음성인식 합성 Overview

서울대학교 전기정보공학부
교수 성원용

2017. Spring

인공지능에이전트

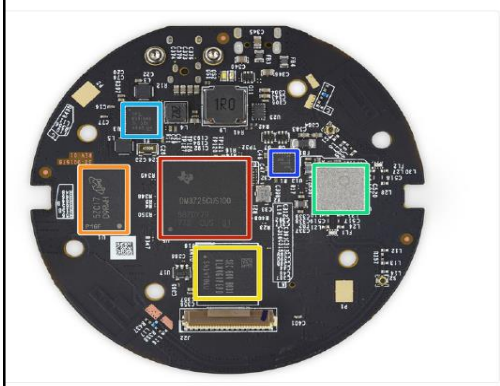
- 아마존 Alexa
- 주요기능
  - 음성 beamforming
    - Seven mic's
  - 음성인식
  - 음성합성
  - 대화생성(Q&A)

  - *음성인식, 합성, 대화생성은 Server에서 이루어짐

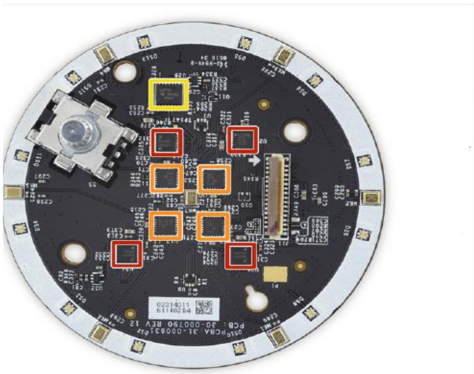
Amazon Echo Dot

8

Edit

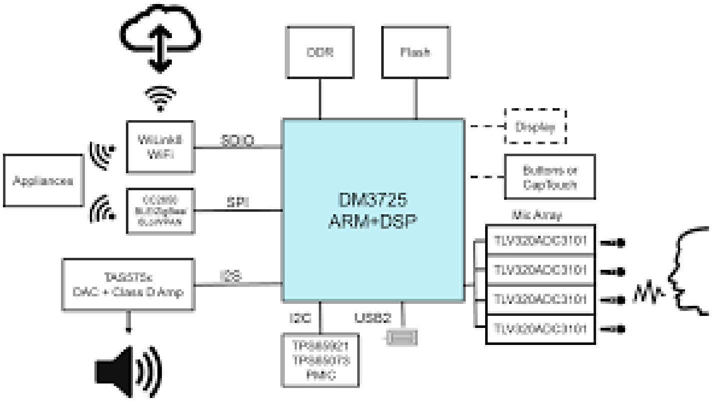


- Chips on one side, ports on the other. Here's what this motherboard is packing:
  - Texas Instruments DM3725 Digital Media Processor
  - Micron MT46H64M32LFBQ 256 MB (16 Meg x 32 x 4 Banks) LPDDR SDRAM
  - Samsung KLM4G1FEPD 4GB High Performance eMMC NAND Flash Memory
  - Qualcomm Atheros QCA6234 Integrated Dual-Band 2x2 802.11n + Bluetooth 4.0 SiP
  - Texas Instruments TPS65910A1 Integrated Power Management IC
  - Texas Instruments DAC



- And buried in the top layer we find another control board, functionally identical to the one we dug up in the original Echo.
  - National Semiconductor LP55231 Programmable 9-Output LED Driver (x4)
  - Texas Instruments TLV320ADC3101 92dB SNR Low-Power Stereo ADC (x4)
  - Texas Instruments SN74LVC74A Dual Positive-Edge-Triggered D-Type Flip-Flops

Amazon Echo block diagrm



Amazon Echo Inside

- https://www.ifixit.com/Teardown/Amazon+Echo+Dot+Teardown/61304

- Texas Instruments DM3725 Digital Media Processor
- Micron MT46H64M32LFBQ 256 MB (16 Meg x 32 x 4 Banks) LPDDR SDRAM
- Samsung KLM4G1FEPD 4GB High Performance eMMC NAND Flash Memory
- Qualcomm Atheros QCA6234 Integrated Dual-Band 2x2 802.11n + Bluetooth 4.0 SiP
- Texas Instruments TPS65910A1 Integrated Power Management IC
- Texas Instruments DAC
- National Semiconductor LP55231Programmable 9-Output LED Driver (x4)
- Texas Instruments TLV320ADC3101 92dB SNR Low-Power Stereo ADC (x4)
- Texas Instruments SN74LVC74A Dual Positive-Edge-Triggered D-Type Flip-Flops

TLV320AIC3101

- Stereo ADC (Analog Digital Converter)

Texas Instruments — TLV320AIC3101

SLAS520E – FEBRUARY 2007 – REVISED DECEMBER 2014

TLV320AIC3101 Low-Power Stereo Audio Codec for Portable Audio/Telephony

1 Features

- Stereo Audio DAC
  - 102-dBA Signal-to-Noise Ratio
  - 16/20/24/32-Bit Data
  - Supports Sample Rates From 8 kHz to 96 kHz
  - 3D/Bass/Treble/EQ/De-Emphasis Effects
  - Flexible Power Saving Modes and Performance are Available
- Stereo Audio ADC
  - 92-dBA Signal-to-Noise Ratio
  - Supports Sample Rates From 8 kHz to 96 kHz
  - Digital Signal Processing and Noise Filtering Available During Record
- Extensive Modular Power Control
- Power Supplies:
  - Analog: 2.7 V–3.6 V.
  - Digital Core: 1.525 V–1.95 V
  - Digital I/O: 1.1 V–3.6 V
- Package: 5-mm × 5-mm 32-Pin QFN

2 Applications

- Digital Cameras
- Smart Cellular Phones

3 Description

The TLV320AIC3101 is a low-power stereo audio codec with stereo headphone amplifier, as well as

DM3725 Digital Media Processor

DM3730, DM3725
Digital Media Processors

Check for Samples: DM3730, DM3725

1 DM3730, DM3725 Digital Media Processors

1.1 Features

- DM3730/25 Digital Media Processors:
  - Compatible with OMAP™ 3 Architecture
  - ARM® Microprocessor (MPU) Subsystem
    - Up to 1-GHz ARM® Cortex™-A8 Core Also supports 300, 600, and 800-MHz operation
    - NEON™ SIMD Coprocessor
  - High Performance Image, Video, Audio (IVA2.2™) Accelerator Subsystem
    - Up to 800-MHz TMS320C64x+™ DSP Core Also supports 260, 520, and 660-MHz operation
    - Enhanced Direct Memory Access (EDMA) Controller (128 Independent Channels)
    - Video Hardware Accelerators
    - Load-Store Architecture With Non-Aligned Support
    - 64 32-Bit General-Purpose Registers
    - Instruction Packing Reduces Code Size
    - All Instructions Conditional
    - Additional C64x+™ Enhancements
      - Protected Mode Operation
      - Expectations Support for Error Detection and Program Redirection
      - Hardware Support for Modulo Loop Operation
  - C64x+™ L1/L2 Memory Architecture
    - 32K-Byte L1P Program RAM/Cache (Direct Mapped)
    - 80K-Byte L1D Data RAM/Cache (2-Way

Digital Signal Processor

ARM CPU

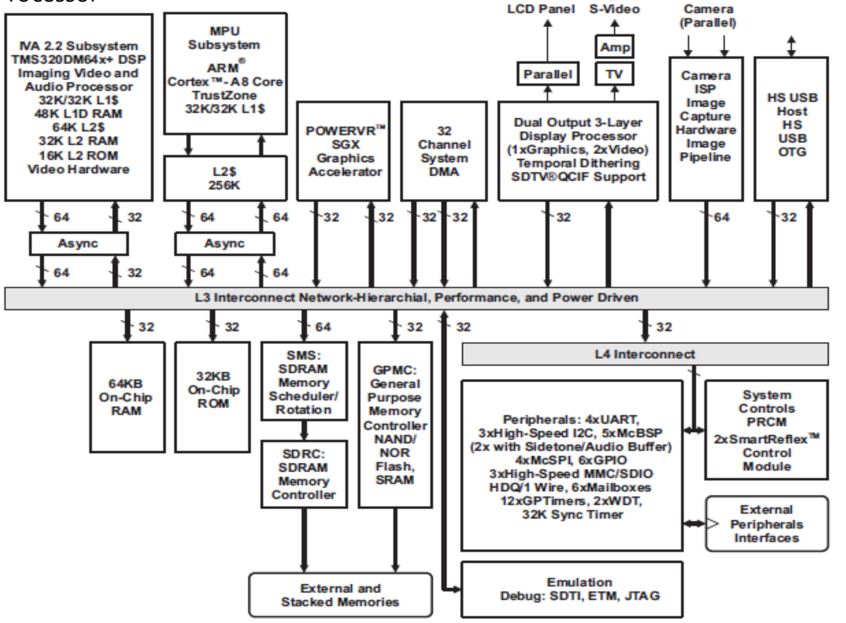


Figure 1-1. DM3730/25 Functional Block Diagram

Dual Processing CPU

- Programmable Digital Signal Processor: 매우 많은 multiply-add 연산을 처리한다. Digital signal processing, 특히 digital filtering을 매우 빠르게 처리 가능하다 (실시간 처리에 필요).
- ARM CPU – 제어 목적의 CPU 또는 메모리 공간이 많이 필요한 video 나 user interface 처리에 필요.

Audio 출력

- Digital Input D class Amp

TEXAS INSTRUMENTS

TAS5780M
SLASEG7 – DECEMBER 2016

TAS5780M Digital Input, Closed-Loop Class-D Amplifier with 96-kHz Processing

1 Features

- Flexible Audio I/O Configuration
  - Supports I^2S, TDM, LJ, RJ Digital Input
  - Sample Rate Support
  - Stereo Bridge Tied Load (BTL) or Mono Parallel Bridge Tied Load (PBTL) Operation
  - 1SPW Amplifier Modulation
  - Supports 3-Wire Digital Audio Interface (No MCLK required)
- High-Performance Closed-Loop Architecture (PVDD = 12 V, R_{SPK} = 8 Ω, SPK_GAIN = 20 dB)
  - Idle Channel Noise = 62 µVrms (A-Wtd)
  - THD+N = 0.2% (at 1 W, 1 kHz)
  - SNR = 103dB A-Wtd (Ref. to THD+N = 1%)

2 Applications

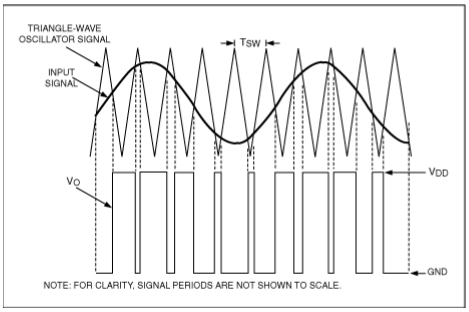
- LCD, LED TV, and Multi-Purpose Monitors
- Sound Bars, Docking Stations, and PC Audio
- Wireless Subwoofers, Bluetooth Speakers, and Active Speakers

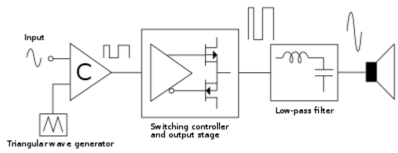
3 Description

The TAS5780M device is a high-performance, stereo closed-loop Class-D amplifier with integrated audio processor with 96-kHz architecture. To convert from digital to analog, the device uses a high performance DAC with Burr Brown™ audio technology. It requires only two power supplies: one DVDD for low-voltage circuitry and one PVDD for high-voltage circuitry. It is controlled by a software control port using standard I^2C communication.

Class D Amp

- Transistor를 on-off로만 빨리 동작해서 analog 신호를 만들어냄
- 매우 전력효율이 좋음.





https://en.wikipedia.org/wiki/Class-D_amplifier

Echo안의 memory 부품

- 에코의 하드웨어 부품에는 텍사스 인스트루먼츠 DM3725 ARM Cortex-A8 프로세서, 256MB의 LPDDR1 RAM, 4GB의 기억공간(non-volatile memory)이 있다.
- LPDDR1 RAM: Dynamic RAM (DRAM) (volatile memory, working space)
- 기억공간: NAND flash memory 등 (non-volatile memory) – program, context, downloaded contents

WiFi network

- 무선 LAN, Amazon Web Service 에 연결
- Low delay network is desired

Speech recognition

- Wake word 인식: keyword spotting 단말기에서 처리
- 음성전처리, audio beamforming (7 mic), feature extraction: 단말기
- 음성인식 및 Q&A: 서버에서 처리

- Echo's voice recognition capability is based on Amazon Web Services and the voice platform Amazon acquired from Yap,[10] Evi, and IVONA[11](a Polish-based specialist in voice technologies used in the Kindle Fire).[12]
- The smart speakers perform well with a 'good' (low latency) Internet connection which minimizes processing time due to minimal communication round trips, streamable responses and geo-distributed service endpoints.
-

https://en.wikipedia.org/wiki/Amazon_Echo#Overview_of_operation

Speech recognition

- Most natural human-computer interface
- It has been studied for more than a few decades, but it is conceived something practical recently.
- Speech recognition is a matter of combining knowledges, such as phonetic information, vocabularies, and language models.

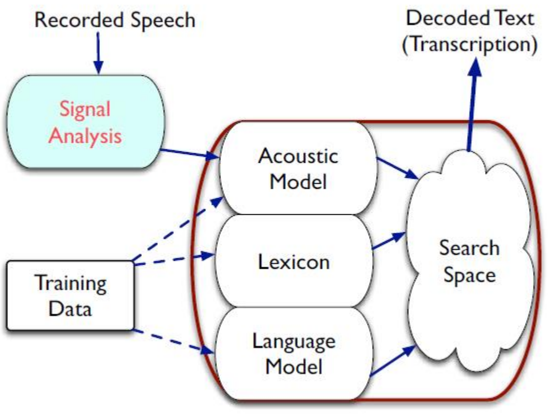
Speech recognition and application

- Large vocabulary speech recognition
  - HMM (Hidden Markov Model) based technique has been used for a long time
  - Neural network based, end to end speech recognition, is now being studied actively.
- Other applications
  - Key-word spotting
  - Speaker identification
  - V/Silence discrimination

Speech recognition history

- Until 1980's: DTW (dynamic time warping) based. Most of researches on speech were speech compression and modeling
- Around 1985: HMM based speech recognition
- From 1990's ~ 2010 Gaussian mixture model + HMM + several different speech features, speaker adaptation
- From 2012 around: DNN + HMM
- 2013: speech recognition with CTC-RNN (end to end training)
- Recent topics: Q&A system

Knowledge integration in speech recognition



Why difficult?

Several sources of variation

Size Number of word types in vocabulary, perplexity

Speaker Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics and accent

Acoustic environment Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)

Style Continuously spoken or isolated? Planned monologue or spontaneous conversation?

Current error rates

Ballpark numbers; exact numbers depend very much on the specific corpus

Task	Vocabulary	Word Error Rate %
Digits	11	0.5
WSJ read speech	5K	3
WSJ read speech	20K	3
Broadcast news	64,000+	5
Conversational Telephone	64,000+	10

HSR versus ASR

Task	Vocab	ASR	Hum SR
Continuous digits	11	.5	.009
WSJ 1995 clean	5K	3	0.9
WSJ 1995 w/noise	5K	9	1.1
SWBD 2004	65K	10?	3-4?

- Conclusions:
  - Machines about 5 times worse than humans
  - Gap increases with noisy speech
  - These numbers are rough, take with grain of salt

Statistical speech recognition

If $\mathbf{X}$ is the sequence of acoustic feature vectors (observations) and $\mathbf{W}$ denotes a word sequence, the most likely word sequence $\mathbf{W}^*$ is given by

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X})$$

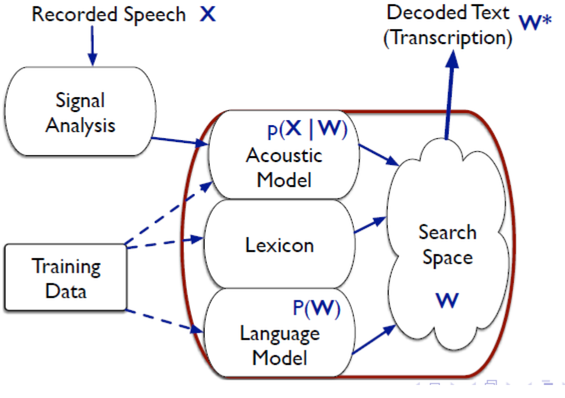
Applying Bayes' Theorem:

$$P(\mathbf{W} \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \mathbf{W})P(\mathbf{W})}{p(\mathbf{X})}$$
$$\propto p(\mathbf{X} \mid \mathbf{W})P(\mathbf{W})$$
$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \underbrace{p(\mathbf{X} \mid \mathbf{W})}_{\text{Acoustic model}} \underbrace{P(\mathbf{W})}_{\text{Language model}}$$

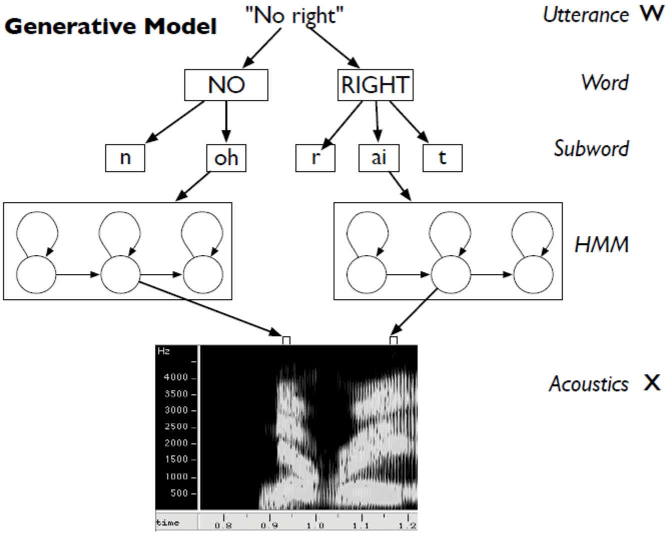
Speech recognition components

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} p(\mathbf{X} \mid \mathbf{W})P(\mathbf{W})$$

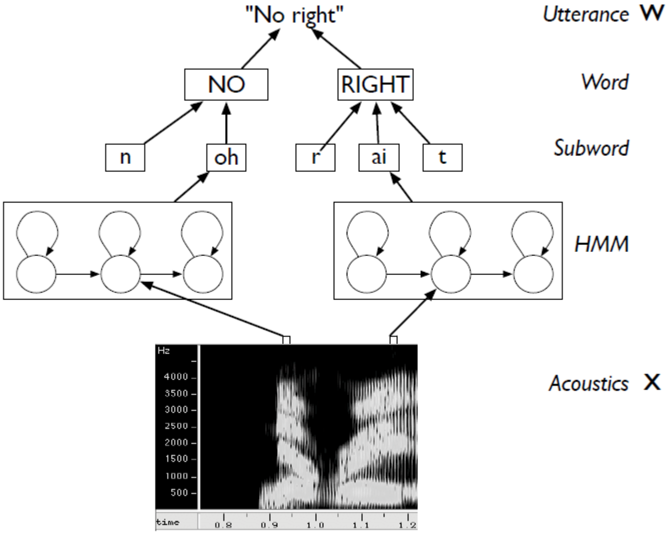
Use an acoustic model, language model, and lexicon to obtain the most probable word sequence $\mathbf{W}^*$ given the observed acoustics $\mathbf{X}$



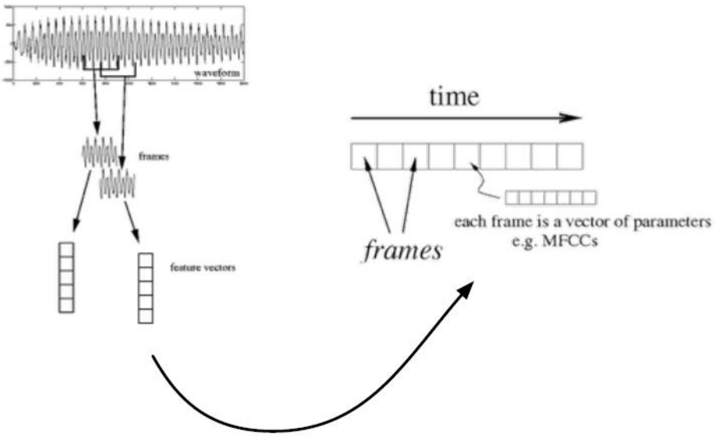
Hierarchical modeling of speech



Hierarchical modeling of speech



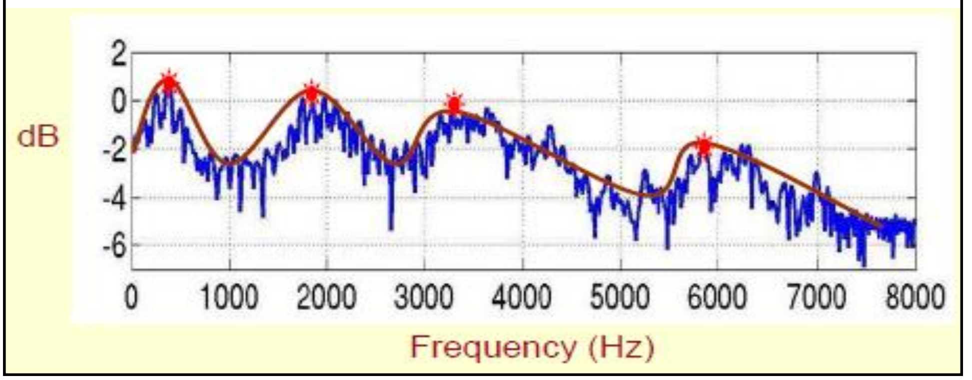
Representing speech (x)



Represent a recorded utterance as a sequence of *feature vectors*

Speech features

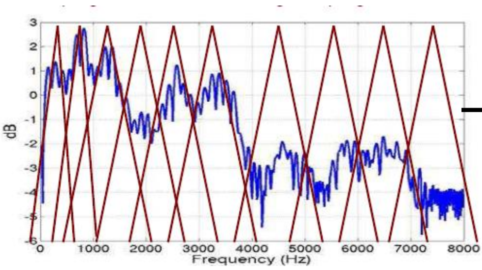
- Spectral envelope이 중요
- Pitch에 의해 생기는 골의 영향이 적어야 함.



Mel Scaled-Filter Bank – based on human auditory systems

Spectrum

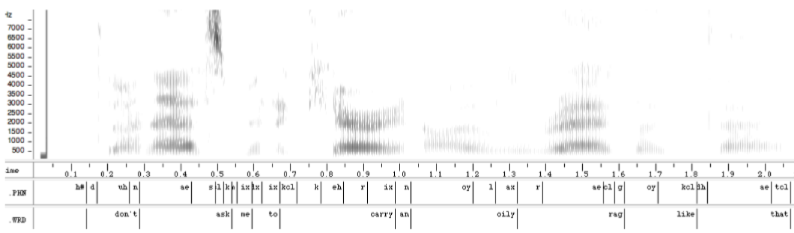
$$F\{f_n\} = F\{\sum_{n=-\infty}^{\infty} h_{m-n} e_n\}$$



Mel spectrum

mel(f)= 2595*log10(1+f/700)

Labeling speech (W)



Labels may be at different levels: words, phones, etc.
Labels may be *time-aligned* – i.e. the start and end times of an acoustic segment corresponding to a label are known

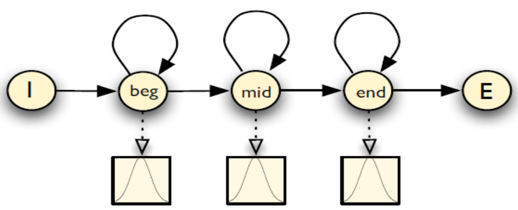
Reading: Jurafsky & Martin chapter 7 (especially sections 7.4, 7.5)

Phones and phonemes

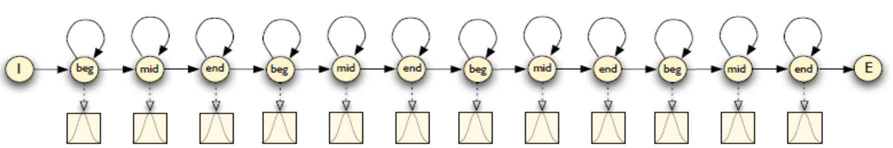
- **Phonemes**
  - abstract unit defined by linguists based on contrastive role in word meanings (eg “cat” vs “bat”)
  - 40–50 phonemes in English
- **Phones**
  - speech sounds defined by the acoustics
  - many *allophones* of the same phoneme (eg /p/ in “pit” and “spit”)
  - limitless in number
- Phones are usually used in speech recognition – but no conclusive evidence that they are the basic units in speech recognition
- Possible alternatives: syllables, automatically derived units, ...

Triphones

Three state phone models (triphone)



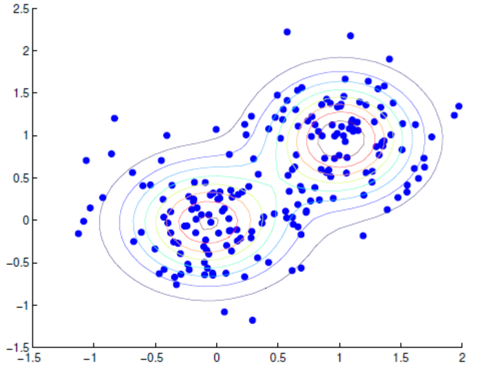
/ih/



/s/ /ih/ /k/ /s/

"six"

Gaussian mixture model of phones



Fitted with a two component GMM using EM

Example: TIMIT Corpus

- TIMIT corpus (1986)—first widely used corpus, still in use
  - Utterances from 630 North American speakers
  - Phonetically transcribed, time-aligned
  - Standard training and test sets, agreed evaluation metric (phone error rate)
- TIMIT phone recognition - label the audio of a recorded utterance using a sequence of phone symbols
  - Frame classification – attach a phone label to each frame data
  - Phone classification – given a segmentation of the audio, attach a phone label to each (multi-frame) segment
  - Phone recognition – supply the sequence of labels corresponding to the recorded utterance

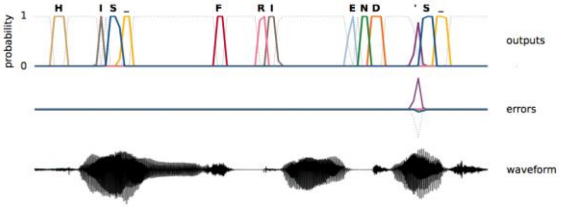
Speech recognition on TIMIT

- Train a classifier of some sort to associate each feature vector with its corresponding label. Classifier could be
  - Neural network
  - Gaussian mixture model
  - ...

  The at test time, a label is assigned to each frame
- Questions
  - What's good about this approach?
  - What the limitations? How might we address them?

CTC-trained RNN AM

- CTC (connectionist temporal classification) objective funct ion allows RNNs to learn sequences rather than frame-wi se targets.
- RNNs can directly learn to generate texts from speech d ata without the linguistic structures or dictionaries.
- Graves and Jaitly, 2014,
  - 5 stacked bidirectional LSTM layers, 500 cells in each layer



End-to-end speech recognition with CTC-RNNs

- No need of frame-wise labels. Just speech and dictated text.
- WSJ can be used for training
  - TIMIT: Recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences (6300 instances).
  - WSJ: The corpus contains approximately 78,000 training utterances (73 hours of speech)
  - Deep Speech 2 from Baidu uses more than 10,000 hours of speech for CTC-RNN training

Language modeling

- n-gram based LM: usually 3 (to 5)-gram which predicts the probabilities of the next words using previous n-1 words.
  - n-gram based LM's are implemented using memory look-up, and demands a large memory space (often GB's)
- RNN based LM
  - Use recurrent neural networks for LM and shows very good performance, but needs a large amount of computation
- LM training: text DB, Wikipedia

Evaluation

- How accurate is a speech recognizer?
- Use dynamic programming to align the ASR output with a reference transcription
- Three type of error: insertion, deletion, substitution
- Word error rate (WER) sums the three types of error. If there are N words in the reference transcript, and the ASR output has S substitutions, D deletions and I insertions, then:

$$\text{WER} = 100 \cdot \frac{S + D + I}{N}\% \qquad \text{Accuracy} = 100 - \text{WER}\%$$

- For TIMIT, define phone error error rate analagously to word error rate
- Speech recognition evaluations: common training and development data, release of new test sets on which different systems may be evaluated using word error rate

Overview of the class

- Part1: 기초 (2 weeks)
  - Digital signal processing (sampling, digital filtering)
  - Probability 기초
  - Matlab 사용법
- Part2: acoustic feature analysis (1 weeks)
  - Acoustic phonetics
  - Feature extraction
  - Lecture material: Own
  - Complementary material: Dan Jurafsky, Stanford CS224S
- Part3: conventional speech recognition (2 weeks)
  - GMM, HMM
  - Material: University of Edinburgh http://www.inf.ed.ac.uk/teaching/courses/asr/
  - Dan Jurafsky, Stanford CS224S http://web.stanford.edu/class/cs224s/
- Part4: DNN based speech recognition (2 weeks)
  - Feed-forward DNN for phoneme extraction
  - CTC RNN for end-to-end speech recognition
  - Recursive neural network for language modeling
  - HMM free speech recognition examples
- Part4: Q&A system, attention model, and memory network etc. (1 week)