Voice morphing

A Seminar Report
submitted in partial fulfillment of the
requirements for the award of the Degree of

MASTER OF COMPUTER APPLICATIONS
under
APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

by
Mohammed Arif T
MES21MCA-2023

DEPARTMENT OF COMPUTER APPLICATIONS
MES COLLEGE OF ENGINEERING
KUTTIPPURAM, MALAPPURAM - 679 582

February 2023

MES COLLEGE OF ENGINEERING

KUTTIPPURAM, KERALA - 679 582
(An Institution With NBA Accredited Departments,
Approved By AICTE, Afficiated to APJ Abdul Kalam Technological University and
Accredited by NAAC With 'A' Grade)

DEPARTMENT OF COMPUTER APPLICATIONS

CERTIFICATE

This is to certify that seminar entitled **Voice morphing** has been prepared and presented by **MOHAMMED ARIF T (MES21MCA-2023)**, fourth semester student of the department, during the academic year 2022-23, in partial fulfillment of the requirements for the award of *Degree of Master of Computer Applications* under *APJ Abdul Kalam Technological University.*

Faculty Guide Head of the Department

Date:

ACKNOWLEDGEMENT

My endeavor stands incomplete without dedicating my gratitude to a few people who have contributed towards the successful completion of my seminar.

I pay my gratitude to the Almighty for His invisible help and blessing for the fulfillment of this work .

At the outset I express my heart full thanks to our **Head of the Department**, **Prof.Hyderali.K**, Associate Professor, for permitting me to present this seminar. I would like to express my sincere gratitude to , our **Seminar coordinator**, **Syed Feroze Ahmed M**, Assistant Professor for his exceptional support and encouragement throughout this project.

I take this opportunity to express my profound gratitude to my **guide**, **Syed Feroze Ahmed M**, Assistant Professor for his valuable guidance, support and help in presenting my seminar.

I am also grateful to all our teaching and non teaching staff for their encouragement, guidance and whole-hearted support.

Last but not least, I am gratefully indebted to my family and friends, who gave me their precious help in presenting my seminar.

Sincerely,
Mohammed Arif T
MES21MCA-2023

SYNOPSIS

Voice morphing is a technique used to alter the characteristics of a person's voice. This can include changing the pitch, tone, and other aspects of the voice to make it sound different.Voice morphing is often used in voice recognition systems and speech synthesis applications,as well as in the entertainment industry for voiceovers and special effects.

The major properties of concern as far as a speech signal is concerned are its pitch and envelope information. These two reside in a convolved form in a speech signal. Hence some efficient method for extracting each of these is necessary. We have adopted an uncomplicated approach namely cepstral analysis to do the same. Pitch and formant information in each signal is extracted using the cepstral approach. Necessary processing to obtain the morphed speech signal include methods like Cross fading of envelope information, Dynamic Time Warping to match the major signal features (pitch) and Signal Re-estimation to convert the morphed speech signal back into the acoustic waveform.

Some voice morphing software also includes advanced features like artificial intelligence and neural networks, machine learning which can be used to create highly realistic and natural-sounding voices.Training and conversion are two important stages in voice morphing. Training: This includes segmentation of the source and target speaker voices in into equal frames then analyzing it is based upon a Linear Predictive Coding model. This is done in order to extract vocal features to be transformed Conversion: This involves features transformation from source to target. In order to have a better alignment between source and the target features, there is an application of Dynamic Time Warping

TABLE OF CONTENTS

1 INTRODUCTION 1

1.1 What is Voice Morphing 1

1.2 History Of Voice Morphing 2

2 WHY Voice Morphing? 3

2.1 Relevance of Voice Morphing 3

2.2 An Introspection Of The Morphing 3

2.3 Process Of Voice Morphing 5

2.3.1 Acoustics of speech production 6

2.4 Real Time Voice Morphing 6

2.4.1 Magic Mic 7

2.4.2 Murf AI 8

3 WORKING OF VOICE MORPHING 9

3.1 Voice Morphing 9

3.1.1 Pre Processing 9

3.1.2 Signal Acquisition 9

3.1.3 Windowing 10

3.1.4 Matching and Warping 10

3.1.5 Dynamic Time Warping (DTW) 11

3.1.6 signal re-estimation 12

3.2 Morphing Stage 12

3.3 Future Of Voice Morphing 14

4 CONCLUSION 15

REFERENCES 16

Chapter 1

INTRODUCTION

1.1 What is Voice Morphing

Voice morphing is a technique used to alter the characteristics of a person's voice. This can include changing the pitch, tone, and other aspects of the voice to make it sound different. Voice morphing is often used in voice recognition systems and speech synthesis applications, as well as in the entertainment industry for voiceovers and special effects.

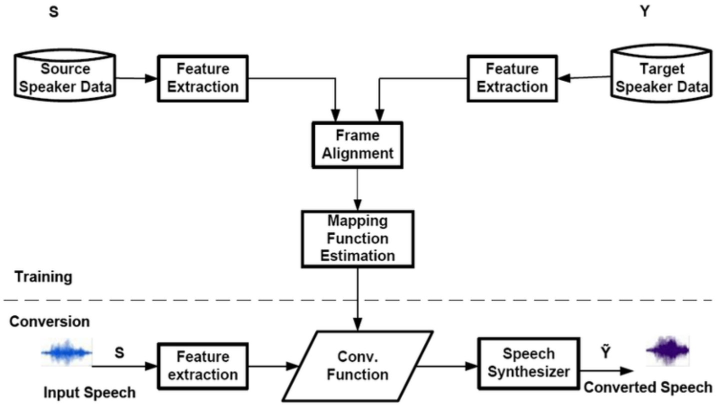

Figure 1.1: Voice morphing Diagram

Voice morphing is a technique used to alter the characteristics of a person's voice. This can include changing the pitch, tone, and other aspects of the voice to make it sound different.Voice morphing is often used in voice recognition systems and speech synthesis applications,as well as in the entertainment industry for voiceovers and special effects.

The major properties of concern as far as a speech signal is concerned are its pitch and envelope information. These two reside in a convolved form in a speech signal. Hence some efficient method for extracting each of these is necessary. We have adopted an uncomplicated approach namely cepstral analysis to do the same. Pitch and formant information in each signal is extracted using the cepstral approach. Necessary processing to obtain the morphed speech signal include methods like Cross fading of envelope information, Dynamic Time Warping to match the major signal features (pitch) and Signal Re-estimation to convert the morphed speech signal back into the acoustic waveform

Some voice morphing software also includes advanced features like artificial intelligence and neural networks, machine learning which can be used to create highly realistic and natural-sounding voices(e.g..murf AI,vall-e,etc..)

1.2 History Of Voice Morphing

Voice morphing is a technology developed at Los Alamos National Laboratory in New Mexico, USA by George Papcun and publicly demonstrated in 1999. With different names, and using different signal processing techniques, the idea of audio morphing is well known in computer music community (Serra, 1994; Tellman, Haken, Holloway, 1995; Osaka, 1995; Slaney, Covell, Lassister, 1996; Settel, Lippe, 1996). Voice morphing enables speech patterns to be cloned and an accurate cop of a persons voice be made which can then say anything the operator wishes it to say.

In 1990, the US department of defense considered using voice morphing to produce a propaganda recording of Iraqi president Saddam Hussein, which could then be distributed throughout the Arab world and Iraq to discredit the Iraqi leader

Chapter 2

WHY Voice Morphing?

2.1 Relevance of Voice Morphing

Voice morphing is used to train speech recognition systems to be more robust and handle a wide range of voices and accents. It can also be used in speech synthesis systems to generate more natural-sounding voices. Voice morphing is used in the entertainment industry to create unique characters and special effects.

It can be used to change a person's voice to sound like someone else, or to create a completely new, artificial voice. Voice morphing can be used to create synthetic speech that can be used to bypass voice-based authentication systems. This is a concern for organizations that rely on voice biometrics for security.

Overall, the relevance of Voice morphing can be used to identify a person who is using a disguised voice, and can be used to match a voice print to a suspect in a criminal investigation.

2.2 An Introspection Of The Morphing

Speech morphing can be achieved by transforming the signal's representation from the acoustic waveform obtained by sampling of the analog signal, with which many people are familiar with, to another representation. To prepare the signal for the transformation, it is split into a number of 'frames'

- sections of the waveform. The transformation is then applied to each frame of the signal. This provides another way of viewing the signal information. The new representation (said to be in the frequency domain) describes the average energy present at each frequency band.

Further analysis enables two pieces of information to be obtained: pitch information and the overall envelope of the sound. A key element in the morphing is the manipulation of the pitch information. If two signals with different pitches were simply cross-faded it is highly likely that two separate sounds will be heard. This occurs because the signal will have two distinct pitches causing the auditory system to perceive two different objects. A successful morph must exhibit a smoothly changing pitch throughout. The pitch information of each sound is compared to provide the best match between the two signals' pitches.

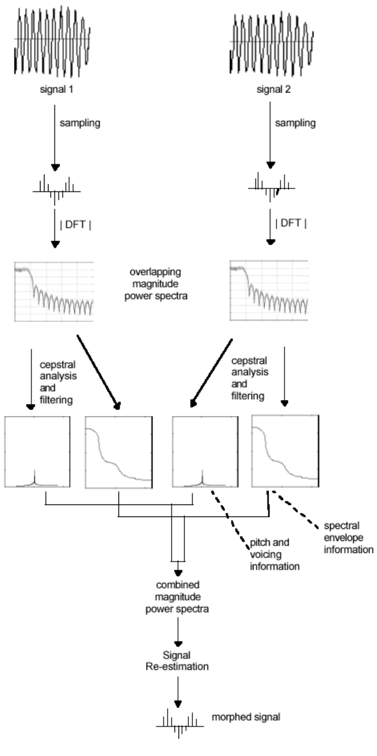
signal 1
signal 2
sampling
sampling
| DFT |
| DFT |
overlapping magnitude power spectra
cepstral analysis and filtering
cepstral analysis and filtering
pitch and voicing information
spectral envelope information
combined magnitude power spectra
Signal Re-estimation
morphed signal

Figure 2.1: Schematic block diagram of the speech morphing process

To do this match, the signals are stretched and compressed so that important sections of each signal match in time. The interpolation of the two sounds can then be performed which creates the intermediate sounds in the morph. The final stage is then to convert the frames back into a normal waveform.

2.3 Process Of Voice Morphing

The algorithm to be used is shown in the simplified block diagram given below. The algorithm contains a number of fundamental signal processing methods including sampling, the discrete Fourier transform and its inverse, cepstral analysis. However the main processes can be categorized as follows.

- Pre-processing or representation conversion: This involves processes like signal acquisition in discrete form and windowing.
- Cepstral analysis or Pitch and Envelope analysis: This process extracts the pitch and formants information in the signal.
- Morphing which includes Warping and interpolation
- Signal Re-estimation

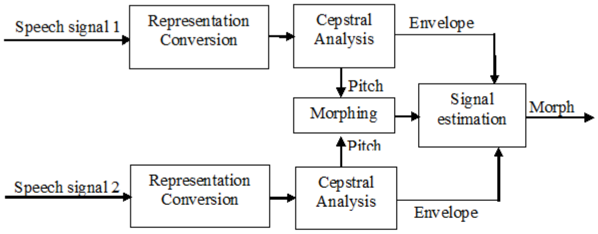

Figure 2.2: Block diagram of the simplified speech morphing algorithm

2.3.1 Acoustics of speech production

Speech production can be viewed as a filtering operation in which a sound source excites a vocal tract filter. The source may be periodic, resulting in voice speech, noisy and a periodic, causing unvoiced speech. As a periodic signal, voice speech has a spectra consisting of harmonics of the fundamental frequency of the vocal cord vibration; this frequency often abbreviated as F0, is the physical aspect of the speech signal corresponding to the perceived pitch. Thus pitch refers to the fundamental frequency of the vocal cord vibrations or the resulting periodicity in the speech signal. This F0 can be determine either from the periodicity in the time domain or from the regularly spaced harmonics in the frequency domain.

The vocal tract can be modelled as an acoustic tube with resonances, called formants, an anti-resonances. Moving certain structures in the vocal tract alters the shape of the acoustic tube, which in turn changes its frequency response. The filter amplifies energy at and near formants. The common method used to extract pitch and formants frequencies is the spectral analysis. This method views speech as the output of a liner, time-varying system (vocal tract) excited by either quasiperiodic pulses or random noise.

Since the speech signal is the result of convolving excitation and vocal tract sample response, separating or ¡de-convolving= the two components can be used. In general, de-convolution of the two signals is impossible, but it work for speech, because the two signals have quite different characteristics. The de-convolution process transforms a product of two signals into a sum of two signals. If the resulting summed signals are sufficiently different spectrally, they may be separated by linear filtering. Now we present a comprehensive analysis of each of the processes involved in morphing with the aid of block diagrams wherever necessary.

2.4 Real Time Voice Morphing

In real time voice morphing what we want is to be able to morph, in real-time user singing a melody with the voice of another singer. It results in an “impersonating” system with which the user can morph his/her voice attributes, such as pitch, timbre, vibrato and articulation, with the ones from a prerecorded target singer. The user is able to control the degree of morphing, thus being able to choose the level of “impersonation” that he/she wants to accomplish.

In our particular implementation we are using as the target voice a recording of the complete song to be morphed. A more useful system would use a database of excerpts of the target voice, thus choosing the appropriate target segment at each particular time in the morphing process. In order to incorporate to the user’s voice the corresponding characteristics of the “target” voice, the system has to first recognize what the user is singing(phonemes and notes), finding the same sounds in the target voice (i.e. synchronizing the sounds), then interpolate the selected voice attributes, and finally generate the output morphed voice.

All this has to be accomplished in real-time.There have been several examples in voice morphing systems around the world. Here are a few examples

2.4.1 Magic Mic

MagicMic is a real-time voice changer so it automatically converts the voice into the selected voice. Main features are voice and sound effects, AI voice cloning, voice customization, keybinds control, voice changing background sound. It comes with 125+ voice-changing options for you to choose from, including AI voices. Some of them come with background sounds.

It can import MP3/MAV audio files Set keybind for voice and sound effects so that you can play and stop them quickly while gaming or live. Customize your special and unique voice and name for it with the voice studio function. Some voice effects come with background sounds so that you can change the environment and change your voice more naturally. It can work for various platforms like PUBG, Second Life, Fortnite, Skype, live streaming, online education, entertainment prank 400+ Voice Effects and 150+ Voice Memes to Use on Your Favorite Programs.

2.4.2 Murf AI

Murf AI is a cloud-based realistic text-to-speech(TTS) platform that can be used to create voiceovers for their content. AI and deep machine learning technology to generate these ultra-realistic voiceovers across a range of 120+ voices in 20+ languages. AI voices are synthetic voices that mimic human speech through a process called deep learning, where artificial intelligence is used to convert text into speech.

AI offers great creative control by letting users fine-tune the punctuation, pitch, interjections, speed, emphasis, and tone of an AI voice. It mainly used to YouTube videos, podcasts, advertisements/ commercials, e-learning content, presentations, audiobooks, etc. . . . . .

Chapter 3

WORKING OF VOICE MORPHING

3.1 Voice Morphing

Voice morphing is a technique used to alter the characteristics of a person's voice. This can include changing the pitch, tone, and other aspects of the voice to make it sound different. Here is a brief overview of how voice morphing works:

3.1.1 Pre Processing

This section shall introduce the major concepts associated with processing a speech signal and transforming it to the new required representation to affect the morph. This process takes place for each of the signals involved with the morph.

3.1.2 Signal Acquisition

Once the speech signal is captured then amplified and filtered to remove any noise or interference. The signal is then digitized using an analog-to-digital converter (ADC) or MIC and CODEC. The analog speech signal is converted into the discrete form by the inbuilt CODEC TLC320AD535 present onboard and stored in the processor memory.

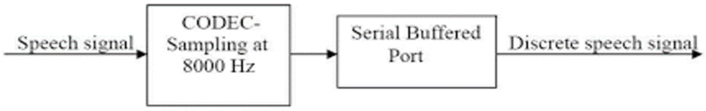

Figure 3.1: Analog-to-Digital converter

3.1.3 Windowing

The windowing process helps to improve the quality of the morphed voice by reducing distortion and artifacts caused by spectral leakage. windowing process involves dividing the speech signal into small frames or windows, typically ranging from 20 to 40 milliseconds in length. Each window is multiplied by a windowing function Hanning, then frames are put back together, which reduces the amplitude of the signal at the beginning and end of the window.
When the frames are put back together, modulation in the signal becomes evident. So a simple method to overcome modulation is to use overlapping windows. In overlapping spectra, as one frame fades out, its successor fades in.

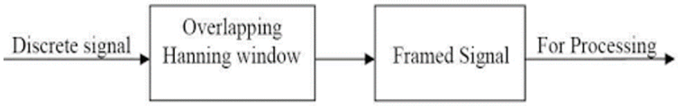

Figure 3.2: Hanning windowing

3.1.4 Matching and Warping

Both signals will have a number of ’time-varying properties’. To create an effective morph, it is necessary to match one or more of these properties of each signal to those of the other signal in some way. The property of concern is the pitch of the signal - although other properties such as the amplitude could be used - and will have a number of features. It is almost certain that matching features do not occur at exactly the same point in each signal. Therefore, the feature must be moved to some point in between the position in the first sound and the second sound. In other words, to smoothly morph the pitch information, the pitch present in each signals needs to be matched and then the amplitude at each frequency cross-faded. To perform the pitch matching, a pitch contour for the entire signal is required. This is obtained by using the pitch peak location in each cepstral pitch slice.

Consider the simple case of two signals, each with two features occurring in different positions as shown in the figure below.

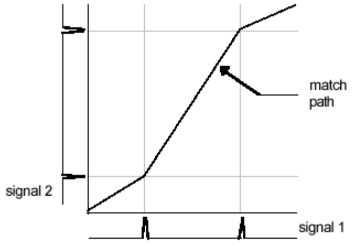

Figure 3.3: The match path between two signals

The match path shows the amount of movement (or warping) required in order aligning corresponding features in time. Such a match path is obtained by Dynamic Time Warping (DTW).

3.1.5 Dynamic Time Warping (DTW)

Speaker recognition and speech recognition are two important applications of speech processing. These applications are essentially pattern recognition problems, which is a large field itself. Some Automatic speech Recognition (ASR) systems employ the normalization. This is the process by which time-varying features within the words are brought in to line. The current method is time-warping in which the time axis of the unknown word is non-uniformly distorted to match its features to those of the pattern word. The degree of discrepancy between the unknown word and the pattern – the amount of warping required to match the two words – can be used directly as distance

measure.
Such time-warping algorithm is usually implemented by dynamic programming and is known as Dynamic Time Warping. DTW is used to find the best match between the features of the two sounds in this case, their pitch. To create a successful morph, major features, which occur at generally the same time in each signal, ought to remain fixed and intermediate features should be moved or interpolated. DTW enables a match path to be created. This shows how each element in one signal corresponds to each element in the second signal

3.1.6 signal re-estimation

signal re-estimation is the process of generating a new speech signal from the transformed speech signal to make it sound more natural and human-like signal re-estimation is to use a technique known as spectral smoothing or spectral shaping. Spectral smoothing involves applying a smoothing function to the spectral envelope of the voice signal.they are.....

- Two signals with different pitches were simply cross-faded which produces two separate sounds .
- The pitch information of each sound is compared to provide the best match between the two signals' pitches.
- To do this match, the signals are stretched and compressed so that important sections of each signal match in time.
- The interpolation of the two sounds can then be performed which creates the intermediate sounds in the morph.
- The final stage is then to convert the frames back into a normal waveform.
- The information lost has to be re estimated for the morphed sound

3.2 Morphing Stage

Now we shall give a detailed account of how the morphing process is carried out. The overall aim in this section is to make the smooth transition from signal 1 to signal 2. This is partially accomplished by the 2D array of the match path provided by the DTW. At this stage, it was decided exactly what form the morph would take. The implementation chosen was to perform the morph in the duration of the longest signal. In other words, the final morphed speech signal would have the duration of the longest signal. In order to accomplish this, the 2D array is interpolated to provide the desired duration.

However, one problem still remains: the interpolated pitch of each morph slice. If no interpolation were to occur then this would be equivalent to the warped cross-fade which would still be likely to result in a sound with two pitches. Therefore, a pitch in- between those of the first and second signals must be created.

The precise properties of this manufactured pitch peak are governed by how far through the morph the process is. At the beginning of the morph, the pitch peak will take on more characteristics of the signal 1 pitch peak - peak value and peak location - than the signal 2 peak. Towards the end of the morph, the peak will bear more resemblance to that of the signal 2 peaks. The variable l is used to control the balance between signal 1 and signal 2. At the beginning of the morph, l has the value 0 and upon completion.

This diagram shows a sample cepstral slice with the pitch peak area highlighted.another sample cepstral slice, again with the same information highlighted. To illustrate the morph process, these two cepstral slices shall be used.

There are three stages:

1. Combination of the envelope information
2. Combination of the pitch information residual - the pitch information excluding the pitch peak
3. Combination of the pitch peak information.

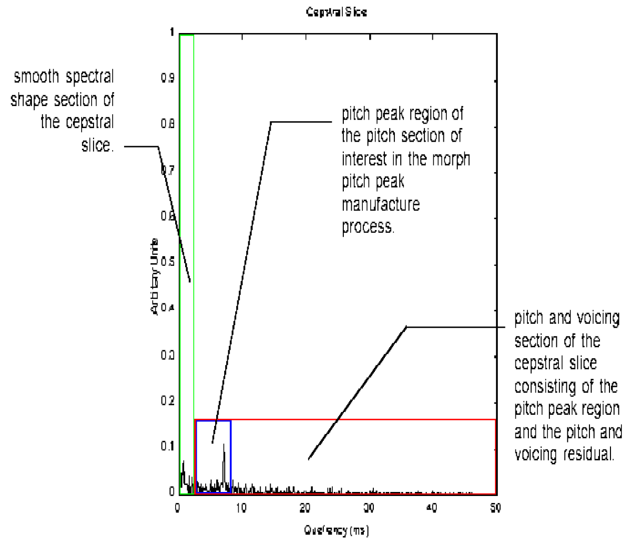

Figure 3.4: sample cepstral slice with the pitch

3.3 Future Of Voice Morphing

Microsoft has introduced VALL-E, a neural codec language model method for text-to-speech synthesis (TTS) that employs audio codec codes as intermediate representations and can replicate anyone's voice VALL-E could be used for high-quality text-to-speech applications, speech editing where a recording of a person could be edited and changed from a text transcript. Simulate anyone's voice with 3 seconds of audio

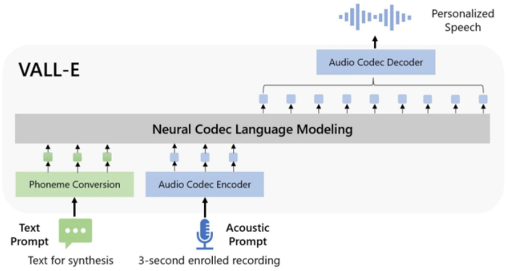

Figure 3.5: Daigram of vall-e

Chapter 4

CONCLUSION

The approach we have adopted separates the sounds into two forms: spectral envelope information and pitch and voicing information. These can then be independently modified. The morph is generated by splitting each sound into two forms: a pitch representation and an envelope representation.

The pitch peaks are then obtained from the pitch spectrograms to create a pitch contour for each sound. Dynamic Time Warping of these contours aligns the sounds with respect to their pitches. At each corresponding frame, the pitch, voicing and envelope information are separately morphed to produce a final morphed frame. These frames are then converted back into a time domain waveform using the signal re-estimation algorithm.

The longest signal is compressed and the morph has the same duration as the shortest signal (the reverse of the approach described here). If one signal is significantly longer than the other, two possibilities arise. However, according to the eventual use of the morph, a number of other types could be produced.

REFERENCES

[1] S. Kim, J. Lee, and Y. Kim - **Real-time Voice Transformation Using Deep Neural Networks**.

[2] Y. Fan, M. Liao, and X. Li - **Voice Conversion and Transformation Using Deep Learning** .

[3] T. E. Martin, L. C. Neumeyer, and J. S. Garofolo - **Voice Morphing** .

[4] **http://mi.eng.cam.ac.uk/ hy216/VoiceMorphingPrj.html** .