.02s is added. (Bottom) Reverberation delay of 0.1s is added. In both examples the decay is set to 0.5.	32
2.10	STFT coefficients of one utterance mapped to (top) 0-255 range to obtain grayscale images and (bottom) 0-16,777,215 range to obtain coloured image. Red indicates higher values.	33
2.11	Different textures found in spectrogram.	35

LIST OF FIGURES

4.1	(Top) MFCC spectral coefficients of TIMIT utterances. (Middle) Spectral difference of MFCC coefficients calculated by subtracting current with previous frame. (Bottom) STM plot with peaks highlighting possible phonetic boundaries. Black vertical lines indicate ground truth phoneme boundaries.	49
4.2	(Top) Spectrogram of the same TIMIT utterances as in Figure 4.1, with window size of 128 samples (8ms) and 50% overlap. (Bottom) Intensity change profile.	50
4.3	(Top) Spectrogram of the same TIMIT utterances, with window size of 32 samples (2ms) and 50% overlap. (Bottom) Intensity change profile. . .	51
4.4	(Top) Spectrogram of the same TIMIT utterances, with window size of 512 samples (32ms) and 50% overlap. (Bottom) Intensity change profile. Black vertical lines indicate ground truth phoneme boundaries.	51
4.5	(1st row) Original spectrogram. (2nd row) Enhanced spectrogram. (3rd row) Intensity change profile without windowing. (4th row) Intensity change profile with windowing.	52
4.6	(Top) Enhanced spectrogram of the same TIMIT utterances. (Middle) STM obtained from MFCC spectrum. (Bottom) Intensity change profile with windowing.	53
4.7	Concatenation of vowel samples, all having high energy in the lower half of the frequency range.	54
4.8	(Left) Intensity profile and (right) standard deviation of the different attribute classes. Spectrogram image Height=129 and corresponds to a frequency range of 0 to 8000 Hz. Black lines correspond to the empirically obtained sub-band divisions.	55
4.9	(Top) Enhanced spectrogram of the same TIMIT utterances. (Middle) STM obtained from MFCC spectrum. (Bottom) Blue: Intensity change profile with windowing without sub-band division and Green: with sub-band division.	58
4.10	(Top) Spectrogram before enhancement. (Bottom) Spectrogram after enhancement and split into the three sub-bands.	60
4.11	(Top) OS plotted against CDR for <i>ISeg</i> . (Bottom) FA plotted against CDR for <i>ISeg</i>	63
4.12	Comparison of segment boundaries obtained from <i>ISeg</i> and <i>HMM-rec</i> . .	67
5.1	Spectrogram generated with 4 ms Hann window, and frequency range of 8000 Hz. (Top) Human speech with no line artifacts. (Bottom) Speech synthesized from <i>MaryTTS</i> [1], a US-based speech synthesis system, with concatenation lines marked in red boxes.	75
5.2	Template used to detect for concatenation lines, with a width of 15 and height N equivalent to the spectrogram height. The central value is set to the width of the template minus one.	76
5.3	Target intensity profile P_{tar}	77

LIST OF FIGURES

5.4	FAR and FRR for Dev 100 and Dev 400 sampled at 16000 Hz, with EER equals to 13.44% and 15.20% respectively.	80
5.5	FAR and FRR for Dev 100 and Dev 400 sampled at 8000 Hz, with EER of 37.35% and 35.50% respectively.	81

LIST OF FIGURES

List of Tables

4.1	Past approaches for speech segmentation	47
4.2	Distribution across different attribute classes	55
4.3	Intensity changes across different attribute transitions and sub-bands. . .	57
4.4	Comparison of <i>ISeg</i> with other unsupervised segmentation algorithms at OS = 0%.	63
4.5	Comparison of results obtained from fusing <i>ISeg</i> with HMM baselines, against <i>HMM-rec</i> and <i>HMM-fa</i>	65
4.6	CDR evaluated at different time window for <i>HMM-rec</i> and <i>ISeg</i>	66
5.1	Database used in experiments.	79
5.2	Definition of various metrics used for evaluation.	79
5.3	Algorithm performance for 16000 Hz Development Sets - Dev 100 and Dev 400.	80
5.4	Algorithm performance on 8000 Hz Development Sets - Dev 100 and Dev 400.	81
5.5	Results on test set.	82

LIST OF TABLES

Summary

The purpose of this research is to examine the use of visual representation and image processing techniques for speech processing applications. This is inspired by the fact that human spectrogram readers can rely solely on visual cues in the spectrogram to perform recognition of words, even in the presence of differing channels and speakers. This suggests that features on the image spectrogram carry important information that can be harnessed for speech processing applications. It is postulated that the image representation better embeds contextual information that is required for human understanding of the speech context. Unfortunately, commonly used speech features, such as the Mel Frequency Cepstral Coefficients(MFCC), are largely frame-based. Therefore, the image representation of speech can serve to complement existing speech features to improve performance of existing speech tasks.

In this work, we developed and applied a solely visual approach to solve speech problems. Two concrete examples of its application are given, so as to perform *unsupervised speech segmentation* and the *detection of unit selection based synthesized speech for anti-spoofing*. We first provide the necessary background by introducing common speech and image processing problems, and then draw the parallel of speech and image processing problems. Next, we introduce an image representation of speech to enable the application of image processing techniques. We then conclude the background with a survey of past attempts that uses image processing techniques on speech and acoustic processing tasks. Next, in the experiments section, we give an in-depth discussion of solving the two example speech tasks using a solely image-based solution. Finally, we wrap up the thesis with conclusions and future work.

LIST OF TABLES

Chapter 1

Introduction

1.1 Motivation

Speech is an important medium of communication for humans. During speech production, the human vocal tract produces acoustic signals that are deciphered by the listener's auditory cortex to make sense of the content spoken. As sound is a signal that is characterized by its constituent frequencies, speech is essentially a signal whose frequency spectrum varies across time, encoding the symbols of human speech communication. The listener's auditory cortex decodes the speech signals by identifying the different frequencies distribution at each time instant.

Motivated by the human auditory process, there have been attempts in the 1960s and 1970s to mimic such approach using the spectrogram [2]. The spectrogram is a 2 dimensional spectral representation, to visualize a signal by plotting the frequency spectral power in the vertical axis and time on the horizontal axis. The work during this period revolves around printing out spectrograms and asking human readers to decipher the speech. One such example of the spectrogram is given in Figure 1.1.

1.1. MOTIVATION

Figure 1.1: An example of a spectrogram with corresponding phonetic transcription.

In [3], trained human spectrogram readers make use of acoustic phonetic knowledge to identify the sounds based on a set of rules and cues such as voice onset time, burst duration, and the frequency distributions of the segment. The spectrogram was also researched by Bell Laboratories researchers who had hoped that spectrographic display could be used to assist college-deaf students to decipher speech [4]. However spectrograms did not prove to be practical due to the complexity of interpreting them. Despite so, it continued to be of research interest in the speech research community [5]. The experiments also highlighted an important observation that visual cues exist in the spectrogram, and it is possible to interpret speech *solely* from a visual representation.

As speech technology progressed, speech processing using spectrogram lost its popularity, and machine learning techniques using spectral representations for speech for automatic speech recognition saw increasing success [6] in the 1980s. To achieve good results and efficiency for such techniques, a good and compact representation of speech is required. These features need to capture the characteristics of the sound, and yet be sampled frequently enough to capture the variations that exist in the human speech. This led to the invention of spectral features such as the Linear Predictive Coefficients (LPC) and derived cepstrum Mel Frequency Cepstral Coefficients (MFCC) [7], and Perceptual Linear Predictive (PLP) coefficients [8]. To extract these features, the signal is first divided into overlapping chunks, or frames, of relatively stationary signals that vary between 20ms to 40ms [7]. These timings are chosen by taking into consideration the speed at which the jaws and tongue can move. Each chunk of signal is analysed using a set of filter banks [7, 8] or a production model [9] to represent each speech chunk by a vector of

1.1. MOTIVATION

coefficients, commonly referred to as a feature frame. The extracted feature frames can then be used to train automatic recognition systems for tasks such as speech recognition [10], language identification [11], speaker identification [12] and etc.. Common machine learning models used include the Hidden Markov Models (HMM) [13], Gaussian Mixture Models (GMM) [14], Deep Neural Network (DNN) [15] to exemplar-based models [16].

Although machine learning techniques with frame-based feature representation have shown impressive results, features such as voice bar, fricatives, voice onset time, are harder to locate accurately from such cepstral representations. Some work illustrating the difficulties to detect these are given in [17, 18, 19]. These features however can be easily located on the image spectrogram as it carries higher level contextual information. For example, fricatives are seen prominently as a noise patch in the higher frequencies on the spectrogram as illustrated in Figure 1.2 labelled ‘sh’. This is because spectrogram is a representation of frequency against time and therefore properties like fricatives, vowels and voice bars, usually identified by spectrogram readers by their frequency distribution, show up better in this representation. In addition, image spectrogram allows the human spectrogram reader to look at its neighbourhood for decision making. For example, although fricatives appear noise-like, presence of surrounding speech-like segments will suggest it is a fricative and eliminate the possibility that it is noise.

Figure 1.2: Fricative in waveform (top) and in spectrogram (bottom).

In summary, despite the effectiveness of machine learning and frame-based features,

early experiments has already highlighted that visual spectrogram contains important cues that can be exploited, and in some circumstances provide more accurate detection for certain speech properties. Its ability to capture contextual relationship can also complement existing frame-based speech features. The motivation of this research is to harness the advantages of visual representation and image processing for speech processing, so as to complement existing speech techniques. This is different to traditional speech research as speech is an *acoustic* signal, but we are looking for a *visual* representation together with appropriate image processing techniques to derive new solutions for existing speech tasks. The proposed framework is generic and can be applied to various speech tasks. Two concrete examples of its application to perform *unsupervised speech segmentation* and *detection of unit selection based synthesized speech for anti-spoofing* are provided in this report.

1.2 Contribution

The contributions of the thesis are:

1. An introduction of how speech problems can be mapped into an image problem in Chapter 2.
2. A literature survey of relevant work that makes use of visual representation and processing to solve speech and acoustic tasks in Chapter 3.
3. A discussion on using the developed image-based solutions for *unsupervised speech segmentation* and *detection of unit selection based synthesized speech for anti-spoofing* in Chapter 4 and 5, published in [20] and [21] respectively.

Although there have been earlier attempts to use visual representations for speech tasks as given in Chapter 3, the testing and tasks are largely limited, and most speech research is still focused on machine learning with frame-based features. This work extends the

use of image representation and techniques to the two speech problems mentioned above, that have not been previously attempted. The results of the experiments showed the effectiveness of using image-based solutions on more speech problems, with competitive results as compared to traditional speech methods. Fusion with existing methods using spectral features is also performed for unsupervised speech segmentation, and achieved improved results, confirming that visual cues carry complementary information which can be exploited.

1.3 Organization of Report

The report is organized as follows.

In Chapter 2, we introduce the readers to how existing speech problems can be converted to image problems. This involves the identification of similar tasks in the image domain, as well an appropriate representation of speech in a visual representation so that image processing techniques can be applied. In this thesis, the image spectrogram is chosen as the representation although other forms of visual representations exist. We then conclude the chapter with a discussion of the properties of the image spectrogram and how acoustic properties manifest itself visually so that readers can better appreciate the relevant work done in using the image spectrogram and image processing for speech and acoustic tasks discussed in Chapter 3.

The next 2 chapters, Chapter 4 and 5, discuss our use of the spectrogram visual representation and developed image processing algorithm to perform phoneme boundary segmentation and detection of unit selection based synthesized speech for anti-spoofing.

Chapter 6 concludes with a discussion on future works and research directions.

1.3. ORGANIZATION OF REPORT

Chapter 2

Converting Speech Problems to Image Problems

To solve a speech problem from an image approach, there is a need to convert an existing speech problem into an image one. We first start with identifying similar speech problems in the image domain, and subsequently discuss about the visual representation of speech so that suitable image processing techniques can be applied. In this thesis, only the image spectrogram is discussed, although other visual representations such as the cochleagram, exists.

2.1 Parallel Between Speech and Image Problems

Many problems are being looked into and addressed in the speech and image research community, but in this thesis we highlight some of the more generic problems found in these two communities, and highlight similar problems as shown in Figure 2.1 below.

2.1. PARALLEL BETWEEN SPEECH AND IMAGE PROBLEMS

Figure 2.1: An illustration of common problems in speech and image research communities.

For speech problems (left of Figure 2.1), the process starts with the extraction of frame-based speech features such as MFCC [7], PLP [8] and LPC [9] from the input speech. These features are then used to train recognition systems to complete the common tasks such as Speaker Identification [12], Language Identification [11], Automatic Speech Recognition (ASR), Voice Activity Detection (VAD) to segment out speech and non-speech portions, Speech Segmentation to segment speech into its constituent sounds, and Enhancement to remove noise in the speech.

Similarly for image problems (right of Figure 2.1) the process starts the extraction of various image features, ranging from lines, edges, texture, intensity and regions. These features are then used to train classifiers to achieve Object Recognition, Image Segmentation to segment out foreground and background, or various objects, as well as Enhancement to sharpen an image, or remove noise.

As illustrated by the yellow boxes in the same figure, the tasks being solved by both the image and speech communities have much in similar. If words spoken in speech are treated as sound objects, then ASR can be treated as an Object Recognition problem. If

2.1. PARALLEL BETWEEN SPEECH AND IMAGE PROBLEMS

we break down an object and treat it as a sum of many parts and matched individually to form the entire object, typically known as ‘Matching by Parts’ [22], a word can also be broken into its constituent sounds and matched individually to form the entire word, which is what is commonly done in ASR techniques [23]. Similarly, a speech segmentation problem to segment out speech into constituent sounds is akin to image segmentation. Enhancement of speech to improve signal-to-noise ratio also has its parallel in image enhancement, where one tries to remove noise from the contaminated image.

It is therefore clear that speech problem has its parallel in the image domain. However, a representation that can represent speech in an image form is required so that the speech’s parallel problem in the image domain can be solved using appropriate image processing techniques. In this thesis, we investigate the spectrogram as the required representation since it has been widely studied since the 1960s and experiments have shown that trained spectrogram readers are able to decipher the spoken content from this visual manifestation. Other forms of visual representation for speech exist as well, such as cochleagram, but are beyond the scope of the thesis. In Figure 2.2, we illustrate how speech problems manifest themselves as image problems on the chosen representation - the image spectrogram. More details on the image spectrogram will be discussed in the following section to understand how speech sounds exhibit themselves on the spectrogram. Understanding of which will allow us to identify the necessary image processing techniques to solve the required tasks.

2.2. IMAGE SPECTROGRAM

Figure 2.2: Manifestation of speech problems as image problems on the spectrogram.

2.2 Image Spectrogram

2.2.1 Spectrogram Production

The spectrogram is widely used to visualize a speech signal in the frequency domain as it varies across time. Since speech signals are mostly digitized, the discrete version of Short-Time Fourier Transform (STFT) is used to generate the spectrogram [24] by

$$X(k, n_s) = \sum_{m=0}^{L-1} w[m]x[n_s + m]e^{-j\frac{2\pi k}{N}m}, k = 0, 1, \dots, N-1 \quad (2.1)$$

where n_s is the starting sample index, $w[m]$ refers to the windowing function, L refers to the window length, and N is the length of the signal. The k values represents the frequency where $\omega_k = 2\pi k/N$ is the frequency at index k . Essentially, STFT takes a small window of the signal and performs Fourier transform. The window is then shifted to the right, with some overlap to the previous window to reduce boundary artefacts.

$X(k, n_s)$ has complex values with magnitude and phase, but the spectrogram is only

2.2. IMAGE SPECTROGRAM

the magnitude of the Fourier Transform coefficients, i.e.

$$\text{spectrogram}(X) = |X(k, n_s)|^2 \quad (2.2)$$

and by taking logarithm on the spectrogram, we have

$$\log - \text{spectrogram}(X) = \log(|X(k, n_s)|) \quad (2.3)$$

The log-spectrogram is used as the human perception of sound is approximately logarithmic in nature [25]. An example of a log-spectrogram is shown in Figure 2.3 below. It represents the TIMIT speech sequence “She had your dark suit in greasy wash water all year”.

Figure 2.3: Log-spectrogram, with window size = 128 samples (8 ms for sampling frequency of 16000Hz). Each pixel $20\log(|X(k, n_s)|)$ represents the power at time n_s and frequency index k , with darker intensities corresponding to higher power.

Spectrograms can be further divided into two types, narrow-band and wide-band spectrogram. Narrow-band spectrograms are generated using a longer window, typically 20 milliseconds [26], during STFT calculations and emphasizes frequency changes. Wide-band spectrograms on the other hand, are generated using a shorter window, usually 2 to 4 milliseconds, to emphasize temporal changes. Depending on the application, different window size can be chosen. For example when the application requires accurate localisation of a sound, then a wide-band spectrogram should be adopted. On the other hand, if the application requires the accurate recognition of the sound based on the frequency distribution, then the narrow-band spectrogram is preferred since frequencies

2.2. IMAGE SPECTROGRAM

show up more distinctly in them. Examples of narrow-band and wide-band spectrograms are given in Figure 2.4.

Figure 2.4: (Top) Narrow-band and (bottom) wide-band spectrogram of same signal in Figure 2.3 with window of 15 milliseconds and 3 milliseconds respectively .

2.2.2 Speech and its Appearance on the Spectrogram

With the understanding that the spectrogram is a frequency versus time representation of the underlying acoustic signals, we now discuss how speech acoustic phenomenon *appear* in the spectrogram. The materials discussed here are collated from various knowledge sources [27, 28, 29, 30].

Different languages comprise of different sounds and the smallest unit of speech that distinguishes a word from another is termed the *phoneme* [31]. For example, sounds like ‘zh’, ‘ch’, ‘sh’ exists only in the Mandarin language, but not in English. To simplify the discussion, we will discuss phonemes in 5 broad attribute classes, which are namely Vowels, Fricatives, Stop/Plosives, Nasals and Approximants, since phonemes in the same class exhibit similar properties. The last 4 classes belongs in the generic ‘Consonant’ classes as well.

2.2. IMAGE SPECTROGRAM

Voicing

Voicing happens when vocal cords vibrate during the production of voiced sounds, which can happen during the production of both vowels and certain consonants. The laryngeal muscle controls the frequency of the vibration and causes the airflow to become air pulses. As such, voicing appears as short vertical lines in the low frequencies on wide-band spectrograms corresponding to the pulsing of the vocal folds. A series of these vertical lines forms a bar which is called the ‘voice bar’ as seen in Figure 2.5. Men and women have different frequencies for this voice bar, but they generally appear from 100 to 250 Hz.

Figure 2.5: Voice bar seen together with voiced stop ‘d’.

Vowels

Formants are dark horizontal lines seen on the spectrogram due to vocal tract resonance and are associated with vowels, but also for nasals and approximants. As these resonances are usually maintained for a period of time as air flows through the vocal tract, the resonating frequencies show up as dark lines in the spectrogram at the corresponding frequencies. These formant frequencies however differ across each individual as everyone has a different vocal tract. Women and children generally have a longer vocal tract than men, leading to higher frequencies of the formants produced.

As vowels are produced by an open vocal tract, they therefore have strong resonance and formant structures. They are also periodic. The first three formant frequencies, usually termed F_1 , F_2 and F_3 , are believed to carry the most important cues to vowel identification, with F_1 giving cues of the jaw openness and F_2 giving cues of the tongue

2.2. IMAGE SPECTROGRAM

position. The different configuration of the articulators therefore affects the shape of the formant frequencies. The structure of these 3 formants helps spectrogram readers identify which vowel it is. However, consonants before and after the vowel modifies the structure of the vowel structures. For example, bilabial stop sounds causes lowering of the formant structures.

Nasals

Nasals such as 'm' and 'n' also exhibit formant structures as there is still resonance in the vocal tract. In nasals, however, the main resonance comes from the nasal cavity as air is mainly forced through the nose as the oral cavity is obstructed by the lips or the tongue. This gives a strong low frequency resonance at around 200 Hz. The oral cavity becomes a side resonance chamber and the effect of having a main and side resonance chambers is the existence of anti-resonance which cancels out some frequencies, causing poles in the frequency spectrum. The effect of this cancellation gives nasal their iconic appearance of a 'hole' in the spectrogram as anti-resonance causes the cancellation of the F_2 .

Approximants

Approximants such as 'l' are consonants that have vowel-like structures, which make them very hard to distinguish between vowels even for spectrogram readers. Identifying approximants in spectrograms can be very challenging as they are vowel-like, and can only be identified by looking at formant transitions.

Stops/Plosives

Stops/plosives such as 'b', 'p' shows up as thin vertical lines in the spectrogram as they are caused by the sudden release of air built up behind closure such as the lips or the teeth. When pronouncing such sounds, the vocal tract does not come into play and no resonance can be observed. They are also aperiodic. Therefore, unlike the vowels, they

2.2. IMAGE SPECTROGRAM

do not have horizontal formant frequencies in their spectrogram appearance.

Fricatives

Fricatives such as 'f' and 's' are produced by forcing turbulent air through a narrow constriction and are associated with fricative sounds. Due to this turbulence, they sound pretty noise-like and are seen as a patch of high frequency noise on the spectrogram.

To better illustrate the different classes and its appearance on the image spectrogram, the different classes are labelled on a portion of the TIMIT sequence in Figure 2.6. The strong formant structures can be seen in the vowel, nasals and approximant classes. Nasals on the other hand can be better distinguished due to the lack of some formants due to anti-resonance as evident when we look at the vowel transition from 'ix' to 'n' in the illustration. Fricatives are easy to detect as they look like noise patches in the high frequencies of the spectrogram. Stops exhibit themselves as a short vertical streak due to the burst.

Figure 2.6: TIMIT sequence “She had your dark suit in greasy wash water all year”. The abbreviation is as follows: vow - vowels, nas - nasals, app - approximants, stop - stop, fri -fricatives.

2.2.3 Variations a Spectrogram Undergoes

Effects of Different Speakers

Properties such as voicing and formant frequencies can vary amongst gender. They can also vary between different speakers. The difference is attributed to the different vocal

2.2. IMAGE SPECTROGRAM

tract for each individual. Other than the length of the vocal tract leading to different frequencies, the difference between the articulatory organs of each individual will affect the appearance of the formant frequencies as the resonance chamber shape is affected. In addition, other properties such as speaking rate and style can also be different. Therefore the spectrogram for each individual can vary.

Figure 2.7: Spectrograms of two different speakers, one male and one female, saying the same utterance “She had your dark suit”.

As can be observed in Figure 2.7, when two speakers speak the same utterance, the spectrogram can look significantly different. The different speaking rates cause the segments to be misaligned in the horizontal axis. Different speaking styles and pronunciation can also cause spectrograms to be different in both time and frequency distribution.

Effects of Noise

The spectrogram image undergoes various changes under the effect of noise as illustrated by Figure 2.8.

2.2. IMAGE SPECTROGRAM

Figure 2.8: Spectrograms of the same utterance under (top) clean conditions, with (middle) white noise of amplitude 0.03 (maximum amplitude 1.0) added to the utterance and (bottom) DTMF tones of amplitude 0.2 added to the utterance.

In the presence of white noise, speech clarity is reduced and appearance of the speech in the spectrogram becomes less obvious. As vowels have formant structures that exhibit themselves as lines in the spectrogram and looked significantly different from the random pattern noise exhibits, the formant lines can still be seen in the noise polluted spectrogram. On the other hand, fricatives and stop have noise-like properties, and hence became consumed by the added white noise and loses its appearance in the spectrogram. Depending on the types of noise encountered, different parts of the spectrogram will be affected. If tonal noise with strong harmonics occurred alongside the speech, formant lines can be badly distorted instead as illustrated by the bottom image in 2.8.

On the other hand, reverberation causes a smearing effect on the spectrogram as illustrated in Figure 2.9. This is expected as reverberation is the phenomenon where multiple copies of the same speech are heard at different delays. This is akin to duplicating the image spectrogram at various time delays, which resulted in the smearing effect. Depending on the time delay, the effect on the spectrogram can range from a

2.2. IMAGE SPECTROGRAM

slight blur to visualizing multiple copies of the spectrogram as observed in the middle and bottom image respectively. As reverberations manifest themselves as a blur effect on the spectrogram, there has been recent efforts to use blur kernel estimation methods used in image processing to perform reverberation time estimation [32].

Figure 2.9: Spectrograms of the same utterance with reverberation added to the middle and bottom utterances. (Middle) Reverberation delay of 0.02s is added. (Bottom) Reverberation delay of 0.1s is added. In both examples the decay is set to 0.5.

It is therefore evident that spectrogram images undergo various changes under the influence of speaker changes and noise.

2.2.4 Features on the Spectrogram Image

With an understanding of how speech exhibit themselves visually, and the variations spectrograms can undergo, we now introduce the various image features that can be extracted from the spectrogram and the relevant image feature extraction techniques.

2.2. IMAGE SPECTROGRAM

Colour Intensity

For camera captured images, the intensity represents the amount of light that falls onto the sensors. However, for spectrogram images, the intensity at each pixel does not measures light, but measures the power of the signal with respect to time align x-axis, and frequency align y-axis. The value can be rescaled to a desired range by quantizing the values either to 1 byte range or 3 bytes, depending on the required level of precision. The two quantization will produce a grayscale or a coloured spectrogram respectively as shown in Figure 2.10.

Figure 2.10: STFT coefficients of one utterance mapped to (top) 0-255 range to obtain grayscale images and (bottom) 0-16,777,215 range to obtain coloured image. Red indicates higher values.

Since the pixel intensity is a directly correlation to the STFT coefficients, colour intensity information can be useful in recognizing areas of speech or non-speech, or amplitude fluctuations in the speech.

Segment

In image processing, segments usually refer to areas with homogeneous properties, such as colours. Region growing techniques are commonly used to segment the images into

2.2. IMAGE SPECTROGRAM

patches of similar properties. Segments are important as they provide many important properties on the object, such as the size and the shape. The relative positions of different segments also help to establish important contextual relationships. For example a blue patch above a yellow patch might suggest a sandy beach. Such concepts can similarly be extended to spectrograms. If a segment on the upper half of the spectrogram is to the left of a segment that occurs on the bottom half of the spectrogram, there is a high possibility this marks a transition from a fricative to a vowel or nasal sound. Most importantly, if segments of homogeneity can be established, such cues can perhaps be incorporate to current speech processing to delineate speech segments that should be considered in totality when incorporating context.

Line

In spectrogram images, lines mainly exist in the formant regions at the lower frequencies as seen in Figure 2.11. In wideband spectrogram vertical lines can also be observed as seen in Figure 2.4. Due to presence of formant lines, Gabor filters can be used to detect and process these lines [33, 34] to improve speech recognition accuracy. For example, [35] used a set of 60 Gabor filters of different orientations are applied to the log-spectrogram to generate a feature vector which is appended to existing features for speech recognition. The results showed that the addition of Gabor filters in tandem with spectral features can improve the robustness of the system and reduced the word error rate.

Other than detecting the presence of lines using Gabor filters, tracking of such lines can also be performed, which can have very important applications to build a phoneme classifier. Spectrogram reading knowledge suggests that spectrogram readers scrutinize the structure of such formant lines to decode the identity of various phonemes. Tracking of lines can also help in identifying speech segments as a discontinuity in the line will suggest the end of a segment. Tracking of formant lines can also be used to detect for pitch fluctuations in speech.

2.2. IMAGE SPECTROGRAM

Texture

Other than structured textures such as lines above, other textures can also exist that do not conform to a fixed pattern, such as a noise-patch due to fricative as that seen in Figure 2.11. Techniques like Gabor filters will not be effective. Hence other techniques such as Local Binary Pattern (LBP) [36], which is commonly used for texture analysis and segmentation, can be applied. The LBP approach represents the underlying texture using a binary pattern with each bit encoding the difference between a pixel and its neighbouring pixels. LBP however is sensitive to noise, but robust to intensity variations in the image, and newer variants of it exist such as the Local Ternary Pattern (LTP) [37]. There have also been attempts to use LBP to detect for synthesized speech [38, 39] as it is observed that synthesis speech exhibits some texture anomalies as compared to natural speech.

Figure 2.11: Different textures found in spectrogram.

In addition, these textures can come in useful for classification of segments into their different attribute classes since only vowels, nasals and approximants exhibit lines, while fricatives exhibits more random texture.

Edge

Edges in images are characterized by a sharp change in intensity, and are important features frequently used in image processing signifying regions of change caused by depth, materials, or object. Many operators exist to perform edge detection, such as the Prewitt

2.2. IMAGE SPECTROGRAM

and Sobel operators [40]. More commonly used edge detection algorithm is the Canny Edge Detector [41].

However, in spectrogram images, sharp changes in intensity are only pronounced in the horizontal direction, which corresponds to the time axis. In clean conditions, sharp intensity changes in the time axis often corresponds to phoneme boundary changes. Changes in the vertical direction corresponding to the frequency axis and is less pronounced as the intensity changes usually smear over a distance. It is challenging to detect edges in the vertical directions.

Chapter 3

Past Works relating to Image Processing Techniques for Speech Processing Tasks

With an understanding of spectrogram and how various speech phenomenon exhibit itself as various image features, we will discuss in this chapter about the attempts at performing speech and acoustic processing using visual representations. These include early attempts at using human spectrogram readers to interpret visual spectrograms as well as later attempts to use image processing on various visual representations to solve speech and acoustic problems. The discussion highlights the feasibility of using visual representation and image processing techniques to perform speech processing, and also helps to highlight the way forward for this research.

3.1 Speech Processing using Spectrogram Reading

Attempts to use spectrograms for speech recognition started as early as 1970s. One early attempt was performed in Haskins Laboratories, by Frances Ingermann and Paul Mermelstein [42]. The purpose of their work was to discover if the cues used by human

3.1. SPEECH PROCESSING USING SPECTROGRAM READING

spectrogram readers for phoneme recognition can be effectively used by an automatic speech recognition system. In their experiment, the subjects were given five printed spectrogram of test sentences, and tasked to locate 10 words also given in printed spectrograms. There were a total of 3 subjects in the experiments, and all achieved high accuracy in the task. However, when the size of the vocabulary increased, the task became more challenging and the accuracy dropped. Another attempt was performed by Dennis H. Klatt and Kenneth N. Stevens [43]. In their experiment, 19 spectrograms of sentences were given to experimenters, who were asked to write down its phonetic transcription and subsequently determine the sentence from a vocabulary of 200 words. A high error rate was obtained, showing that it is a challenging task.

In 1977 [44], the possibility of using a parametric representation of the spectrogram to overcome the need to store raw spectrogram was proposed. Two human spectrogram analysts were each asked to list a set of 16 spectrogram features which they felt were useful for them to perform word identification. These features were then applied for a word retrieval task. Findings showed that 16 features were insufficient to capture all the phonetic variation that a word can undergo and more features would be required to improve retrieval performance. In addition, the manual labelling of the 16 features was extremely time consuming and have limited this approach's feasibility.

In 1979, Zue also conducted spectrogram experiments [45]. Instead of performing spectrogram matching, his experiments required the test subject to produce the phonetic labelling from the spectrogram. The experiments were first performed by Zue and he obtained an 85% accuracy. In addition, he trained 5 graduate students over a period of 13 weeks, 3 hours a week. At the end of the course, the students were asked to perform spectrogram reading on 5 spectrograms. The students were also able to achieve good performance, with an accuracy of 83% was reported. However, the students had required much more time as compared to Zue to decipher the spectrogram. The experiments showed that information can be extracted from the spectrogram and phoneme identification can be taught and realized.

3.1. SPEECH PROCESSING USING SPECTROGRAM READING

With these encouraging results, Zue collaborated with Lamel to realize an automatic speech recognition system based on spectrogram reading knowledge [46]. This work was very much motivated by [47] which had proposed the idea of a three-person team speech spectrogram expert system. They were namely the vision expert, who extracts visual cues on the speech spectrogram; the acoustic-phonetic expert, who associates visual features to phonemes; and the phonetics expert, who decides the allowable phoneme sequences. In [46], the features were first extracted manually by experts from the spectrogram. These features include voice onset time, burst location and strength, and formant transitions before closure and after release. A total of 26 acoustic features were used. Next, rules on how to differentiate the various phonemes from this feature set were applied and refined iteratively to build an expert system for stop consonants identification. Results of the experiments showed that the system performed as well as the human expert on the training set. However, in the testing conditions, the system performed slightly worse than human. In [48], the performance of this knowledge-based framework was compared to the performance of an HMM-based system trained on the TIMIT corpus, and results showed that knowledge-based system out-performed the HMM system.

Other attempts to build such an expert system include [49] and [50]. In [49], acoustic features were automatically extracted from the digitized spectrogram using peak detection. The frequency and time associated with the peaks in the spectrogram intensity were used as features. These features were then used to build an expert system for phoneme recognition. Although these results showed that the extracted features from spectrogram for phoneme recognition is feasible, the results are preliminary as the testing set was very limited, only 3 test utterances. In [50], a Japanese phoneme segmentation expert system was built based on acoustic features extracted from the spectrogram such as frequency range, time range and etc.. The reported results showed that the produced segmentation is as good as the human labelled ones.

The many attempts from the early works have shown that visual cues exist in the spectrogram that can be used for speech recognition, without even listening to the audio

signal.

3.2 Speech Processing using the Image Spectrogram

3.2.1 Image Processing for Speech Signal Processing

The works earlier discussed in Section 3.1 have relied on humans to process the spectrogram. In this section, we discuss works that use image processing to automatically process visual representations of speech to solve speech tasks.

To automatically extract features from the spectrogram, [51] used the output from edge detection on the spectrogram to create a feature representation of the spectrogram that resembled a “cartoonized” representation of the phonemes. This feature was used to perform simple vowel recognition with 112 vowel tokens from 8 speakers. Results were promising, achieving a first choice accuracy of 83%. However, the authors recognized that conclusions cannot be drawn based on the limited test set.

Other attempts [52, 53] to automatically extract features using image processing techniques on spectrograms exist. In these two works, morphological operations were used to obtain features to describe phonetic segments. The difference being the former used a skeleton of the spectrogram, while the latter argued that the blob was a more useful description for phonetic segments. In both works, the authors concluded that the extracted features can be used for phoneme recognition to a certain extent.

Other than using the blobs extracted from spectrogram to directly perform speech recognition, there were also attempts to use image processing techniques to identify relevant regions in the spectrogram to assist in current Automatic Speech Recognition (ASR) techniques. E.g. [54, 55, 56, 57] use region growing methods to extract relevant speech portions in a Gaussian polluted noise signal. This is achieved by first dividing the spectrogram into grids and calculating the local mean and variance in each grid cell. Points with the highest signal-to-noise ratio in a region are chosen as seeds and a region is grown from it, with the region corresponding to a segmented speech portion.

3.2. SPEECH PROCESSING USING THE IMAGE SPECTROGRAM

The authors believed that speech segmentation using this approach is more robust to Gaussian noise. The segmented portions are then used as a mask to extract the speech portions used to perform ASR based on the Aurora experimental framework [58] and was shown to have improved performance over using the entire noisy signal.

More complex feature descriptors from the image spectrogram have also been attempted in [59]. In this work, Fourier descriptors were obtained from the spectrogram to approximate shapes of different semivowels, which are also known as glides and are sounds such as “w”, to perform semivowel classification. Fourier descriptors are frequently used in image processing for shape matching and have the promising property of being scale, shift and rotation invariant [60]. Simply speaking, Fourier descriptors take the outline of a shape and approximate it using Fourier coefficients. As an extension to this work [61], the Fourier descriptor was combined with other features to perform semivowel classification. Principal Component Analysis [62] was also used to reduce the feature dimension.

Filters have also been used to extract features from the spectrogram. In [63], 291 manually designed filters corresponding to line patterns on the spectrogram were applied. Boosting [64] and feature selection were augmented to extract the more discriminating lines to perform isolated digit classification and achieve results better than HMM. However, the authors highlighted that HMM models were used to perform time alignment on the utterances as a pre-processing step and hence the comparison to HMM is unfair.

Recently, Dennis et al [65] experimented with using Generalized Hough Transform to perform phoneme recognition. Generalized Hough Transform (GHT) is commonly used in image processing for object recognition as it can accumulate spatial evidence surrounding the location of the target class. This idea is similarly adopted for speech image features. In this case, evidence was derived from MFCCs, and spatial locations were only restricted to the 1-Dimensional time index. The accumulated evidence was used to form a feature vector to train a Neural Network phoneme classification. This framework achieved state-of-the-art performance, with a lower frame-level error rate as

3.2. SPEECH PROCESSING USING THE IMAGE SPECTROGRAM

compared to DNN approaches.

There are also attempts to use spectrograms for singing speech. In the work discussed in [66], the authors used spectrogram matching to detect 5 Thai phonemes in singing clips. Although results are promising, the test set was very limited.

Beyond recognition, there is also an attempt recently to use image processing deblurring techniques to estimate the reverberation time [32] in reverberant speech. This technique treats the whole spectrogram as a blurred image, and attempts to find the image kernel that can return the blurred image to its focused form. The amount of blurring, which is given by the estimated blur kernel gives the reverberation timing.

3.2.2 Image Processing for Acoustic Signal Processing

In addition to speech signals, image processing techniques have also been applied in the area of acoustic signal processing. In the survey performed in 2010 [67], the authors summarized the image processing techniques that had been done to perform underwater mammals and vehicles classification using spectrograms. Common image processing techniques such as edge detector, morphological operator and particle filter were discussed. Other than underwater acoustic signals, spectrogram images have also been exploited for music genre classification. In [68] and [69], textures from spectrograms are used as features. In [25], similar texture features were used to classify monophonic sequences to musical instrument.

Recently, Dennis et al. [70, 71, 72, 73, 74] had also published many works regarding the use of spectrogram features for acoustic event detection. They believed that using spectrogram features for sound event classification would be more robust to noisy channel as compared to conventional methods using MFCC. In [70] [71], Local Spectrogram Features (LSF) was proposed as a feature to detect prominent points in the spectrogram. Specifically, neighbourhood of prominent points were used with GHT and Neural Network for event detection. In addition, they also propose the Spectrogram Image Feature (SIF) [72] which is a vector summarizing the intensity statistics of 9 by 9 pixel blocks as

feature. To improve robustness, the sub-band power distribution (SPD) image [73] [74] was further proposed. Essentially, instead of the time-frequency representation used in SIF, the SPD image was a power-frequency representation on a segment of signal.

3.3 Summary

In this chapter, a review of early works that make use of spectrogram reading for speech processing is given. Although the success has only been shown on limited task, these work highlighted the possibility of “seeing speech” and using visual cues on the spectrogram for speech processing. We then gave an overview of speech and acoustic processing task that made use of image processing solutions in recent years. It can be seen that despite the attempts, the work is limited, and most speech research is still focused on machine learning with frame-based features. We therefore hope to extend the usage of image processing techniques for speech to more speech problems.

In the next two chapters, we discuss our developed image-based solutions for two speech tasks, namely unsupervised speech segmentation and the detection of unit selection based synthesized speech for anti-spoofing, which has not been previously attempted.

3.3. SUMMARY

Chapter 4

Unsupervised Speech Segmentation

4.1 Introduction

This chapter examines the use of visual cues for unsupervised speech segmentation, in other words, to automatically find the boundaries, often phonetic boundaries, of acoustic elements in the speech. Our contribution is published in [20]. This problem has been of constant interest in the speech community due to the importance of having accurate and consistent segment boundaries for acoustic model training for Automatic Speech Recognition, as well as for speech synthesis [75].

Traditionally, phonetic boundaries are obtained by human experts manual annotating. This is laborious, and may vary across different experts. In addition, it is also hard to find experts for low resource languages [76]. Hence, attempts for automatic speech segmentation have been proposed and these techniques can be divided into two classes, supervised versus unsupervised. Supervised algorithms involve the use of labelled training data to first train acoustic models, such as Hidden Markov Models (HMMs) [13], and subsequently use them for segmentation. Examples include those discussed in [77, 78, 79, 80, 81, 82]. As this class of techniques require manually transcribed data, they are less attractive for low-resource languages. The other class of techniques, the unsupervised methods do not require labelled training data [83, 75, 84, 76, 85, 86].

4.2. RELATED WORKS

The proposed framework in this chapter belongs to the unsupervised category and examines visual cues for speech segmentation. Other than evaluating the results on the TIMIT corpus, fusion of the the proposed algorithm's result with that obtained by HMM showed that spectrogram image features provides additional information that can improve current speech segmentation.

The rest of the chapter is organized as follows. Section 4.2 discusses past speech segmentation works. Section 4.3 presents an analysis of the signal characteristics of phonetic boundaries in image spectrogram. Section 4.4 presents the proposed algorithm and experimental results are provided in Section 4.5. Finally, the conclusion and future works are presented in Section 4.6.

4.2 Related Works

This section discusses 9 recent works for speech segmentation. It is however difficult to compare the different algorithms as they use different evaluation metrics. Even for accuracy, the definition of hit varies across different works and some deemed a detection as a hit if it is within 10 ms of a ground truth boundary, while others use different values. Fortunately, most works provided results for 20 ms and hence we also adopted this threshold to dictate a hit. We compared the performances of the proposed methods using the following 3 metrics defined in [84], which are *correct detection rate (CDR)*, *over-segmentation (OS)* and *false alarm (FA)*, defined in Eqs. (4.1) to (4.3) below:

$$\text{CDR} = \frac{\text{Number of correct boundaries}}{\text{Number of true boundaries}} \times 100\%, \quad (4.1)$$

$$\text{OS} = \left(\frac{\text{Number of boundaries found}}{\text{Number of true boundaries}} - 1 \right) \times 100\%, \quad (4.2)$$

$$\text{FA} = \left(1 - \frac{\text{Number of true boundaries found}}{\text{Number of boundaries found}} \right) \times 100\%, \quad (4.3)$$

4.2. RELATED WORKS

Table 4.1 summarized the performance for various methods. As most published works only provided CDR values, but not OS and FA statistics, we had to leave out some of the fields. The last two columns indicate if these methods are supervised/unsupervised algorithms, and the use of TIMIT corpus to evaluate their results.

Method	CDR (%)	OS (%)	FA (%)	Supervised?	TIMIT?
Svendsen & Kvale (1990) [81]	88.1	-	-	Yes	No
Micallef & Chilton (1997) [80]	73.8	-	-	Yes	Yes
Pellom & Hansen (1998) [82]	71.0	-	-	Yes	Yes
Cucchiarini & Strik (2003) [79]	85.9	-	-	Yes	Yes
Dusan & Rabiner (2006) [75]	75.3	12.8	33.2	No	Yes
Aversano, Espito,	73.6	0	-	No	Yes
Espito & Marinaro (2001) [83]	90.0	63.0			
Estevan, Wan	76.0	0	26.0	No	Yes
& Scharenborg (2007) [84]	90.3	75.5	47.0		
Qiao, Shimomura,	76.7	0	-	No	Yes
Minematsu (2008) [85]	77.5	-	-		
Niekard & Barnard (2008) [76]	75.0 - 81.0	-	-	No	No

Table 4.1: Past approaches for speech segmentation

In [76], the accuracy reported was evaluated using 16ms time window for 3 different languages. Therefore, a range of values was given in the table.

Generally, supervised methods that use HMM like those in [79] achieve considerable high accuracy. If the text transcription without time labels is available, forced alignment can be used and can achieve a reasonably good alignment. Hence this method has been widely adopted for speech segmentation and alignment. Despite so, many researchers have reported that there are still significant discrepancies between the found boundaries obtained by HMM methods versus manually labelled boundaries [76]. This explains why

4.2. RELATED WORKS

research is still ongoing to improve speech segmentation.

Unsupervised methods on the other hand rely purely on signal properties for segmentation. Much of the methods in Table 4.1 use acoustic properties such as spectral transition, intensity changes, F0 voicing, and etc. for phonetic boundaries detection. Spectral transition is the most common feature and studies have shown that they correlate closely with phonetic boundaries [75]. Most approaches first calculate the change in spectral properties at each time instance and then either adopt a peak selection algorithm to sieve out potential phonetic boundaries, or use clustering methods to arrive at the final segmentation [84, 85]. In [84], Maximum Margin Cluster (MMC) was used to separate frames within a time window into one, or two clusters. The window is slid across the whole utterance and a plot of cluster labels against time is derived. Phonetic boundaries exhibit themselves as slopes in the plot and such structures are detected to identify segment boundaries. In [85], the authors obtained a rate distortion measure, based on information rate distortion theory, computed on the MFCC feature to perform segmentation. Clustering algorithm is then used to optimize on this objective measure to find the final segmentation. A comparison between the supervised and unsupervised algorithms showed that unsupervised approaches generally suffer a 10% degradation in CDR (at 0% OS), sitting at around mid-70%, as compared to the HMM segmentation, with 85.9% CDR. Much higher CDR can be observed if higher OS and FA rates can be tolerated.

In this work, we examine visual cues with an unsupervised approach to complement existing algorithms to improve speech segmentation, and compare against the MMC [84] and Rate Distortion (RD-DIA) [85] as these are the two most recent unsupervised works that achieved the best results. If visual cues such as intensity changes and edges can provide additional information on top of acoustic cues, the output can be fused to achieve better speech segmentation. In the following chapter, an analysis of how phonetic boundaries exhibit themselves in the image spectrogram as compared to the commonly used spectral features is given.

4.3 Signal Analysis

4.3.1 Spectral versus visual features for boundary detection

It is reported in the literature that phonetic boundaries corresponds to areas of high spectral changes in the MFCC representation [75]. In Figure 4.1, two examples of the MFCC spectral coefficients for two truncated TIMIT sequences is shown.

Figure 4.1: (Top) MFCC spectral coefficients of TIMIT utterances. (Middle) Spectral difference of MFCC coefficients calculated by subtracting current with previous frame. (Bottom) STM plot with peaks highlighting possible phonetic boundaries. Black vertical lines indicate ground truth phoneme boundaries.

The MFCC configuration used here is a 25 ms window and 10 ms frame shift. In the last row, the spectral transition measure (STM) as calculated in [75] and given in Equations 4.4 and 4.5 is plotted. As can be observed, the true phonetic boundaries, indicated by black vertical lines, do correspond to areas of high spectral changes as reported in literature. However, noise at the beginning of the sequences from frame 0 to

4.3. SIGNAL ANALYSIS

15 also gave rise to prominent peaks that can result in false alarms.

$$STM(m) = (\sum_{i=1}^D a_i^2(m))/D \quad (4.4)$$

$$a_i(m) = (\sum_{n=-I}^I MFCC_i(n+m) * n) / (\sum_{n=-I}^I n^2) \quad (4.5)$$

where m is the frame index and I represents the number of frames used to compute the regression coefficient $a_i(m)$. $I=2$ here.

Figure 4.2 shows the image spectrogram of the two examples discussed in Figure 4.1 and shows that phonetic transitions exhibit themselves as regions of high intensity change in spectrogram. The intensity change is calculated by evaluating difference between two adjacent frames. The results of found boundaries using this difference is plotted with the ground truth boundaries indicated with black vertical lines.

Figure 4.2: (Top) Spectrogram of the same TIMIT utterances as in Figure 4.1, with window size of 128 samples (8ms) and 50% overlap. (Bottom) Intensity change profile.

The window size used to generate the image spectrogram is an important parameter for the experiment. In the above example, the window size is set to 128 which corresponds to 8ms as TIMIT sequences are sampled at 16000 Hz. When the window size is decreased to 32 (2ms) as in Figure 4.3, noise in the spectrogram increases to the extent that

4.3. SIGNAL ANALYSIS

phoneme boundaries become hard to distinguish from intensity changes caused by noise. On the other hand, a larger window size of 512 (32 ms) results in poor time localization on the spectrogram, causing some phoneme boundaries to be missed, as illustrated in Figure 4.4. An appropriate window is therefore important.

Figure 4.3: (Top) Spectrogram of the same TIMIT utterances, with window size of 32 samples (2ms) and 50% overlap. (Bottom) Intensity change profile.

Figure 4.4: (Top) Spectrogram of the same TIMIT utterances, with window size of 512 samples (32ms) and 50% overlap. (Bottom) Intensity change profile. Black vertical lines indicate ground truth phoneme boundaries.

It can be observed that the image spectrogram (Figure 4.2) generated from the STFT

4.3. SIGNAL ANALYSIS

coefficients may introduce grainy artefacts which causes noise in the intensity profile. They can be removed by image enhancement methods such as median filtering [87], which is effective for removing salt-and-pepper noise in images and can be applied in this case. Together with histogram equalization [88] to boost the contrast of the image, an enhanced spectrogram can be obtained as in the 2nd row in Figure 4.5. Sharper intensity change profile can then be obtained from the enhanced spectrogram as shown by the 3rd row in Figure 4.5.

Figure 4.5: (1st row) Original spectrogram. (2nd row) Enhanced spectrogram. (3rd row) Intensity change profile without windowing. (4th row) Intensity change profile with windowing.

To further reduce spurious peaks, windowing can be used when computing intensity

4.3. SIGNAL ANALYSIS

change at each time instance to obtain the 4th row. The final intensity change profile has peaks that correspond closely with phoneme boundaries and spurious peaks in the early portions of the utterance are not observed.

Comparing the two approaches, one using MFCC and one using the image spectrogram, the algorithm resulted in different boundary decisions. More interestingly, the two examples suggest that both spectral and image representations can complement each other, as indicated in Figure 4.6. Fusion of these two techniques can possibly improve the individual systems.

Figure 4.6: (Top) Enhanced spectrogram of the same TIMIT utterances. (Middle) STM obtained from MFCC spectrum. (Bottom) Intensity change profile with windowing.

4.3.2 Sub-band Division to Enhance Boundary Detection

Next, we examine if the division of the spectrogram into multiple frequency sub-bands will help to improve boundary detection and make phoneme transitions more prominent.

To decide on the sub-bands, intensity profiles of 4 attribute classes, namely stops,

4.3. SIGNAL ANALYSIS

fricative, vowels and nasals, is obtained. As sounds belonging to the same class tend to exhibit similar characteristics, we will only focus on the attribute level. Approximants are not considered as they exhibit similar characteristics to vowel, and we therefore merge them with the vowel class.

To plot the intensity profiles, the spectrograms of all the TIMIT utterances are enhanced with median filtering and segments of phonemes that belongs to the same attribute class as indicated by the TIMIT ground truth are concatenated as in Figure 4.7. Next, concatenated image is averaged across the horizontal time-axis to form an intensity profile which plots the intensity mean against the image spectrogram height, which corresponds to the 129 FFT frequency bins, 1 pixel per bin. Essentially, the intensity profile shows which are the frequencies an attribute has strong intensity concentrations in. This is performed for all the attribute classes to obtain the graph in Figure 4.8. The standard deviation across the samples is also calculated in a similar manner and the number of samples across the 4 attribute classes in the TIMIT training set is given in Table 4.2.

Figure 4.7: Concatenation of vowel samples, all having high energy in the lower half of the frequency range.

4.3. SIGNAL ANALYSIS

Attribute	No. of frames/columns in spectrogram
Stops	991745
Fricatives	465238
Vowels	1513862
Nasals	179442

Table 4.2: Distribution across different attribute classes

Figure 4.8: (Left) Intensity profile and (right) standard deviation of the different attribute classes. Spectrogram image Height=129 and corresponds to a frequency range of 0 to 8000 Hz. Black lines correspond to the empirically obtained sub-band divisions.

Inspecting the intensity mean (left of Figure 4.8) across the 4 different attribute classes, it can be observed that nasals have the highest intensity concentration in the lower 10% of the frequency range and vowels also have intensity concentrated in the lower half of the frequency range. This is consistent with that discussed in Section 2.2.2. Fricatives on the other hand have intensity concentrated in the upper half. Stops are different from the other 3 classes, with no obvious high intensities concentration in any frequency band as their energy spreads across the entire frequency spectrum. A slightly

4.3. SIGNAL ANALYSIS

higher intensity is observed in the lower 10% of the frequency range due to voice bars. Analysing the intensity deviation (right of Figure 4.8), the stops and fricative classes exhibits the highest variation in the lower 10% of the frequency range as there can be both voiced and unvoiced stops and fricatives. The existence and non-existence of the voiced bar will cause intensity variations in the lower frequency range. The nasals and vowels are all voiced, which explains the low intensity variation in this lower 10% of the frequency range. With these observations, a good sub-band division should be at the 10% and 60% of the frequency range, and are indicated by the black lines shown in Figure 4.8.

Next, with the sub-band division, we investigate the intensity changes that occur at phoneme boundaries given the manually labelled ground truth in the TIMIT database. If the sub-band division designed is ideal, a high intensity change should be observed in at least one of the sub-bands that enables good detection of the boundary. An analysis over the different types of phoneme transitions will also enable us to understand if certain types of transitions are harder to detect. Since we have 4 attributes here, there is a maximum of 16 different kinds of attribute transitions that can occur. For each phoneme boundary in the training corpus of the TIMIT database, we find the intensity change on the image spectrogram that occurs over the 3 sub-bands using the equation in 4.6, where c here corresponds to the ground truth boundary. With that, we obtain Table 4.3 tabulating all the intensity change over different types of attribute transitions across the 3 sub-bands, s_1 to s_3 . The percentage distribution of change at a sub-band i is obtained by taking the intensity change at sub-band i and dividing it by the total amount of change observed all 3 sub-bands. With this % we can conclude if the change in that particular sub-band is significant.

4.3. SIGNAL ANALYSIS

transitions	count	Mean Intensity Change			% dist. of change		
		s_1	s_2	s_3	s_1	s_2	s_3
stop-fricative	6608	37.50	53.49	62.09	0.24	0.35	0.41
stop-vowel	22856	24.72	28.82	23.21	0.32	0.38	0.30
stop-nasal	1724	42.33	37.15	20.51	0.42	0.37	0.21
stop-stop	19822	66.16	75.19	55.85	0.34	0.38	0.28
fricative-vowel	13244	32.59	29.60	35.61	0.33	0.30	0.36
fricative-stop	6243	37.78	59.63	68.78	0.23	0.36	0.41
fricative-fricative	831	24.44	28.50	34.60	0.28	0.33	0.40
fricative-nasal	560	35.09	36.35	47.00	0.30	0.31	0.40
vowel-vowel	26948	4.73	17.82	16.64	0.12	0.45	0.42
vowel-stop	20501	55.37	55.98	28.70	0.40	0.40	0.20
vowel-fricative	11745	37.10	28.77	48.74	0.32	0.25	0.43
vowel-nasal	10168	12.71	32.52	20.79	0.19	0.49	0.31
nasal-stop	4440	32.94	29.26	20.56	0.40	0.35	0.25
nasal-vowel	6315	10.14	31.21	22.72	0.16	0.49	0.35
nasal-fricative	1694	31.12	39.28	63.03	0.23	0.29	0.47
nasal-nasal	196	4.91	16.30	12.73	0.14	0.48	0.38

Table 4.3: Intensity changes across different attribute transitions and sub-bands.

From the table, it can be observed that intra-class changes experience minimal changes as compared to inter-class intensity changes, with the exception of stop-stop changes. This is because stops sometimes exhibit themselves by a lack of intensity in the spectrogram, and sometimes as a burst of energy across all frequencies in the spectrogram. Transitions to/from fricative classes to/from other classes also experience the highest amount of intensity changes in the last sub-band s_3 , which is consistent with our intuition as they are the only attribute that have energy concentrated in the higher frequencies. As stops have no specific concentration in any frequency bands, when tran-

4.3. SIGNAL ANALYSIS

siting from stop to another phoneme, the largest change will depend on the phoneme transited into. For example when transiting from stop to fricative, the largest change occurs in the last sub-band. Lastly, it can be observed that the changes experienced in vowel-nasal boundaries are less significant as compared to other transitions, as both vowels and nasals are characterised by strong formant signatures in the spectrogram. Visually, they are more similar than different, which makes them harder to differentiate as compared to other transitions.

From this analysis, it can be concluded that changes in intensities at phoneme boundaries become more prominent when we track for sub-band intensity changes, as illustrated in Figure 4.9. This makes peak selection easier to locate potential boundaries.

Figure 4.9: (Top) Enhanced spectrogram of the same TIMIT utterances. (Middle) STM obtained from MFCC spectrum. (Bottom) Blue: Intensity change profile with windowing without sub-band division and Green: with sub-band division.

4.4 Segmentation Algorithm

This section presents the algorithm proposed for phoneme segmentation, starting with the generation of an image spectrogram of the speech signal. With the intuition that significant intensity changes along the time-axis indicate a possible phoneme change, we detect the locations of such differences to find possible phonetic boundaries. As spurious detections can exist, post processing is done to find the best set of boundaries by using a minimum distortion reduction criterion.

4.4.1 Spectrogram Generation

The spectrogram is obtained by performing *STFT* [89] on the speech signal using a window size of 128 samples with a 64-sample overlap size. Since TIMIT sequences are used in this experiment, 128 samples for the sequences sampled at 16000 Hz correspond to an 8 ms window. To preserve the frequency details, 256 FFT bins is used. The logarithm is then applied on the *power spectral density (PSD)* matrix, where each column corresponds to the PSD for each window, from which we obtain the image representation of the speech.

Observing the spectrogram image generated (Figure 4.10, top), the contrast of the formants with the background is not distinct, making boundary detection challenging. Therefore, we apply histogram equalization [88] on the whole spectrogram to boost the contrast. Also seen from the same image, regions of silence does not appear smooth, but as grainy patches of grey regions, which can generate spurious detections resulting in high false alarms. Therefore, a median filter [87] with a window size of B is applied to each column of the image to remove such salt-and-pepper noise, with the effect of B discussed in Section 4.5. The filtering is not done row-wise to avoid blurring the phoneme boundaries which we wish to detect. Figure 4.10 shows a spectrogram before and after enhancement. After enhancement, regions of silence appear as white smooth regions, and the formants appear stronger.

4.4.2 Boundary Detection

To detect for possible phoneme boundaries, the algorithm locates significant intensity changes along the time-axis in three sub-bands discussed in Section 4.3.2. The first sub-band, s_1 , is set as 1 to $(0.1 \times h)$, where h refers to the image height. Voice bars [90] are usually found in this sub-band and can be seen as the high values observed in the corresponding range on the energy profiles. These voice bars help to differentiate between silence and speech portions. The second sub-band, s_2 , is set as $(0.1 \times h + 1)$ to $(0.6 \times h)$, and is where formants of vowels can be detected. As evident from the energy profiles, vowel have energies concentrated in the lower frequency spectrum and are significantly higher than that observed in the other classes. The last sub-band, s_3 , refers to $(0.6 \times h + 1)$ to h , and is useful for the detection of fricatives. Although fricative energy can spread over the whole frequency spectrum as they exhibit noise-like characteristics, they frequently have energies concentrated in the high frequencies. Due to the lack of high frequency components for other classes like vowels and nasals, the last sub-band becomes crucial to separate fricatives from the other classes. An example of the sub-band divisions on the spectrogram is shown in Figure 4.10, bottom image.

Figure 4.10: (Top) Spectrogram before enhancement. (Bottom) Spectrogram after enhancement and split into the three sub-bands.

To avoid many spurious detections, a 5 columns window is used for frame differencing.

4.4. SEGMENTATION ALGORITHM

$\mathbf{d}(s, c)$, which is the difference of each sub-band s at column c , is computed as such,

$$\mathbf{d}(s, c) = \frac{\sum_{r=r_1}^{r_2} \sum_{w=1}^5 |\mathbf{I}(r, c - 5 + w) - \mathbf{I}(r, c + w)|}{(r_2 - r_1 + 1) \times 5} \quad (4.6)$$

where $\mathbf{I}$ refers to the image spectrogram, and r_1 to r_2 defines the rows in sub-band s of interest.

We subsequently compute the image difference profile, $\mathbf{D}$, in Eq. (4.7).

$$\mathbf{D}(c) = \max(\mathbf{d}(s_1, c), \mathbf{d}(s_2, c), \mathbf{d}(s_3, c)) \quad (4.7)$$

Only the sharpest intensity change across the sub-bands is recorded in $\mathbf{D}$. The peaks in $\mathbf{D}$ with values above $f \times \text{mean}(\mathbf{D})$ are chosen as boundary locations. Effect of f is discussed in Section 4.5.

4.4.3 Boundary Selection

Next, we remove spurious boundaries detected above due to noise by using a minimum distortion reduction criterion. Given a segment on the image, the distortion E , of this segment is given by Eq. (4.8),

$$E = \frac{\sum_{r=1}^m \sum_{c=2}^n |\mathbf{I}(r, c) - \mathbf{I}(r, c - 1)|}{m \times (n - 1)} \quad (4.8)$$

where m and n refer to the number of rows and columns in spectrogram $\mathbf{I}$, and E indicating the intensity change accumulated in the segment. Intuitively, if the segment does contain a boundary, splitting the image at this boundary would reduce the accumulated distortion from the two resultant segments since the biggest contributor to the distortion measure, which is the boundary, is removed and no longer contribute to E .

With this intuition in mind, the boundary selection process starts with the whole spectrogram being one segment. The best boundary from the set of detected boundaries is chosen and the image is split into two segments. The algorithm is repeated on the

4.5. EXPERIMENTAL RESULTS AND DISCUSSIONS

resultant segments and the process repeated until no more boundaries can be chosen from the current set of segments.

There are three terminating conditions where no more boundaries can be selected:

- no more detected boundaries on the segment, or
- the resultant distortion reduction is less than the specified threshold of 0.0025 in our algorithm, or
- the selection of the boundary would result in a segment less than 5 column in width. This threshold is set to coincide with the window length used in Section 4.4.2.

Once the boundary selection process terminates, the selected set of boundaries gives the final segmentation of the spectrogram into its phoneme segments.

4.5 Experimental Results and Discussions

4.5.1 Image-based Phoneme Segmentation

The proposed algorithm, termed *ISeg*, was used to segment the 1344 utterances in the TIMIT testing set. The results were compared against the results of [84] using the evaluation metrics defined in Section 4.2.

Figure 4.11 gives the segmentation algorithm's performance. To obtain the curves, the peak selection threshold f described in Section 4.4.2 is first arbitrarily fixed as 0.5 and the median filter window size B varied. B affects the level of details preserved and therefore the number of boundaries detected. The best window size was found to be 8. We then fix B and vary f .

4.5. EXPERIMENTAL RESULTS AND DISCUSSIONS

Figure 4.11: (Top) OS plotted against CDR for *ISeg*. (Bottom) FA plotted against CDR for *ISeg*.

The results obtained were similar to those found in [84] and [85], with slight degradation, and were tabulated in Table 4.4. Comparison of the results can only be done using the CDRs obtained at 0% OS, as only this performance figure was provided in both [84] and [85].

Method	CDR
MMC [84]	76.0%
RD-IDA [85]	76.7%
Proposed Method - <i>ISeg</i>	75.0%

Table 4.4: Comparison of *ISeg* with other unsupervised segmentation algorithms at OS = 0%.

4.5. EXPERIMENTAL RESULTS AND DISCUSSIONS

CDR at 78.07%, OS at 7.74% and FA at 27.54% were obtained at the best operating point found at $B = 8$ and $f = 0.625$. This result would be used for subsequent comparisons. Although the proposed method did not outperform existing unsupervised methods, results are still encouraging as it showed the potential for image features on spectrograms to be used for speech processing tasks. With simple image processing techniques applied to the spectrogram, the algorithm attained a similar performance to other unsupervised segmentation methods.

4.5.2 Fusion with HMM-based Segmentation

Next, fusion of the results of *ISeg* with those obtained by HMM phoneme recognition, *HMM-rec*, and those obtained by HMM forced alignment, *HMM-fa*, is performed to investigate if performance improvement can be achieved. An accuracy gain would imply that visual cues could provide an additional feature to improve the performance of speech processing tasks. To obtain HMM-based segmentation, the HTK toolkit and the 4120 utterances in the TIMIT training set were used to train context-independent HMMs for the 61 phonemes in the TIMIT corpus as context-independent HMM was believed to give better segmentation results [77]. Each HMM contained 5 states with left-to-right transitions, and each state contained 16 Gaussian mixture components. The HMMs were then used to obtain HMM segmentations for *HMM-rec* and *HMM-fa*.

Subsequently, we performed a fusion of the results by first comparing all the boundaries found in *ISeg* with those obtained by *HMM-rec* or *HMM-fa*. Identical boundaries were ignored and the remaining boundaries discovered by *ISeg* fused into the boundary set discovered by HMMs. The results were indicated by (*raw*) in Table 4.5. This simplistic fusion resulted in many false alarms as detections that were one column apart were also deemed as different boundaries.

In addition, we did not double count true detections in our evaluation. If a boundary is less than 20ms from a true boundary, a match for the true boundary had been fulfilled and accounted for in the true detection count. If another detection within 20ms from the

4.5. EXPERIMENTAL RESULTS AND DISCUSSIONS

same true boundary was found, the second detection will be regarded as a false alarm. This caused the extremely high false alarm counts for the (*raw*) results. The purpose of the results however is to highlight the improvement in CDR that *ISeg* can achieve through fusion.

Method	CDR(%)	OS(%)	FA(%)
ISeg	78.07	7.74	27.54
HMM-rec	84.57	7.73	21.50
HMM-fa	86.60	0.00	13.40
ISeg-HMM-rec (raw)	94.90	104.24	53.53
pruned	89.75	33.98	33.01
strict (f=0.75)+pruned	88.98	27.17	30.03
ISeg-HMM-fa (raw)	96.28	97.04	51.14
pruned	91.85	30.25	29.48
strict (f=0.75)+pruned	91.29	23.06	25.82

Table 4.5: Comparison of results obtained from fusing *ISeg* with HMM baselines, against *HMM-rec* and *HMM-fa*.

As observed from Table 4.5 above, fusing the results of *HMM-rec* with *ISeg*, can achieve a CDR of 94.90%, which was about 10% higher than that obtained by *HMM-rec* alone. Similarly, when the result of *ISeg* was fused with those from *HMM-fa*, close to 10% improvement of CDR was also observed. This confirmed the hypothesis that visual cues on the spectrogram can provide an additional feature representation to existing ones for speech processing.

To understand how the proposed method complements the HMM approaches, we compare the detections obtained by *ISeg* with that obtained by HMM in Figure 4.12. It could be observed that the boundaries obtained by *ISeg* in Figure 4.12(c) are closer to the true phoneme boundaries shown in Figure 4.12(b). The ones obtained from *HMM-rec* in Figure 4.12(d) are very often a few columns off from the true boundaries. This is further confirmed by Table 4.6, where we tabulate the CDR obtained at different time window. Previously, the definition of a correct detection is when the detection is 20ms away from the manually labelled ground truth. When the time window is changed to 0,

4.5. EXPERIMENTAL RESULTS AND DISCUSSIONS

Time Window (ms)	HMM-rec CDR (%)	ISeg CDR (%)
0	12.9	18.5
4	33.4	45.7
12	69.2	70.6
20	86.6	78.1
32	94.3	81.6
40	96.1	82.8

Table 4.6: CDR evaluated at different time window for *HMM-rec* and *ISeg*.

4 and 12 ms, it can be observed that *ISeg* has higher accuracies than *HMM-rec*, which shows that detections from *ISeg* are closer to the truth boundaries. However, it can be observed that the performance of *ISeg* saturates at higher window sizes, suggesting that some boundaries simply cannot be detected. *HMM-rec* on the other hand managed to detect close to 96% of the boundaries, but some boundaries are more than 40ms away. Visual cues can therefore give a more accurate segmentation of phoneme boundaries which contributed to the improvement of the CDR in the fused results. However, there is a limit on the performance of *ISeg*, which will be elaborated later. Note also in Figure 4.12(a) that there was a distinct noise spike seen as a black vertical line, and highlighted in a red box. *ISeg* did not generate a false detection due to median filtering of the spectrogram. In other words, *ISeg* was able to overcome transient impulse noise spikes in the speech with suitable filtering. However, *HMM-rec* produced a false detection in the same segment as seen in Figure 4.12(d).

ISeg, of course, has its drawbacks. Closer observation of Figure 4.12 revealed a series of miss detections in the middle portion of Figure 4.12(c). When compared to the ground truth in Figure 4.12(a), the miss detection corresponds to a sequence of vowels and approximants which exhibit themselves very similarly in spectrograms as discussed in [27] and Section 2.2.2. Such similarities make it is very challenging to find vowel-vowel or approximant-vowel boundaries as no distinct transitions on the image spectrogram could be observed as concluded in Table 4.3. *ISeg* would be more effective in segmenting syllabic languages with clear consonant-vowel-consonant structures.

4.5. EXPERIMENTAL RESULTS AND DISCUSSIONS

(a) Spectrogram from Wavesurfer

(b) Ground truth segments labeled in black

(c) *ISeg* detected labels in blue. $f=0.625$

(d) *HMM-rec* detected labels in blue

(e) Fused results of *ISeg* and *HMM-rec* with false alarm pruning

Figure 4.12: Comparison of segment boundaries obtained from *ISeg* and *HMM-rec*

In order to overcome the high false alarm rate due to the simplistic fusion, we removed some of the unrealistic false alarms by running a sliding window with a size of 5 columns

4.6. CONCLUSION AND FUTURE WORK

through all the detected boundaries. If the window contained more than one boundary, we selected the one with the highest intensity change at the boundary. If the intensity change at this selected boundary was less than a threshold τ , we rejected this selected boundary as it should be a false alarm. τ is set to 3 here. This eradicates segment boundaries that are less than 5 columns from each other and the result of such a false alarm pruning is illustrated in Figure 4.12(e).

The performance of such a pruning strategy on the evaluation metrics are shown in rows indicated by *pruned* in Table 4.5. Although the post-processing helped to remove some of the false alarms, a significant amount remained. Perhaps a better enhancement step should be performed on the spectrogram so that the extracted features are more robust to spectral changes and spurious detections reduced. A better fusion method is also required to exploit the advantages of both *ISeg* and HMM without generating too many false alarms.

A stricter version of *ISeg* can also be used if slight loss in CDR can be tolerated. In rows indicated by *strict* ($f=0.75$), *pruned* in Table 4.5, the peak selection threshold in *ISeg* was set to $f=0.75$, higher than the optimal $f=0.625$ as discussed previously. The results of this configuration was fused with HMM segmentation and false alarm pruning was also performed. A slight degradation in CDR was observed, but both OS and FA were lowered with this setup. The CDR now was still above that achieved by *ISeg* or HMM alone.

4.6 Conclusion and Future Work

In summary, we have presented an image processing approach to speech phoneme segmentation with comparable results to other unsupervised phoneme segmentation methods, showing that image-based feature can be a possible alternative to phoneme segmentation. We also fused the results of the proposed segmentation algorithm with HMM results and observed an improvement to the segmentation accuracy. This showed that image features does contain additional information that can be used to improve the performance

4.6. CONCLUSION AND FUTURE WORK

of speech processing tasks.

In the future, we wish to explore better image features that are more invariant to spectral variations to reduce the false alarms. Different kind of spectrogram representations can also be investigated. Fusing the results of the proposed algorithm with other language-independent detectors, such as the manner and place attribute detector, is also a possible direction.

4.6. CONCLUSION AND FUTURE WORK

Chapter 5

Detection of Unit Selection Based Synthesized Speech for Anti-spoofing

5.1 Introduction

This chapter investigates the use of an image approach to detect for synthesized speech generated by unit selection(US)-based synthesis technique [91], with the work published in [21]. US-based synthesis technique works by selecting and concatenating speech units from a speech database to form a synthetic sequence. In order to produce natural sounding speech, the database needs to be sufficiently large to contain a good selection of speech units to choose from.

The need to detect for synthesized speech arises from the fact that they can be used in spoofing attacks, which tries to imitate a target speaker in order to fool speaker verification (SV) systems, compromising security [92, 93, 94, 95, 96]. Although different spoofing techniques are available, such as voice mimicry, playback, voice conversion or speech synthesis [96], recent advances of voice conversion and speech synthesis technologies and prevalence of open-source tool-kits to facilitate voice spoofing [93], are posing

5.2. RELATED WORKS

serious threats to automatic SV systems. In fact, state-of-the-art HMM-based speech synthesizers now require only a few minutes of a speaker’s data to perform model adaptation [97], making spoofing techniques easily available. The false acceptance rate, referring to the rate synthetic speech is falsely accepted as natural speech, could be as high as 85.5% for synthetic speech obtained from Hidden Markov Model (HMM)-based speech synthesizer trained on Wall Street Journal corpus [92], and as high as 98.08% [93] for a corpus synthesized using the *MARY Text-to-Speech Synthesis (MaryTTS)* system [1] based on US. The threats posed by synthetic speech and the high false acceptance rate of 98.08% for US-based synthetic makes it crucial to have effective measures to detect for them.

In Section 5.2, we introduce the past attempts of detecting synthesized speech in general, and highlight the poor performance attained for US-based synthetic speech detection by existing techniques due to the method used to generate them. In Section 5.3, we highlight how US-based synthetic speech shows artifacts on the image spectrogram which can be exploited to detect for them. Section 5.4 and 5.5 will discuss the algorithm that is training-free and experiments before we conclude this chapter in Section 5.6.

5.2 Related Works

The threat to SV systems caused by synthetic speech has motivated research to detect spoofing. Various attempts are made, such as the use of inter-frame difference as a discriminator for human and synthetic speech [98] with synthetic speech observing lower scores than human speech. In [99], the authors used the lower variance observed in higher order Mel-cepstral coefficients of synthetic speech to detect for them. However, these two methods assume that the HMM-based speech synthesizer used in the spoofing attempt is known [96, 100], which is an unrealistic assumption for real anti-spoofing scenarios. The more common feature used to discriminate between natural and synthetic speech is the phase information [92, 101, 102, 103, 104, 105, 106, 107, 108]. Since human auditory is insensitive to phase information [109], vocoders do not reconstruct them accurately,

5.2. RELATED WORKS

resulting in the lack of speech-like phase information in synthetic speech. However, the drawback of most of these studies is the dependence of the detection algorithm on known vocoder seen in training [92, 103, 104, 110]. When the testing speech used a different vocoder, the detector accuracy decreases sharply from 90% to as low as 6.3% in [92], which limits the usefulness in unforeseen attacks. Therefore, vocoder-independent algorithms would be a possible future research direction. Other features used to detect synthetic speech include the use of prosodic pitch features [111, 112, 113] and local binary patterns [102, 114, 115]. The ASVspoof 2015 challenge [116] also saw authors using wavelets [107], phoneme posterior probability [113] and cochlear filters [117].

Much of the work discussed so far were tested on HMM-based synthesized speech that can be easily adapted to a target speaker [92, 98, 99, 101, 114, 115], with most of them [92, 98, 99, 101] requiring knowledge of the speech synthesizers or the vocoders used. Although US-based speech synthesis systems is less appealing for attackers due to its high database requirement [96], the threat it poses to SV systems due to the high quality of speech produced [118] should not be overlooked, as highlighted by the high false acceptance rate of 98.08% reported in [93].

In the ASVspoof 2015 challenge, US-based synthesized speech was included as a testing condition and results [102, 103, 104, 110, 105, 113, 106, 117, 107] highlighted the difficulty in detecting US-based synthesized speech using existing features which are predominantly phase-based. The equal error rate (EER) obtained by most systems for voice conversion and HMM-based systems are all below 1%, but that for US-based speech ranges from 20% to 40%. The poor performance can be due to the lack of training data to train models for US-based synthesized speech, as well as the lack of phase artifacts in US-based synthesized speech since vocoders are not used. Even though local binary patterns (LBP) [114, 115] were said to be less reliant on prior knowledge, and proposed to provide a more universal countermeasure for spoofing attacks, it was not effective for US-based synthesized speech as well [102]. One work in the challenge [117] did manage to obtain an EER of 8.49% by using cochlear filters and instantaneous frequency. Despite so, it is

5.3. OBSERVATIONS

still much higher than that obtained for the other testing conditions. In a recent work exploring various high-dimensions features for spoofing detections [119], 0% EER can be obtained for all category of spoofing attack, except for US-based synthesized speech, which gave an EER of 27.79%. This highlights the challenge in US-based synthesized speech detection and the need for better techniques to deal with them, which concurred with the authors' opinions in [102, 105, 119].

In this work, we exploit the artifacts observed on the spectrogram image caused by the concatenation process unique to US-based synthesizer. The proposed detection algorithm used is training-free to overcome the reliance of the detector's performance on the availability of target speaker's training data, which can be a considerable amount for US-based speech synthesis [96] so as to generate natural sounding synthetic speech. This makes the training data preparation for such synthesizers considerably taxing.

5.3 Observations

As US-based speech synthesis works by concatenation, it does not involve vocoders and exhibits no phase artifacts. However, concatenation artifacts can be seen on the spectrogram as concatenation causes spectral discontinuities that show up as vertical lines in the spectrogram. This is illustrated in Figure 5.1 and we refer to this line artifact as *concatenation line* and use it to perform spoofing detection.

5.4 Algorithm

The algorithm begins by detecting for vertical lines in the spectrogram and subsequently uses a verification stage to remove false detections that do not corresponds to concatenation artifacts.

5.4. ALGORITHM

Figure 5.1: Spectrogram generated with 4 ms Hann window, and frequency range of 8000 Hz. (Top) Human speech with no line artifacts. (Bottom) Speech synthesized from *MaryTTS* [1], a US-based speech synthesis system, with concatenation lines marked in red boxes.

5.4.1 Spectrogram Generation

STFT is performed on the speech signal using a 4 ms Hann window and 50% overlap to generate a wide band spectrogram. As concatenation lines caused by spectral discontinuities are instantaneous, they show up better in wide band spectrograms than narrow band spectrogram which uses a typical 20 ms [26] window size. In addition, Hann window [120] is used instead of the usual Hamming window, as it is zero at both ends and thus eliminates spectral discontinuities during spectrogram generation that can result in artifacts that can be confused with the concatenation lines. The number of FFT bins used is 512.

Next, we enhance the spectrogram by performing histogram equalization [88] to bring the intensity values to the range of 0 to 255.

5.4.2 Lines Detection

Template matching [121] is used to perform concatenation line detection, a technique commonly used to detect for specific structures. The template illustrated in Figure 5.2 is used, which has a height equivalent to the spectrogram image height since the observed artifact as a result of spectral discontinuity during concatenation spans across any frequency range and exhibits itself across the height of the spectrogram. The width

5.4. ALGORITHM

of the template is decided empirically at 15 pixels. Narrow templates produce higher detection rates, but will result in high false alarm rates. Having a wider template will take into account a wider neighbourhood and favour more well-defined line structures, reducing false detections. However, it can be too strict and result in low detection rates.

Row 1	-1	-1	-1	-1	-1	-1	-1	14	-1	-1	-1	-1	-1	-1	-1
	-1	-1	-1	-1	-1	-1	-1	14	-1	-1	-1	-1	-1	-1	-1
Row N															
	-1	-1	-1	-1	-1	-1	-1	14	-1	-1	-1	-1	-1	-1	-1
	-1	-1	-1	-1	-1	-1	-1	14	-1	-1	-1	-1	-1	-1	-1

Figure 5.2: Template used to detect for concatenation lines, with a width of 15 and height N equivalent to the spectrogram height. The central value is set to the width of the template minus one.

The template is convolved horizontally across the spectrogram $\mathbf{I}$ to obtain the detection profile $\mathbf{D}$ as described by Equation 5.1.

$$\mathbf{D}(c) = \frac{\sum_{r=1}^N \sum_{i=-7}^7 \mathbf{I}(r, c+i) \times \mathbf{T}(r, 8+i)}{N \times 15} \quad (5.1)$$

where c and r refer to the column and row of $\mathbf{I}$ respectively. N is the height of $\mathbf{I}$ and $\mathbf{T}$ is the template.

After obtaining $\mathbf{D}$, we detect for peaks in $\mathbf{D}$ and only accept those with value more than the mean of $\mathbf{D}$. The peak positions correspond to location on the spectrogram with line-like structures. Note however they may not correspond to concatenation lines as natural speech also exhibit vertical lines in the spectrogram. Further verification is needed.

5.4.3 Detections Verification

Observing the spectrograms in Figure 5.1, concatenation lines are usually isolated from its surroundings with lower intensities in its neighbourhood, while non-concatenation lines are usually accompanied by higher intensities in its surrounding due to speech

5.4. ALGORITHM

signals. In this stage, we verify the surrounding intensities at the detected positions to ensure that it conforms to the target profile described. The intensity profile at each detected position is constructed using Equation 5.2.

$$\mathbf{P}_{det}(c) = \sum_{r=1}^N \mathbf{I}(r, pos + c) \quad (5.2)$$

where *pos* refers to a previously detected position, and *c* refers to the column that ranges from -5 to 5. $\mathbf{P}$ therefore forms a vector of 11 values which is matched with the target profile in Figure 5.3. The result of the matching is computed in Equation 5.3.

-1	-1	-1	-1	0	1	0	-1	-1	-1	-1
----	----	----	----	---	---	---	----	----	----	----

Figure 5.3: Target intensity profile P_{tar} .

$$mVal = \sum_{c=1}^{11} P_{det}(c) \times P_{tar}(c). \quad (5.3)$$

The strict target profile strongly penalizes values that occurs in the surroundings, with strong intensity in the neighbourhood causing *mVal* to be negative. The target profile size is decided empirically, with a wider profile considering a wider neighbourhood, enforcing more ‘isolatedness’ of the detected line. The two positions adjacent to the central position are set to zero so that concatenation line that are more than one column wide do not affect the computation. Detection lines that fail to exceed threshold *t* will be discarded and utterances with any detected lines will be classified as a spoof utterance.

5.4.4 Finding Best Region for Verification

Strong intensities in the lower frequencies of the spectrogram can pollute the intensity profile. It is therefore non-ideal to use the entire height of the spectrogram to construct $\mathbf{P}_{det}$ and renders the need to select a least polluted section to be used for verification instead. To find the optimal region, we inspect the entire height about the detected

position, together with the 5 columns left and right of it. From this section, we wish to obtain a 40 rows slice that has the least intensity deviation, with the central column excluded in the deviation calculation as it contains the concatenation line with strong intensities. The choice of 40 rows is decided empirically. This 11 columns by 40 rows slice corresponds to the section with the least amount of noise and will hence be used for verification. The values are finally normalized to $[0, 1]$ and used in place of the entire height of the spectrogram in Equation 5.2.

5.5 Experiments and Discussions

3 sets of experiments are performed to evaluate the effectiveness the image-based algorithm for US-based synthetic speech detection. Firstly, we ran our algorithm on 16000 Hz data on two development sets, so as to find the optimal threshold t and check the sensitivity of the algorithm to the size of the development sets. The same experiment is repeated on 8000 Hz speech utterances as such speech signals are also encountered in real spoofing scenarios. Therefore there is a need to verify the impact on the performance due to the truncated higher frequencies on the 8000 Hz spectrogram, where concatenation lines are more prominent. Lastly, we apply the algorithm on a larger 16 000 and 8000 Hz test sets to verify the robustness of threshold t .

5.5.1 Corpus

The corpus from the ASVspoof 2015 Challenge [116] is used in the experiments, with test portion from *Genuine* subset and *S10* being used as our human speech and US-based synthesized speech data sets respectively. Utterances in *S10* are synthesized using *MaryTTS* [1] and both sets are sampled at 16000 Hz. The development sets and test set are summarized in Table 5.1. The 400 sequences are randomly chosen from each database, and the 9000 sequences in the test set is taken from the rest of the subsets. In addition, 100 sequences from the 400 utterances are randomly chosen to create another

5.5. EXPERIMENTS AND DISCUSSIONS

development set. The same data sets are downsampled to 8000 Hz.

Subset	Dev 100	Dev 400	Test
Genuine	100	400	9000
S10 - US-based Synthesis	100	400	9000

Table 5.1: Database used in experiments.

5.5.2 Experimental Results

For 16000 Hz and 8000 Hz Development Sets

The two development sets are used to find the optimal t for $mVal$ for both the 16000 Hz and 8000 Hz experiments. Table 5.3 and 5.4, together with Figure 5.4 and 5.5, show the results of running the algorithm at different t . The metrics used for evaluation is summarized in Table 5.2 below.

Metric	Definition
<i>False Acceptance Rate (FAR)</i>	Rate at which synthetic speech is not detected and falsely accepted as natural speech
<i>False Rejection Rate (FRR)</i>	Rate at which natural speech is falsely rejected and being detected as synthetic speech
<i>EER</i>	Rate at which FAR equals FRR .

Table 5.2: Definition of various metrics used for evaluation.

5.5. EXPERIMENTS AND DISCUSSIONS

	Dev 100		Dev 400	
t	FAR(%)	FRR(%)	FAR(%)	FRR (%)
none	0.00	100.00	0.00	100.00
-1	7.00	64.00	7.75	57.25
-0.8	11.00	28.00	12.75	26.25
-0.7	13.00	15.00	15.25	15.00
-0.6	15.00	8.00	17.25	7.00
-0.4	22.00	1.00	25.50	0.50
-0.2	35.00	0.00	33.75	0.00
0	42.00	0.00	41.25	0.00
EER(%)	13.44		15.20	
Optimal t	-0.6778		-0.7018	

Table 5.3: Algorithm performance for 16000 Hz Development Sets - Dev 100 and Dev 400.

Figure 5.4: FAR and FRR for Dev 100 and Dev 400 sampled at 16000 Hz, with EER equals to 13.44% and 15.20% respectively.

5.5. EXPERIMENTS AND DISCUSSIONS

	Dev 100		Dev 400	
t	FAR(%)	FRR(%)	FAR(%)	FRR (%)
none	0.00	100.00	0.00	100.00
-1	28.00	45.00	30.50	40.50
-0.9	39.00	36.00	38.00	33.00
-0.8	48.00	25.00	46.25	22.50
-0.6	57.00	13.00	59.00	10.00
-0.4	76.00	5.00	76.25	3.25
-0.2	83.00	1.00	86.25	0.75
0	92.00	0.00	93.75	0.00
EER(%)	37.35		35.50	
Optimal t	-0.9150		-0.9333	

Table 5.4: Algorithm performance on 8000 Hz Development Sets - Dev 100 and Dev 400.

Figure 5.5: FAR and FRR for Dev 100 and Dev 400 sampled at 8000 Hz, with EER of 37.35% and 35.50% respectively.

5.5. EXPERIMENTS AND DISCUSSIONS

For Test Set

The t obtained from Dev 400 is applied on the test set with results summarized in Table 5.5. Threshold from Dev 400 is chosen since it contains more data and should produce more reliable parameters. Results obtained are close to that obtained on the development sets given in Table 5.3 and 5.4.

Sampling Rate (Hz)	t used	FAR (%)	FRR (%)
16000	-0.7	14.12	16.06
8000	-0.93	38.90	35.37

Table 5.5: Results on test set.

5.5.3 Discussions

Results of the experiments showed that concatenation lines observed on spectrograms of speech utterances synthesized by US can be used to detect spoofing attacks. The performance is stable between the two development sets, with 15.20% and 35.50% EER obtained for 16000 Hz and 8000 Hz signals respectively. The 20% degradation in performance for 8000 Hz highlighted the importance of higher frequencies in the spectrogram to detect concatenation lines. The EER of 15.20% is lower than most of that from other systems in the ASVspoof 2015 challenge, although still unable to beat the 8.49% obtained in [117]. The results are therefore encouraging, showing that concatenative lines detection is an effective feature for US-based synthesized speech detection. Moreover, the proposed algorithm is training-free and relied on simple thresholding. A follow-up experiment would be to apply this algorithm and threshold to speech synthesized by other US-based synthesis systems to verify its robustness.

We believe that the output from the proposed algorithm can be fused with existing anti-spoofing algorithms to further improve the detection performance for speech synthesized by US. A feature representation so as to enable fusion with existing speech features and machine learning techniques used for spoofing detection will also be useful. We hope

that the fusion will achieve comparable performance to that obtained for other spoofing attacks.

5.6 Conclusion and Future Work

Natural sounding US-based synthesized speech poses a serious threat to SV systems. Although detection algorithms for synthesized speech exists, current detection algorithms rely predominantly on phase-based features to detect for synthesized speech. As US-based speech synthesizers do not use vocoders and exhibit phase anomalies, they can bypass existing detection algorithms, as shown by the ASVspoof 2015 challenge [116] results. This highlights the challenge in detecting for US-based synthesized speech.

In this work, we exploit the fact that US-based speech synthesizer uses concatenation and hypothesized that artifacts on the spectrogram caused by the concatenation can be used to perform spoofing detection. A training-free algorithm that makes use of line detection is proposed to detect the artifacts. Experiments tested on both 16000 and 8000 Hz utterances have produced encouraging results and showed the effectiveness of the algorithm, obtaining an EER of 15.20% and 35.50% respectively, which is lower than most reported in the ASVspoof 2015 challenge .

In future, we wish to evaluate the robustness of the algorithm and threshold on speech produced by other US-based speech synthesis or voice conversion algorithms. Fusion of the algorithm with existing anti-spoofing algorithms should also be performed to improve the performance for US-based synthetic speech detection. Other features such as F0 discontinuities may also be useful for US-based synthesizers and should also be explored.

5.6. CONCLUSION AND FUTURE WORK

Chapter 6

Conclusion and Way Ahead

The objective of the research is to investigate into visual representation and image processing techniques for speech processing. It is believed that visual cues carries important information for speech processing that can be harnessed to complement existing speech processing. The image representation also incorporates contextual information more easily as compared to current speech features which are predominantly frame-based. The thesis discussed the mapping of speech problems into an image problem, with the use of the image spectrogram as the visual representation for speech. A literature survey was provided in Chapter 3, summarizing past works on spectrogram readings, as well as recent attempts to use image processing for speech and acoustic processing. However, work done in this area remained to be a feasibility test on limited tasks and this work seeks to extend the use of image-based approach to two speech problems: *Unsupervised Speech Segmentation*, *Detection of Unit Selection Based Synthesized Speech for Anti-spoofing*.

We conducted experiments to perform unsupervised speech segmentation, and the detection of synthesized speech for anti-spoofing using purely image-based solutions. For phoneme segmentation, the quality of the boundaries obtained were closer to the true boundary labelled by humans, as compared to that provided by HMM recognition, showing that image features can value-add to speech task and visual features do carry complementary information to aid existing speech tasks. For the detection of synthesized

speech, the use of visual features and techniques have been shown to produce optimistic results, surpassing many of the results obtained using traditional speech features, except for one work.

Results from the experiment are encouraging but more work remains to be done. Currently, intensity values and differences are used as features to act as visual cues. More features on the spectrogram should be explored, such as lines that represent formant tracks and blobs for fricative area. Unlike conventional features such as MFCC, these representations are segment based, and can take into account more contextual information. However challenges remains on how to obtain such features accurately, especially in the presence of noise found in the spectrogram.

In addition, the work so far is only based on the time-frequency spectrogram; there are other forms of visual representations that can also be considered. The waveform, although hard to interpret, might carry important information of speech onset and offset that can be harnessed. Another interesting representation will be the cochleagram time-frequency representation of speech [122], which attempts to mimic the outer and middle hear functions. Visual cues on the cochleagram might be able to give additional cues that can be harnessed for existing speech tasks. Wavelets [107], that are popular representations for both image and speech signals can also be exploited. In addition, like in MFCC, which considers change, and change of change, a similar concept can be adopted here for visual representations to capture contextual changes. Lastly, many of the visual representations discussed rely on the window to generate the different representation. The window affects the resolution, and thus the level of details preserved. Perhaps, multiple window sizes can be used to generate multiple representations, and information fused, like multi-resolution algorithms in image processing.

In addition to exploring novel representations, we can also explore other novel applications using these visual features. E.g., with the phoneme segmentation algorithm proposed in the algorithm, one possible extension is to use this algorithm to perform unsupervised attribute recognition for improving word recognition performance. Since

properties like vowels and fricatives can be easily recognized on spectrogram, the image spectrogram can be used to perform such recognition to reduce ambiguity in word recognition to improve performance.

From attribute recognition, we can also extend the use of visual cues to perform low resource keyword spotting. In this task, we wish to spot the keyword in a database of speech sequences given a visual representation of the spoken keyword. This is similar to the object recognition problem in image processing as previously mentioned in Chapter 2.1. One solution is to build all possible templates for the object to be recognized. However, different perspectives of the object and objects containing moving parts, example includes humans, animals, etc. poses serious challenges, and imply that a large repository of templates will be required to cover all the variations. Therefore, there has been work to match objects by first matching the different parts, and then making sure that the relationship between the different parts conforms to a structure to conclude a match. This technique is commonly referred to as “Matching by parts” [22]. This idea can be adopted to perform keyword spotting. Each formant track or fricative blob can be thought of as a part, with different structures and arrangement of these parts forming different words. Due to different individuals speaking differently, and pronouncing words differently, these structures may vary, but still fall within certain structures enforced by the phonemes composition of the words.

Lastly, another application of visual cues can be in the area of source classification, such as music and speech segmentation. This is because music exhibits rigid structures, unlike speech. Such texture cues can be used to perform segmentation. If music and speech segmentation is achievable, more challenging tasks such as noise and speech segmentation, and even noise classification can be attempted. Noise classification results can then be extrapolated to do channel estimation should certain channel exhibits consistent noise characteristics.

Publications

1. S. J. Leow, E. S. Chng and C. H. Lee, “Language-resource Independent Speech Segmentation Using Cues From a Spectrogram Image”, *ICASSP 2015*, Accepted.
2. S. J. Leow, E. S. Chng and C. H. Lee, “Zero Resource Anti-spoofing Detection for Unit Selection Based Synthetic Speech Using Spectrogram Artifacts”, *APSIPA 2016*, Accepted.

References

- [1] M. Schröder and J. Trouvain, “The german text-to-speech synthesis system mary: A tool for research, development and teaching,” *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.
- [2] W. Koenig, H. K. Dunn, and L. Y. Lacy, “The sound spectrograph,” *The Journal of the Acoustical Society of America*, vol. 18, pp. 244, 1946.
- [3] V. W. Zue, “Speech spectrogram reading: An acoustic study of english words and sentences : Special summer course, lecture notes and spectrograms,” *Massachusetts Institute of Technology*, July 14-18 1986.
- [4] M. Mills, “Deaf jam: From inscription to reproduction to information,” *Social Text*, vol. 28, no. 1 102, pp. 35–58, 2010.
- [5] A. A. S. Farani, E. H. S. Chilton, and R. Shirtley, “Dynamical spectrogram, an aid for the deaf,” *The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference Australia*, 1998.
- [6] B. H. Juang and L. R. Rabiner, “Automatic speech recognition - a brief history of the technology development,” *Elsevier Encyclopedia of Language and Linguistics*, 2005.
- [7] S.B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, , no. 4, pp. 357–366, 28 1980.
- [8] H. Hermansky, “Perceptual linear predictive (plp) analysis for speech,” *J. Acoust. Soc. Amer.*, pp. 1738–1752, 1990.
- [9] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” *J. Acoust. Soc. Am.*, pp. 637–655, 1971.
- [10] A. G. Adami, “Automatic speech recognition: from the beginning to portuguese language,” *PROPOR*, 2010.

REFERENCES

- [11] K. Sarmah and U. Bhattacharjee, “Gmm based language identification using mfcc and sdc features,” *International Journal of Computer Application*, vol. 85, no. 5, 2014.
- [12] K. S. R. Murty and B. Yegnanarayana, “Combining evidence from residual phase and mfcc features for speaker recognition,” *IEEE Signal Processing Letters*, vol. 13, no. 1, 2006.
- [13] L. Rabiner and B. H. Juang, “An introduction to hidden markov models,” *ASSP-Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, Jan 1986.
- [14] M. N. Stuttle and M. J. E. Gales, “Combining a gaussian mixture model frontend with mfcc parameters,” *Proceedings ICSLP*, 2002.
- [15] G. Hinton, L. Deng, D. Yu, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, Nov 2012.
- [16] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067 – 2080, Feb 2011.
- [17] D. N. S. Rajendran and B. Yegnanarayana, “Features for automatic detection of voice bars in continuous speech,” *Interspeech*, 2008.
- [18] C.-Y. Lin and H.-C. Wang, “Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection,” *J. Acoust. Soc. Am.* 130, p. 514–525, 2011.
- [19] A. Waibel, *Prosody and Speech Recognition*, Pitman Publishing, London, 1988.
- [20] S. J. Leow, E. S. Chng, and C. H. Lee, “Language-resource independent speech segmentation using cues from a spectrogram image,” *ICASSP*, 2015.
- [21] S. J. Leow, E. S. Chng, and C. H. Lee, “Zero resource anti-spoofing detection for unit selection based synthetic speech using image spectrogram artifacts,” *APSIPA*, 2016.
- [22] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627 – 1645, Sep 2010.
- [23] T. S. Lau, I. F. Chen, and C.-H. Lee, “Tunable keyword-aware language modeling and context dependent fillers for lvcsr-based spoken keyword search,” *Interspeech*, 2015.

REFERENCES

- [24] J. J. McClellan, R. W. Schafer, and M. A. Yoder, *Signal Processing First*, Prentice Hall, 2003.
- [25] G. Yu and J. J. Slotine, “Audio classification from time-frequency texture,” *ICASSP*, 2009.
- [26] National Center for Voice and Speech (2007 May 24), “The spectrograph & spectral analysis,” [Online]. Available: <http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/spectral.html>.
- [27] R. Hagiwara, (2009, Sep, and 19), “How to read a spectrogram,” [Online], Available: <http://home.cc.umanitoba.ca/~robh/howto.html>.
- [28] W. F. Katz, *Phonetics For Dummies*, John Wiley & Sons, Inc., New Jersey, 2013.
- [29] Macquarie University, (2014, Mar, and 7), “Speech acoustics,” [Online], Available: <http://clas.mq.edu.au/speech/acoustics/index.html>.
- [30] P. Ladefoged and I. Maddieson, *The Sounds of the World’s Languages*, lackwell Publishers Ltd, Oxford, 1996.
- [31] Encyclopedia Britannica, (2015, May, and 16), “Phoneme,” [Online], Available: <http://global.britannica.com/EBchecked/topic/457241/phoneme>.
- [32] F. Lim, M. R. P. Thomas, and I. J. Tashev, “Blur kernel estimation approach to blind reverberation time estimation,” *ICASSP*, 2015.
- [33] T. P. Weldon, W. E. Higgins, and D. F. Dunn, “Efficient gabor filter design for texture segmentation,” *Pattern Recognition*, vol. 29, no. 12, pp. 2005–2015, Dec 1996.
- [34] B.S. Manjunathi and W.Y. Ma, “Texture features for browsing and retrieval of image data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837 – 842, Aug 1996.
- [35] M. Kleinschmidt and D. Gelbart, “Improving word accuracy with gabor feature extraction,” *Proc. ICSLP*, 2002.
- [36] M. Inen, M. Pietikäinen, A.Hadid, G. Zhao, and T. Ahonen, *Computer Vision Using Local Binary Patterns*, vol. 40, Computational Imaging and Vision, 2011, 978-0-85729-747-1.
- [37] J. Ren, X. Jiang, and J. Yuan, “Noise-resistant local binary pattern with an embedded error-correction mechanism,” *IEEE Transactions on Image Processing*, vol. 22, no. 10, Oct 2013.

REFERENCES

- [38] F. Alegre, A. Amehraye, and N. Evans, “A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns,” *BTAS*, pp. 1–8, 2013.
- [39] F. Alegre, R. Vippera, A. Amehraye, and N. Evans, “A new speaker verification spoofing countermeasure based on local binary patterns,” *Interspeech*, 2013.
- [40] R. Gonzalez and R. Woods, *Digital Image Processing*, Addison Wesley, 1992.
- [41] J. F. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.
- [42] F. Ingermann, P. Mermelstein, and Haskin Laboratories, “Speech recognition through spectrogram matching,” *Journal of the Acoustical Society of America*, vol. 57, no. 1, January 1975.
- [43] D. H. Klatt and K. N. Stevens, “On the automatic recognition of continuous speech: Implications from a spectrogram-reading experiment,” *IEEE Transactions on Audio and Electroacoustics*, vol. AU-21, no. 3, June 1973.
- [44] P. W. Nye, F. S. Cooper, and P. Mermelstein, “Interactive experiments with a digital pattern playback,” *Haskin Laboratories: Status Report on Speech Research SR-49*, 1977.
- [45] V. W. Zue and R. A. Cole, “Experiments on spectrogram reading,” *ICASSP*, pp. 116–119, 1979.
- [46] V. W. Zue and L. F. Lamel, “An expert spectrogram reader: A knowledge-based approach to speech recognition,” *ICASSP*, 1986.
- [47] J. Johannsen, J. MacAllister, T. Michalek, and S. Rose, “A speech spectrogram expert,” *ICASSP*, 1983.
- [48] L. F. Lamel, “A knowledge-based system for stop consonant identification based on speech spectrogram reading,” *Computer Speech and Language*, 1993.
- [49] J. H. Connolly, E. A. Edmonds, J. J. Guzy, S. R. Johnson, and A. Woodcock, “Automatic speech recognition based on spectrogram reading,” *International Journal of Man-Machine Studies*, vol. 24, pp. 611–621, 1986.
- [50] K. Hatazaki, Y. Komori, T. Kawabata, and K. Shikano, “Phoneme segmentation using spectrogram reading knowledge,” *ICASSP*, 1989.
- [51] H. C. Leung and V. W. Zue, “Visual characterization of speech spectrograms,” *ICASSP*, 1986.

REFERENCES

- [52] J. F. Hemdal and R. M. Loughheed, “Morphological approaches to the automatic extraction of phonetic features,” *IEEE Transactions on Signal Processing*, vol. 39, no. 2, February 1991.
- [53] R. Steinberg and D. O’Shaughnessy, “Segmentation of a speech spectrogram using mathematical morphology,” *ICASSP*, 2008.
- [54] C. Hory, N. Martin, and A. Chehikian, “Spectrogram segmentation by means of statistical features for non-stationary signal interpretation,” *IEEE Transactions on Signal Processing*, vol. 50, pp. 2915–2925, 2002.
- [55] C. Hory and N. Martin, “Mixture densities formulation of a spectrogram segmentation task,” *Proceedings of EUSIPCO*, pp. 427–430, 2002.
- [56] F. R. Rodriguez, W. Liu, N. Evans, and J. Mason, “Morphological filtering of speech spectrograms in the context of additive noise,” *Eurospeech*, 2003.
- [57] W. Liu, V. R. Bastante, F. R. Rodriguez, N. Evans, and J. Mason, “Morphological filtering of spectrograms for automatic speech recognition,” *Proc. Visual., Imag., Image Process.*, 2004.
- [58] H.G. Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” *ISCA ITRW ASR2000 ‘Automatic Speech Recognition: Challenges for the Next Millennium’*, 2000.
- [59] B. Pinkowski, “Multiscale fourier descriptors for classifying semivowels in spectrograms,” *Pattern Recognition*, vol. 26, no. 10, pp. 1593–1602, 1993.
- [60] A. Folkers and H. Samet, “Content-based image retrieval using fourier descriptors on a logo database,” *Proceedings of International Conference on Pattern Recognition*, vol. 3, pp. 521–524, Aug 2002.
- [61] B. Pinkowski, “Principal component analysis of speech spectrogram images,” *Pattern Recognition*, vol. 30, no. 5, pp. 777–787, 1997.
- [62] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
- [63] K. Schutte and J. Glass, “Speech recognition with localized time-frequency pattern detectors,” *Automatic Speech Recognition and Understanding Workshop*, 2007.
- [64] Y. Freund and E. R. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *European Conference on Computational Learning Theory*, pp. 23–27, 1995.
- [65] J. Dennis, H. D. Tran, H. Li, and E. S. Chng, “A discriminatively trained hough transform for frame-level phoneme recognition,” *ICASSP*, 2014.

REFERENCES

- [66] P. Khunarsal, C. Lursinsap, and T. Raichareon, “Singing voice recognition based on matching of spectrogram pattern,” *Proceedings of International Joint Conference on Neural Networks*, Atlanta, Georgia, USA, 2009.
- [67] T. A. Lampert and S. E. M. O’Keefe, “A survey of spectrogram track detection algorithms,” *Applied Acoustics*, vol. 71, pp. 87–100, 2010.
- [68] H. Deshpande, R. Singh, and U. Nam, “Classification of music signals in the visual domain,” *Proceedings of the COST-G6 Conference on Digital Audio Effects*, 2001.
- [69] M. J. Wu, Z. S. Chen, J. S. R. Jang, Y. H. Li, and C. H. Lu, “Combining visual and acoustic features for music genre classification,” *International Conference on Machine Learning and Applications*, pp. 124–129, Dec 2011.
- [70] J. Dennis, H. D. Tran, and H. Li, “Spectrogram image feature for sound event classification in mismatched conditions,” *IEEE Signal Processing Letters*, vol. 18, no. 2, 2011.
- [71] J. Dennis, H. D. Tran, and E. S. Chng, “Image feature representation of the sub-band power distribution for robust sound event classification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 2, 2013.
- [72] J. Dennis, H. D. Tran, and H. Li, “Image representation of the subband power distribution for robust sound classification,” *Interspeech*, 2011.
- [73] J. Dennis, H. D. Tran, and E. S. Chng, “Overlapping sound event recognition using local spectrogram features and the generalised hough transform,” *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1085–1093, 2013.
- [74] J. Dennis, Q. Yu, H. Tang, H. D. Tran, and H. Li, “Temporal coding of local spectrogram features for robust sound recognition,” *ICASSP*, 2013.
- [75] S. Dusan and L. Rabiner, “On the relation between maximum spectral transition positions and phone boundaries,” *Interspeech*, 2006.
- [76] D.R. Van Niekerk and E. Barnard, “Acoustic cues identifying phonetic transitions for speech segmentation,” *Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2008.
- [77] D. Torre, L. A. H. Gomez, and L. V. Grand, “Automatic phonetic segmentation,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 617–625, 2003.
- [78] J. Adell, A. Bonafonte, J. A. Gomez, and M. J. Castro, “Comparative study of automatic phone segmentation methods for tts,” *ICASSP*, 2005.

REFERENCES

- [79] C. Cucchiarini and H. Strik, “Automatic phonetic transcription: An overview,” *Proc. of Int. Congress of Phonetic Sciences*, pp. 347–350, 2003.
- [80] P. Micallef and T. Chilton, “Automatic identification of phoneme boundaries using a mixed parameter model,” *Eurospeech*, 1997.
- [81] T. Svendsen and K. Kvale, “Automatic alignment of phonemic labels with continuous speech,” *ICSLP*, 1990.
- [82] B. L. Pellom and J. H. L. Hansen, “Automatic segmentation of speech recorded in unknown noisy channel characteristics,” *Speech Communications*, vol. 25, pp. 97–116, 1998.
- [83] G. Aversano, A. Esposito, A. Esposito, and M. Marinaro, “A new text-independent method for phoneme segmentation,” *Proc. of the 44th IEEE Midwest Symposium on Circuits and Systems*, vol. 2, pp. 526–529, 2001.
- [84] Y. P. Estevan, V. Wan, and O. Scharenborg, “Finding maximum margin segments in speech,” *ICASSP*, 2007.
- [85] Y. Qiao, N. Shinomura, and N. Minematsu, “Unsupervised optional phoneme segmentation: Objectives, algorithm and comparisons,” *ICASSP*, 2008.
- [86] N. J. Shah, B. B. Vachhani, H. B. Sailor, and H. A. Patil, “Effectiveness of plp-based phonetic segmentation for speech synthesis,” *ICASSP*, 2014.
- [87] W. K. Pratt, “Median filtering,” *Semiannual Report, Image Processing Institute University of Southern California*, pp. 116–123, 1975.
- [88] R. E. Woods and R. C. Gonzales, “Real-time digital image enhancement,” *Proceedings of IEEE*, vol. 69, no. 5, pp. 634–654, 1981.
- [89] J. B. Allen, “Short term spectral analysis, synthesis, and modification by discrete fourier transform,” *ICASSP*, 1977.
- [90] P. G. Garn-Nunn and J. M. Lynn, *Calvert’s Descriptive Phonetics, Third Edition*, NY:Thieme Medical Publishers, Inc., 2004.
- [91] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” *ICASSP*, vol. 1, pp. 373–376, May 1996.
- [92] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, “Evaluation of speaker verification security and detection of hmm-based synthetic speech,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, 2012.
- [93] M. Wester, Z. Wu, and J. Yamagishi, “Human vs machine spoofing detection on wideband and narrowband data,” *INTERSPEECH*, 2015.

REFERENCES

- [94] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, , and S. King, “Sas: A speaker verification spoofing database containing diverse attacks,” *ICASSP*, 2015.
- [95] Z. Wu and H. Li, “Voice conversion and spoofing attack on speaker verification systems,” *APSIPA*, 2013.
- [96] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, “Spoofing and countermeasures for speaker verification: a survey,” *Speech Communication*, pp. 130–153, 2015.
- [97] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
- [98] T. Satoh., T. Masuko, T. Kobayashi, and K. Tokuda, “A robust speaker verification system against imposture using an hmm-based speech synthesis system,” *INTERSPEECH*, pp. 759–762, September 2001.
- [99] L. W. Chen, W. Guo, and L. R. Dai, “Speaker verification against synthetic speech,” *ISCSLP*, pp. 309–312, November 2010.
- [100] N. Evans, T. Kinnunen, and J. Yamagishi, “Spoofing and countermeasures for automatic speaker verification,” *INTERSPEECH*, pp. 925–929, 2013.
- [101] Z. Wu, X. Xiao, E. S. Chng, and H. Li, “Synthetic speech detection using temporal modulation feature,” *ICASSP*, pp. 7234–7238, May 2013.
- [102] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, “Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing,” *INTERSPEECH*, September 2015.
- [103] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, “Development of crim system for the automatic speaker verification spoofing and countermeasures challenge 2015,” *INTERSPEECH*, September 2015.
- [104] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, “Spoofing detection with dnn and one-class svm for the asvspoof 2015 challenge,” *INTERSPEECH*, September 2015.
- [105] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, “Spoofing speech detection using high dimensional magnitude and phase features: the ntu approach for asvspoof 2015 challenge,” *INTERSPEECH*, September 2015.
- [106] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, “Relative phase information for detecting human speech and spoofed speech,” *INTERSPEECH*, September 2015.

REFERENCES

- [107] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, “Stc anti-spoofing systems for the asvspoof 2015 challenge,” *arXiv:1507.08074*, 2015.
- [108] Z. Wu, E. S. Chng, and H. Li, “Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition,” *INTERSPEECH*, September 2012.
- [109] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*, Prentice-Hall Inc., 2002.
- [110] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, “The aholab rps ssd spoofing challenge 2015 submission,” *INTERSPEECH*, September 2015.
- [111] P. L. De Leon, B. Stewart, and J. Yamagishi, “Synthetic speech discrimination using pitch pattern statistics derived from image analysis,” *INTERSPEECH*, September 2012.
- [112] A. Ogihara, U. N. N. O. Hitoshi, and A. Shiozaki, “Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification,” *IE-ICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 88, no. 1, pp. 280–286, 2005.
- [113] S. Weng, S. Chen, L. Yu, X. Wu, W. Cai, Z. Liu, and M. Li, “The sysu system for the interspeech 2015 automatic speaker verification spoofing and countermeasures challenge,” *arXiv:1507.06711*, 2015.
- [114] F. Alegre, A. Amehraye, and N. Evans, “A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns,” *BTAS*, pp. 1–8, September 2013.
- [115] F. Alegre, R. Vippera, A. Amehraye, and N. Evans, “A new speaker verification spoofing countermeasure based on local binary patterns,” *INTERSPEECH*, August 2013.
- [116] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. S., and A. Sizov, “Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” *INTERSPEECH*, September 2015.
- [117] T. B. Patel and H. A. Patil, “Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech,” *INTERSPEECH*, September 2015.
- [118] W. Y. Wang and K. Georgila, “Automatic detection of unnatural word-level segments in unit-selection speech synthesis,” *ASRU*, pp. 289–294, December 2011.
- [119] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, “Spoofing detection from a feature representation perspective,” *ICASSP*, 2016.

REFERENCES

- [120] P. Podder, T. Z. Khan, M. H. Khan, and M. M. Rahman, “Comparative performance analysis of hamming, hanning and blackman window,” *International Journal of Computer Applications*, vol. 96, no. 18, 2014.
- [121] R. Brunelli, *Template matching techniques in computer vision: theory and practice*, John Wiley & Sons, 2009.
- [122] P. Cosi and E. Zovato, “Lyon’s auditory model inversion: a tool for sound separation and speech enhancement,” *Proceedings of ESCA Workshop on ‘The Auditory Basis of Speech Perception’*, pp. 194–197, 1996.