Acoustic Modeling

송치성

Intro

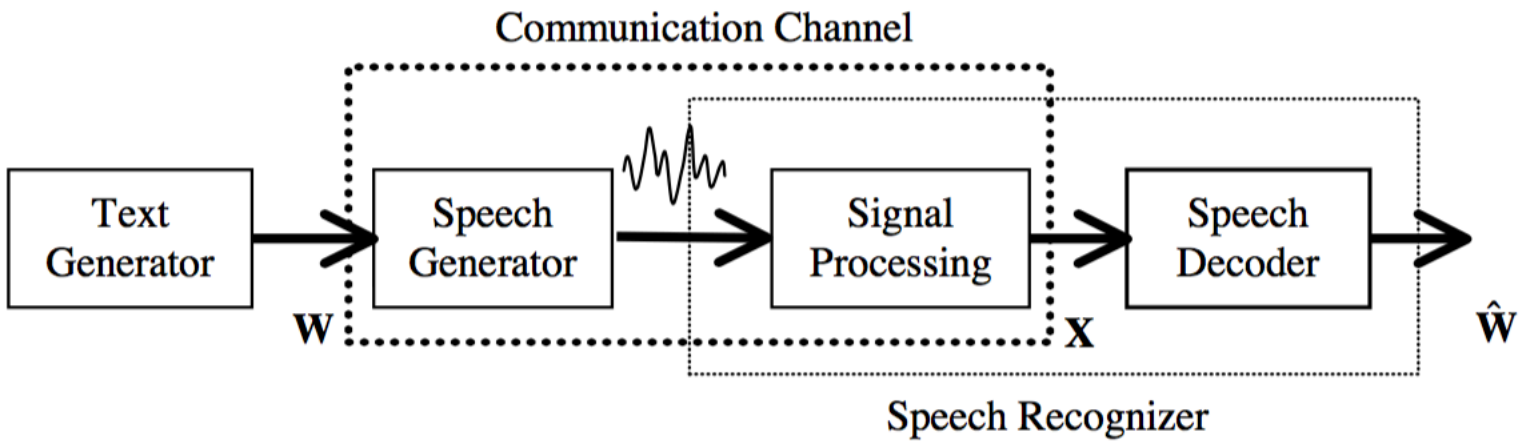


Figure 1.1 A source-channel model for a speech recognition system [15].

- the speaker's mind decides the source word sequence W that is delivered through his/her text generator.
- The source is passed through a noisy communication channel.
- The speech decoder aims to decode the acoustic signal X into a word sequence W^, which is hopefully close to the original word sequence W.

Intro

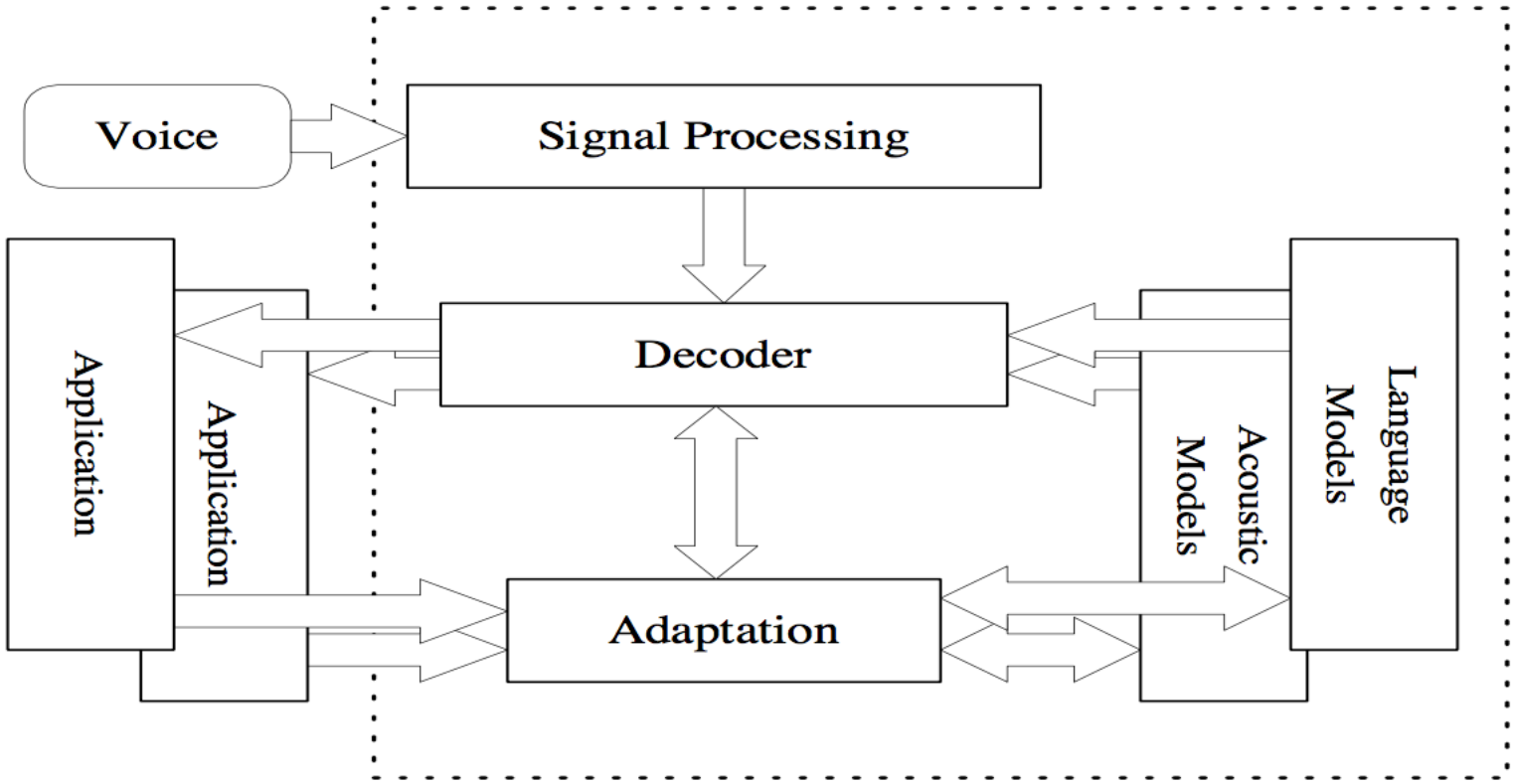
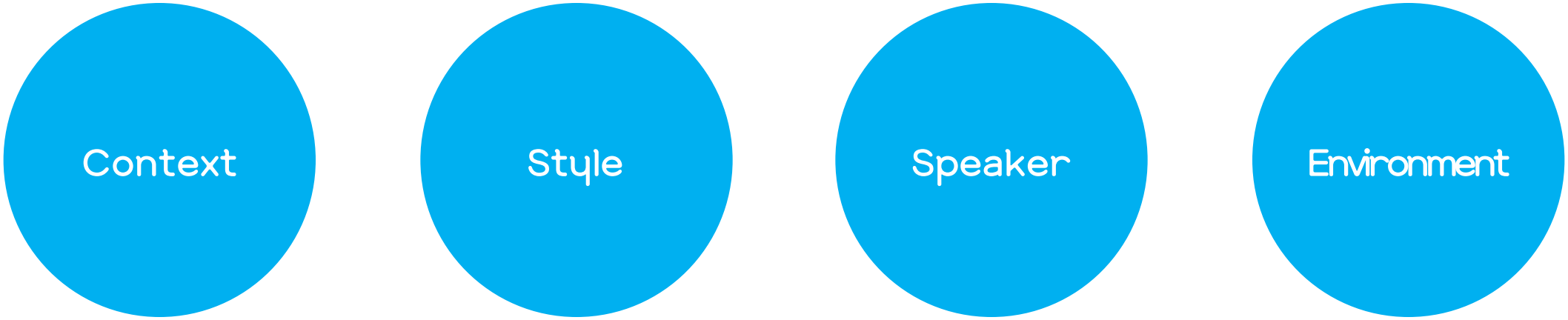


Figure 1.2 Basic system architecture of a speech recognition system [12].

Contents

Variability in the Speech Signal

- Although we can build a very accurate speech recognizer for a particular speaker, in a particular language and speaking style, in a particular environment, and limited to a particular task, it remains a research challenge to build a recognizer that can essentially understand anyone's speech, in any language, on any topic, in any free-flowing style, and in almost any speaking environment.



Variability in the Speech Signal

1) Context Variability

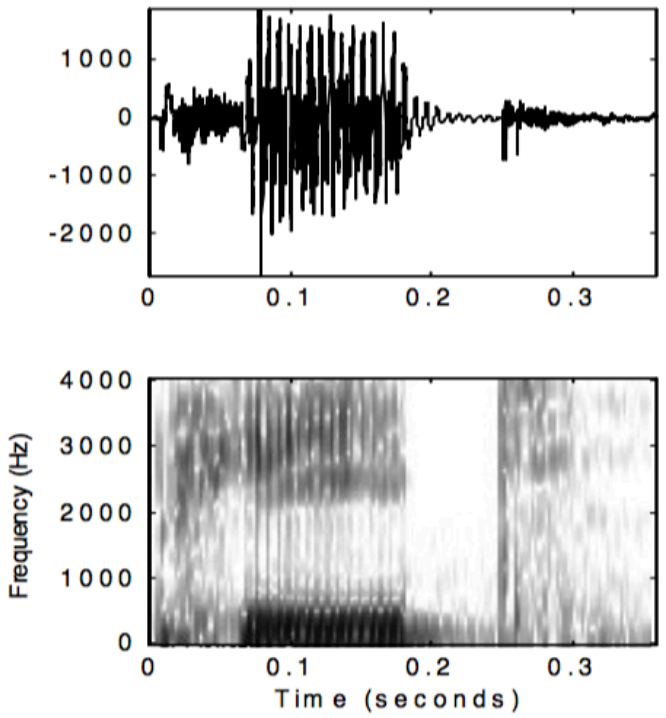
Mr. Wright should write to Ms. Wright right away about his Ford or four door Honda.

- Spoken language interaction between people requires knowledge of word meanings, communication context, and common sense.
- 발음은 같지만 그 의미는 다를 수 있음 : Wright, write, right
- 발음도 비슷한데 의미적 유사성이 있을 수도 있음. : Ford or, four door

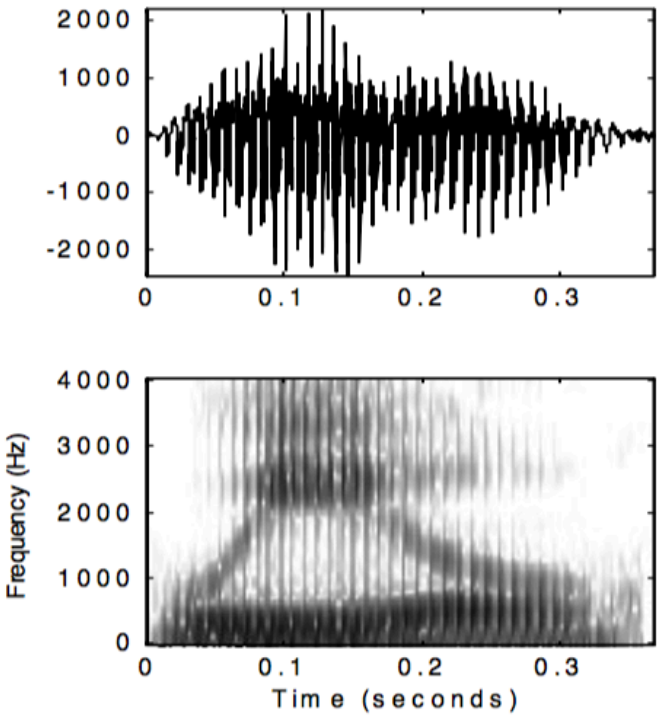
Variability in the Speech Signal

1) Context Variability

- the acoustic realization of phoneme /ee/ for word peat and wheel depends on its left and right context. The dependency becomes more important in fast speech or spontaneous speech conversation, since many phonemes are not fully realized.



〈 Peat 〉



〈 Wheel 〉

Variability in the Speech Signal

2) Style Variability

- we can have an isolated speech recognition system, in which users have to pause between each word.
- the pause provides a clear boundary for the word
  -> easily eliminate errors such as Ford or and Four Door.
- Isolated speech provides a correct silence context to each word so that it is easier to model and decode the speech, leading to a significant reduction in computational complexity and error rate.
  (In practice, WER of an isolated speech recognizer can typically be reduced by more than a factor of three (from 7% to 2%))
- The rate of speech also affects the word recognition rate. It is typical that the higher the speaking rate (words/minute), the higher the error rate. -> 진짜..?

Variability in the Speech Signal

3) Speaker Variability

- Every individual speaker is different.
  - vocal tract size
  - length and width of the neck
  - a range of physical characteristics
  - age
  - sex
  - dialect
  - health
  - education
  - and personal style
  - ...

Variability in the Speech Signal

3) Speaker Variability

- Speaker-independent : 화자 독립
  - use more than 500 speakers to build a combined model.
  - speakers with accents have a tangible error-rate increase of 2 to 3 times.
  - 성능을 향상시키기 위해 제약조건이 필요함 (등록된 화자당 30분 이상의 발화문)
- Speaker-dependent : 화자 종속
  - not only improved accuracy but also improved speed,
    -> since decoding can be more efficient with an accurate acoustic and phonetic model.
  - A typical speaker- dependent speech recognition system can reduce the word recognition error by more than 30%
- it is important to make use of both speaker-dependent and speaker-independent data using speaker-adaptive training techniques

Variability in the Speech Signal

4) Environment Variability

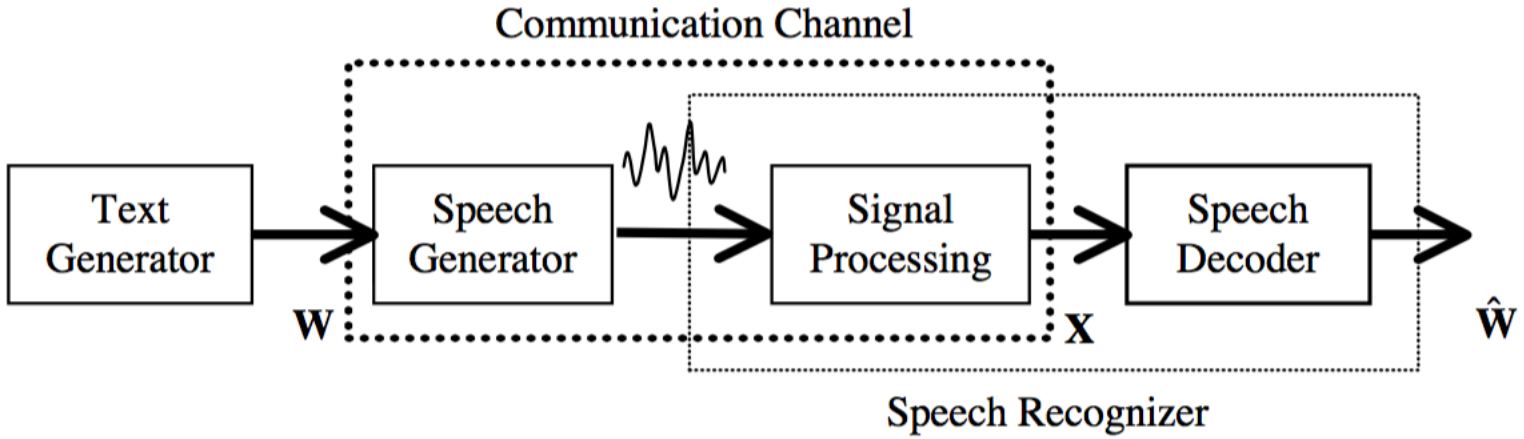


Figure 1.1 A source-channel model for a speech recognition system [15].

- Background noise : 문닫는 소리, 핸드폰 소리 등 환경 소음
  - External parameters, such as the characteristics of the environmental noise and the type and placement of the microphone, can greatly affect speech recognition system performance.
- noises made by speakers such as lip smacks and noncommunication words.
- Noise may also be present from the input device itself, such as the microphone and A/D interference noises.

How to Measure Speech Recognition Errors

- When you compare different acoustic modeling algorithms, it is important to compare their relative error reduction.
- need to have a test data set that contains more than 500 sentences (with 6 to 10 words for each sentence) from 5 to 10 different speakers to reliably estimate the recognition error rate.
- it is important to evaluate performance of a test set after you decide the optimal parameter setting.
- The test set should be completely new with respect to both training and parameter tuning.

How to Measure Speech Recognition Errors

- There are typically three types of word recognition errors in speech recognition:
  - **Substitution** : an incorrect word was substituted for the correct word
  - **Deletion** : a correct word was omitted in the recognized sentence
  - **Insertion** : an extra word was added in the recognized sentence

Correct: *Did mob mission area of the Copeland ever go to m4 in nineteen eighty one*
Recognized: *Did mob mission area ** the* ***copy*** *land ever go to m4 in nineteen* ***east*** *one*

How to Measure Speech Recognition Errors

$$\text{Word Error Rate} = 100\% \times \frac{Subs + Dels + Ins}{\text{No. of word in the correct sentence}}$$

- To determine the minimum error rate, you can't simply compare two word sequences one by one.
  - Ex) *The effect is clear* recognized as *Effect is not clear*.
  - If you compare word to word, the error rate is 75% (The vs. Effect, effect vs. is, is vs. not).
    In fact, the error rate is only 50% with one deletion (The) and one insertion (not).
- you need to align a recognized word string against the correct word string and compute the number of substitutions (Subs), deletions (Dels), and insertions (Ins).
- This alignment is also known as the maximum substring matching problem, which can be easily handled by the dynamic programming algorithm discussed in Chapter 8.

How to Measure Speech Recognition Errors

- R[i, j] : the minimum error of aligning substring w1w2 …wn against substring w^1w^2 …w^m
- The optimal alignment and the associated word error rate R[n,m] for correct word string w1w2Lwn and the recognized word string w^1w^2 Lw^m are obtained via the dynamic programming algorithm

ALGORITHM 9.1: THE ALGORITHM TO MEASURE THE WORD ERROR RATE

Step 1: *Initialization* $R[0,0] = 0$ $R[i,j] = \infty$ if $(i < 0)$ or $(j < 0)$ $B[0,0] = 0$

Step 2: *Iteration*

for $i = 1,\ldots,n$ {

for $j = 1,\ldots,m$ {

$$R[i,j] = \min \begin{bmatrix} R[i-1,j]+1 \text{ (deletion)} \\ R[i-1,j-1] \text{ (match)} \\ R[i-1,j-1]+1 \text{ (substitution)} \\ R[i,j-1]+1 \text{ (insertion)} \end{bmatrix}$$

$$B[i,j] = \begin{cases} 1 & \text{if deletion} \\ 2 & \text{if insertion} \\ 3 & \text{if match} \\ 4 & \text{if substitution} \end{cases} \quad \} \}$$

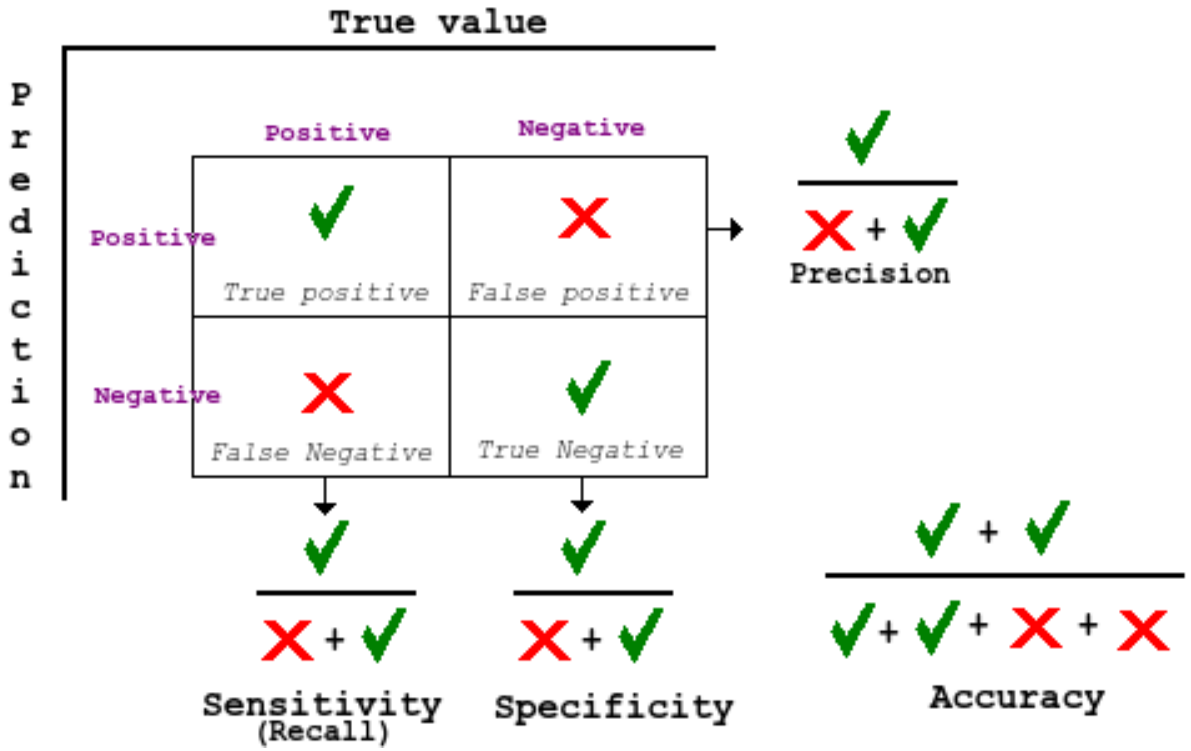
Step 3: *Backtracking and termination*

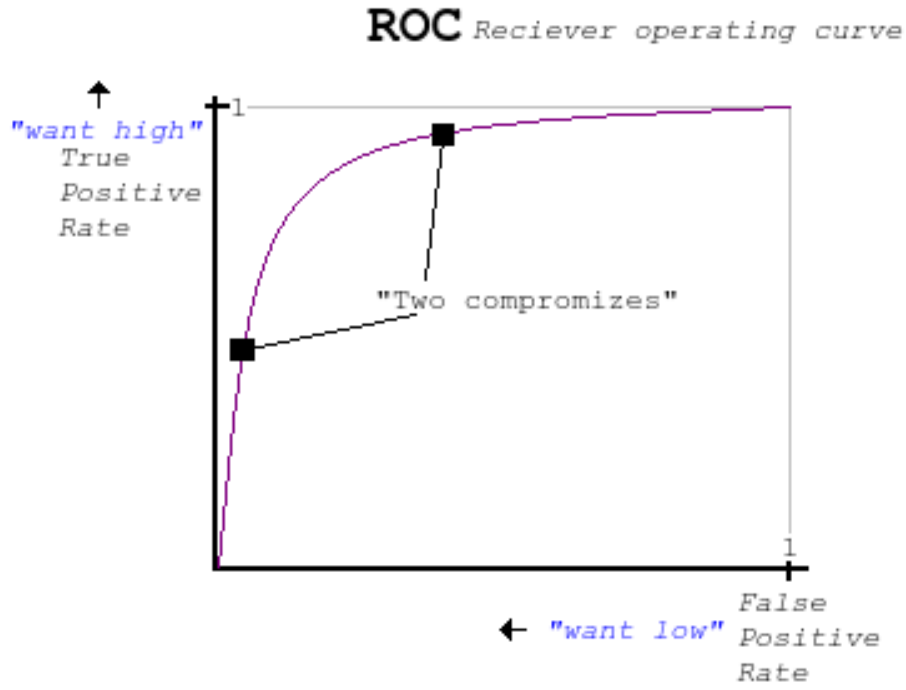
$$\text{word error rate} = 100\% \times \frac{R(n,m)}{n}$$

optimal backward path = $(s_1, s_2, \ldots, 0)$

where $s_1 = B[n,m]$, $s_t = \begin{bmatrix} B[i-1,j] \text{ if } s_{t-1} = 1 \\ B[i,j-1] \text{ if } s_{t-1} = 2 \\ B[i-1,j-1] \text{ if } s_{t-1} = 3 \text{ or } 4 \end{bmatrix}$ for $t = 2,\ldots$ until $s_t = 0$

How to Measure Speech Recognition Errors





- For applications involved with rejection, such as word confidence measures as discussed in Section 9.7, you need to measure both false rejection rate and false acceptance rate.
- In speaker or command verification, the false acceptance of a valid user/command is also referred to as Type I error, as opposed to the false rejection of a valid user/command (Type II).

Signal Processing - Extracting Features

1) Signal Acuisition

- To perform speech recognition, a number of components
  - such as digitizing speech,
  - feature extraction and transformation,
  - acoustic matching,
  - language model-based search
- Those are can be pipelined time-synchronously from left to right.

Signal Processing - Extracting Features

1) Signal Acuisition

- Most operating systems can supply mechanisms for organizing pipelined programs in a multitasking environment.
- Most operating systems can supply mechanisms for organizing pipelined programs in a multitasking environment. Buffers must be appropriately allocated so that you can ensure time-synchronous processing of each component.
- Slow Machine -> potential delays in processing an individual component -> Large buffers
- The right buffer size can be easily determined by experimentally tuning the system with different machine load situations to find a balance between resource use and relative delay.

Signal Processing - Extracting Features

1) Signal Acuisition

Table 9.1 Relative error rate reduction with different sampling rates.

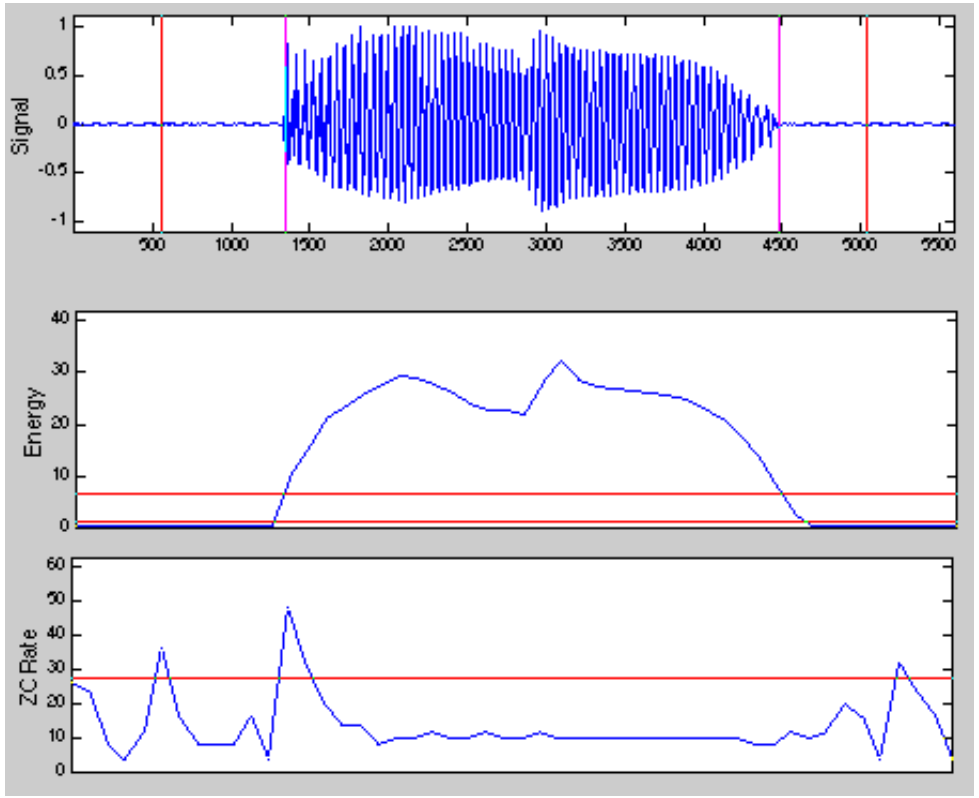
Sampling Rate	Relative Error-Rate Reduction
8 kHz	Baseline
11 kHz	+10%
16 kHz	+10%
22 kHz	+0%

- buffer typically ranges from 4 to 64 kB with 16kHz speech sampling rate and 16-bit A/D precision.
- In practice, 16kHz sampling rate is sufficient for the speech bandwidth (8 kHz).
- Reduced bandwidth, such as telephone channel, generally increases speech recognition error rate.
- most of the salient speech features are within 8kHz bandwidth.

Signal Processing - Extracting Features

2) End-Point Detection (EPD)

- To activate speech signal capture, you can use a number of modes including either push to talk or continuously listening.
- The push-to-talk mode uses a special push event to activate or deactivate speech capturing, which is immune to the potential background noise and can eliminate unnecessary use of processing resources to detect speech events.
- The continuously listening model listens all the time and automatically detects whether there is a speech signal or not.
- often based on the energy threshold that is a function of time.



Signal Processing - Extracting Features

2) End-Point Detection (EPD)

- It is not critical for the automatic end-point detector to offer exact end-point accuracy. The key feature required of it is a low rejection rate (i.e., the automatic end-point detector should not interpret speech segments as silence/noise segments).
- false rejection : leads to an error in the speech recognizer.
- false acceptance (i.e., the automatic end-point detector interprets noise segments as speech segments) : may be rescued by the speech recognizer later if the recognizer has appropriate noise models, such as specific models for clicks, lip smacks, and background noise.

Signal Processing - Extracting Features

2) End-Point Detection (EPD)

- Explicit end-point detectors work reasonably well with **recordings exhibiting a signal-to-noise ratio of 30 dB or greater**, but they fail considerably on noisier speech.
- speech recognizers can be used to determine the end points by aligning the vocabulary words preceded and followed by a silence/noise model. This scheme is generally much more reliable than any threshold-based explicit end-point detection, because recognition can jointly detect both the end points and words or other explicit noise classes, but requires more computational resources.
- A compromise is to use a simple adaptive two-class (speech vs. silence/noise) classifier to locate speech activities (with enough buffers at both ends) and notify the speech recognizer for subsequent processing.

Signal Processing - Extracting Features

2) End-Point Detection (EPD)

- For the two-class classifier, we can use both the log-energy and delta log-energy as the feature. Two Gaussian density functions, $\{\Phi 1, \Phi 2\} = \Phi$, can be used to model the background stationary noise and speech, respectively.
- enough frames before the *tb*, for the speech recognizer to minimize the possible detection error.
- enough noise/silence frames are detected at *te*, we should keep providing the speech recognizer with enough frames for processing before declaring that the end of the utterance has been reached.

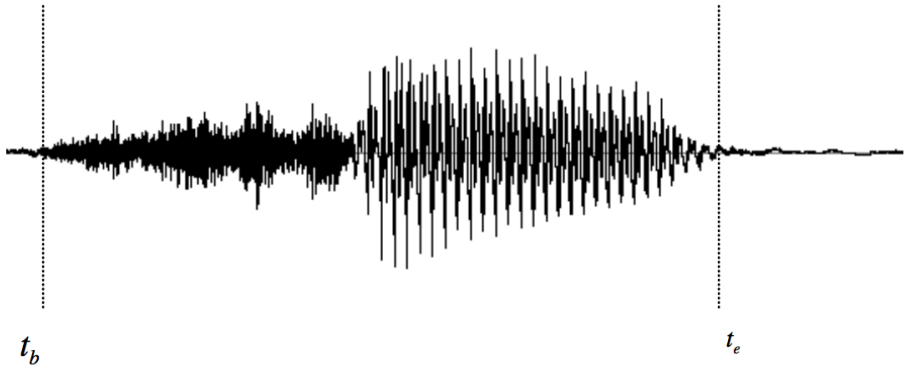


Figure 9.2 End-point detection boundary t_b and t_e may need extra buffering for subsequent speech recognition.

Signal Processing - Extracting Features

2) End-Point Detection (EPD)

- these parameters can be dynamically adapted using the EM algorithm during runtime.
- EM algorithm can iteratively estimate the Gaussian parameters without having a precise segmentation between speech and noise segments.
- This is very important, because we need to keep the parameters dynamic for robust end-point detection in constantly changing environments.

Signal Processing - Extracting Features

2) End-Point Detection (EPD)

- we use an exponential window to give weight to the most recent signal:

$$w_k = \exp(-\alpha k)$$

- α : constant that controls the adaptation rate
- K : the index of the time.
- In fact, you could use different rates for noise and speech when you use the EM algorithm to estimate the two-class Gaussian parameters.
- It is advantageous to use a smaller time constant for noise than for speech.

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=-\infty}^{t} w_i \frac{c_k P(\mathbf{x}_i \mid \Phi_k)\mathbf{x}_i}{\sum_{k=1}^{2} P(\mathbf{x}_i \mid \Phi_k)}}{\sum_{i=-\infty}^{t} w_i \frac{c_k P(\mathbf{x}_i \mid \Phi_k)}{\sum_{k=1}^{2} P(\mathbf{x}_i \mid \Phi_k)}}, k \in \{0,1\}$$

Phonetic Modeling - Selecting Appropriate Units

- phonetic system은 특정언어에 한정되어 있음
- general-purpose large-vocabulary speech recognition에서 whole-word models이 안되는 이유:
  - Every new task contains novel words(신조어) without any available training data, such as proper nouns(고유명사) and newly invented jargons(특수용어/은어).
  - 너무 많은 단어들이 있고, 각각의 다른 단어들은 다른 형태의 acoustic realizations을 함. -> context-dependent word models을 만들기 위한 단어의 반복이 충분히 없을 확률이 높음.

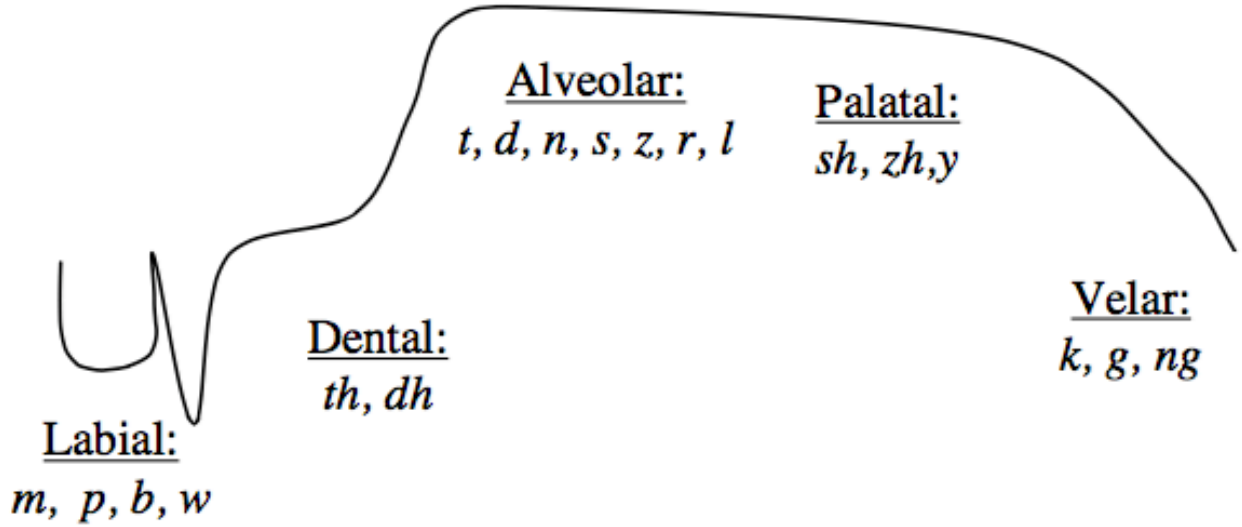


Figure 2.21 The major places of consonant articulation with respect to human mouth.

Phonetic Modeling - Selecting Appropriate Units

- How to select the most basic units to represent salient acoustic and phonetic information for the language is an important issue in designing a workable system. At a high level, there are a number of issues we must consider in choosing appropriate modeling units.
  - The unit should be **accurate**, to represent the acoustic realization that appears in different contexts.
  - The unit should be **trainable**. We should have enough data to estimate the parameters of the unit. Although words are accurate and representative, they are the least trainable choice in building a working system, since it is nearly impossible to get several hundred repetitions for all the words, unless we are using a speech recognizer that is domain specific, such as a recognizer designed for digits only.
  - The unit should be **generalizable**, so that any new word can be derived from a predefined unit inventory for task-independent speech recognition. If we have a fixed set of word models, there is no obvious way for us to derive the new word model.

Phonetic Modeling - Selecting Appropriate Units

1) Comparison of Different Units

- English : typically considered as a principal carrier of meaning and are seen as the smallest unit that is capable of independent use. -> 음성인식에 whole-word model 이 주로 사용됨.
- word models의 장점 : 단어들 사이에 가지고 있는 coarticulation(동시조음)을 포착 할 수 있음.
  Coarticulation : 조음기관들을 겹치게해서 변화를 쉽게 해서 단어 발화의 시간을 줄여줌.
  - Ex) Soon (/Sun/) : '쓰운'이라고 발음하지 않고, 미리 입을 오무려서(rounded) '쑨'이라고 발음함.]
- When the vocabulary is small, we can create word models that are context dependent.
- for small vocabulary recognition -> both **accurate and trainable**, and there is **no need to be generalizable** -> whole-word models are widely used.

Phonetic Modeling - Selecting Appropriate Units

2) Context Dependency

- If we make units context dependent, we can significantly improve the recognition accuracy, provided there are enough training data to estimate these context-dependent parameters.
- Context-dependent phonemes have been widely used for large-vocabulary speech recognition, thanks to its significantly improved accuracy and trainability.
  (A context usually refers to the immediately left and/or right neighboring phones)
- triphone model : 인접한 좌/우 음소를 고려하는 phonetic model. 만일 두개의 음소가 같더라도 하나가 다르면 그건 다른 triphones.
  (We call different realizations of a phoneme allophones. Triphones are an example of allophones.)
- Modeling interword context-dependent phones is complicated.
  - Ex) speech (/s p iy ch/) : both left and right contexts for /p/ and /iy/ are known, while the left context for /s/ and the right context for /ch/ are dependent on the preceding and following words in actual sentences.
  - The juncture effect on word boundaries is one of the most serious coarticulation phenomena in continuous speech. (a / the)
- Even with the same left and right context identities, there may be significantly different realizations for a phone at different word positions (the beginning, middle, or end of a word)
  - Ex) /t/ in that rock is almost extinct, while the phone /t/ in the middle of theatrical sounds like /ch/.
    -> This implies that different word positions have effects on the realization of the same triphone.

Phonetic Modeling - Selecting Appropriate Units

2) Context Dependency

- stress also plays an important role in the realization of a particular phone.
- Stressed vowels tend to have longer duration, higher pitch, and more intensity, while unstressed vowels appear to move toward a neutral, central schwa-like phoneme.
  (cf. schwa : about의 a나 moment의 e처럼, 한 단어에서 강세가 주어지지 않는 모음)
- stress can be used as a unique feature to distinguish a set of word pairs
  - ex) import vs. import, and export vs. export.
- /t/ in word Italy vs Italian is pronounced differently in American English due the location of the stress. (triphone context는 동일하더라도..)

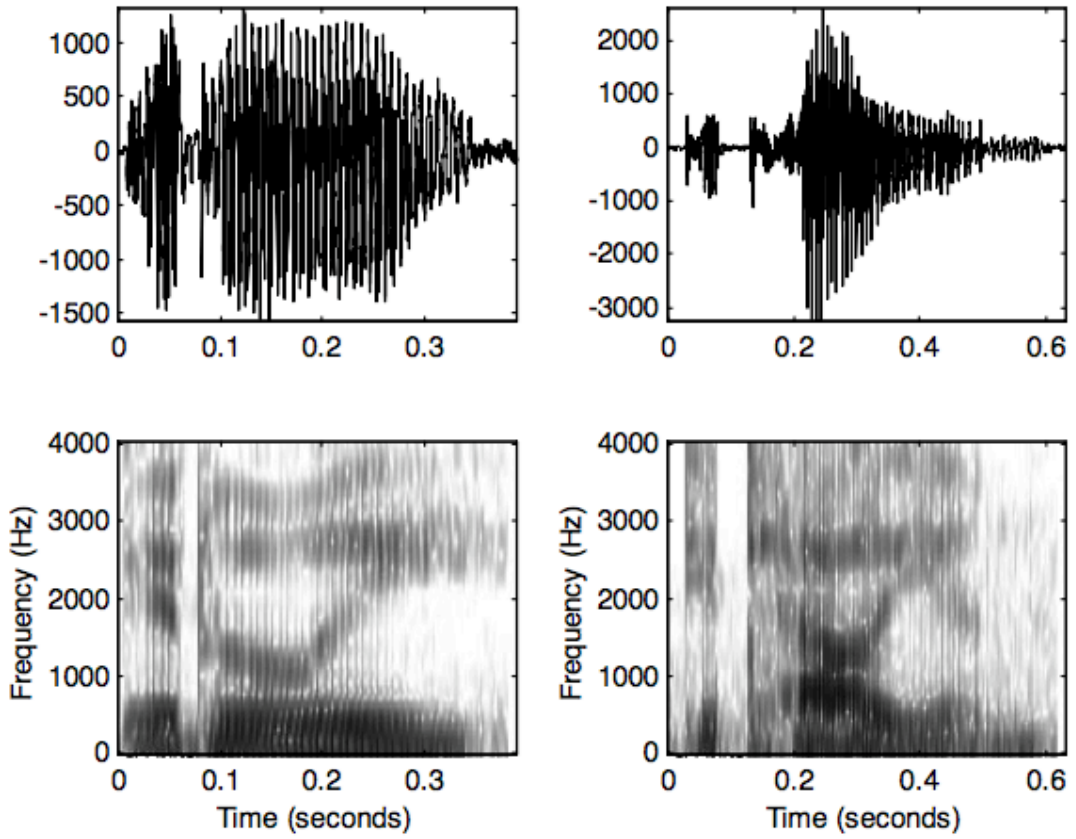


Figure 9.3 The importance of stress is illustrated in *Italy* vs. *Italian* for phone /t/.

Phonetic Modeling - Selecting Appropriate Units

3) Clustered Acoustic-Phonetic Units

- Triphone modeling assumes that every triphone context is different.
- Actually, many phones have similar effects on the neighboring phones.
- The position of our articulators(조음기관) has an important effect on how we pronounce neighboring vowels.
- both phonetic and subphonetic units have the same benefits, as they share parameters at unit level.
  -> This is the <u>key benefit</u> in comparison to the word units.

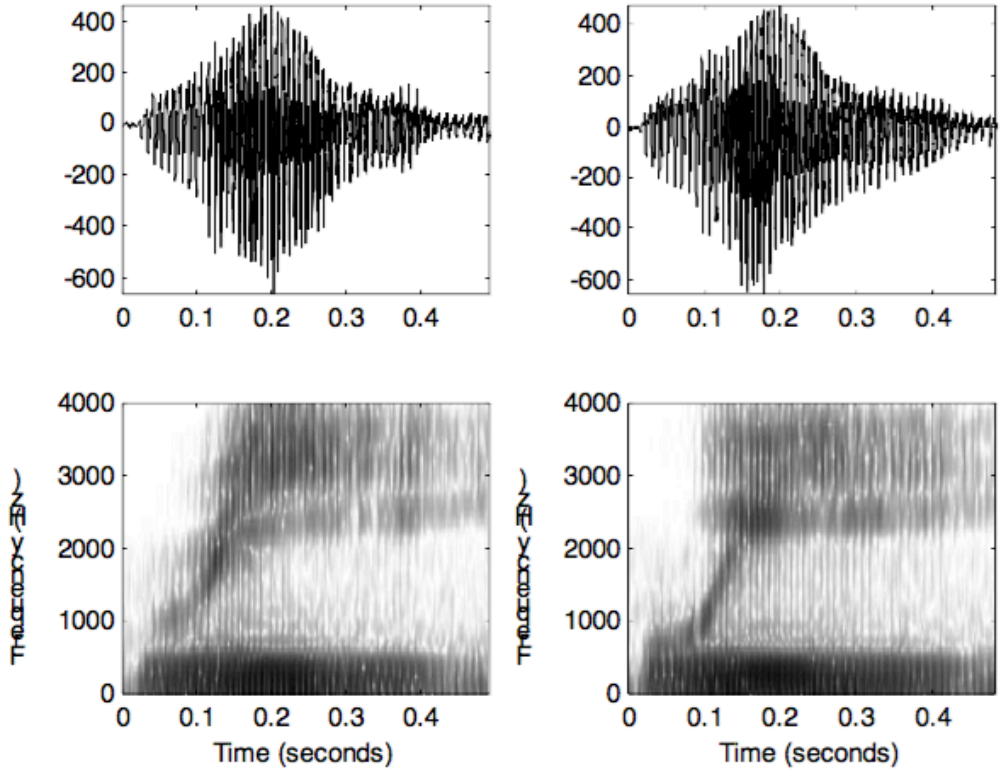


Figure 9.4 The spectrograms for the phoneme /iy/ with two different *left-contexts* are illustrated. Note that /r/ and /w/ have similar effects on /iy/. This illustrates that different left-contexts may have similar effects on a phone.

Phonetic Modeling - Selecting Appropriate Units

3) Lexical Baseforms

- When appropriate subword units are used, we must have the correct pronunciation for each word so that concatenation of the subword unit can accurately represent the word to be recognized.
- The dictionary represents the standard pronunciation used as a starting point for building a workable speech recognition system.
- We also need to provide alternative pronunciations to words such as tomato that may have very different pronunciations.
- we must also use phonologic rules to modify interword pronunciations or to have reduced sounds.
- Assimilation(동화 ; 음소가 이웃하는 음성적 특성을 띠게됨) is a typical coarticulation(동시조음 ; 발화중 조음기관에 일어나는 중복 ex. would you) phenomenon—a change in a segment to make it more like a neighboring segment.

Phonetic Modeling - Selecting Appropriate Units

3) Lexical Baseforms

- In continuous speech recognition, we must also use phonologic rules to modify inter-word pronunciations or to have reduced sounds. Assimilation is a typical coarticulation phenomenon—a change in a segment to make it more like a neighboring segment.
  - did you /d ih jh y ah/
  - set you /s eh ch er/
  - last year / l ae s ch iy r/
  - because you've /b iy k ah zh uw v/
- Deletion is also common in continuous speech.
  - /t/ and /d/ are often deleted before a consonant.

Phonetic Modeling - Selecting Appropriate Units

3) Lexical Baseforms

- We can use a probabilistic finite state machine to model each word's pronunciation variations.
- The probability with each arc indicates how likely that path is to be taken, with all the arcs that leave a node summing to 1. As with HMMs, these weights can be estimated from real corpus for improved speech recognition. In practice, the relative error reduction of using probabilistic finite state machines is very modest (5–10%).

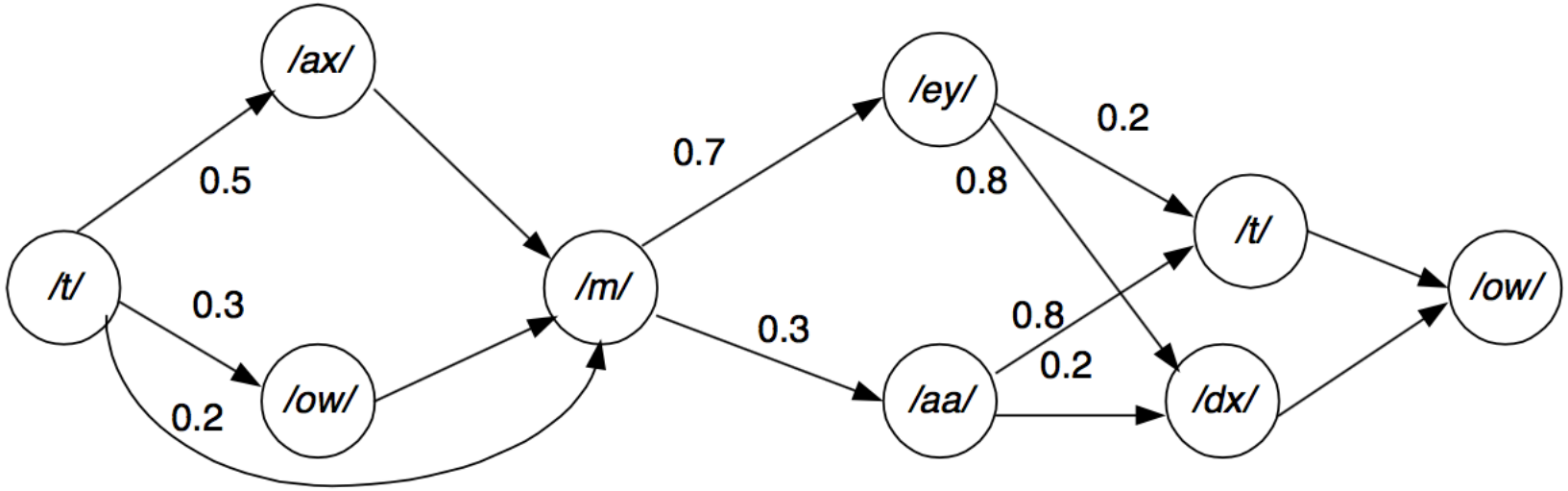


Figure 9.7 A possible pronunciation network for word *tomato*. The vowel /*ey*/ is more likely to flap, thereby having a higher transition probability into /*dx*/.

Acoustic Modeling - Scoring Acoustic Features

- After feature extraction, we have a sequence of feature vectors, X, such as the MFCC vector, as our input data.
- We need to estimate the probability of these acoustic features, given the word or phonetic model, W, so that we can recognize the input data for the correct word.
- This probability is referred to as acoustic probability, $P(X \mid W)$.

Acoustic Modeling - Scoring Acoustic Features

1) Choice of HMM Output Distribution

- continuous HMM :
  - with a large number of mixtures offers the best recognition accuracy, although its computational complexity also increases linearly with the number of mixtures.
- Discrete HMM :
  - is computationally efficient, but has the worst performance among the three models.
- Semicontinuous HMM :
  - provides a viable alternative between system robustness and trainability.

- discrete or semicontinuous HMM : it is helpful to use multiple codebooks for a number of features for significantly improved performance. Each codebook then represents a set of different speech parameters.

Acoustic Modeling - Scoring Acoustic Features

2) Isolated vs Continuous Speech Training

- It is not necessary to have precise end-point detection, because the silence model automatically determines the boundary if we concatenate silence models with the word model in both ends.
- If subword units, such as phonetic models, are used, we need to share them across different words for large-vocabulary speech recognition. These subword units are concatenated to form a word model, possibly adding silence models at the beginning and end.

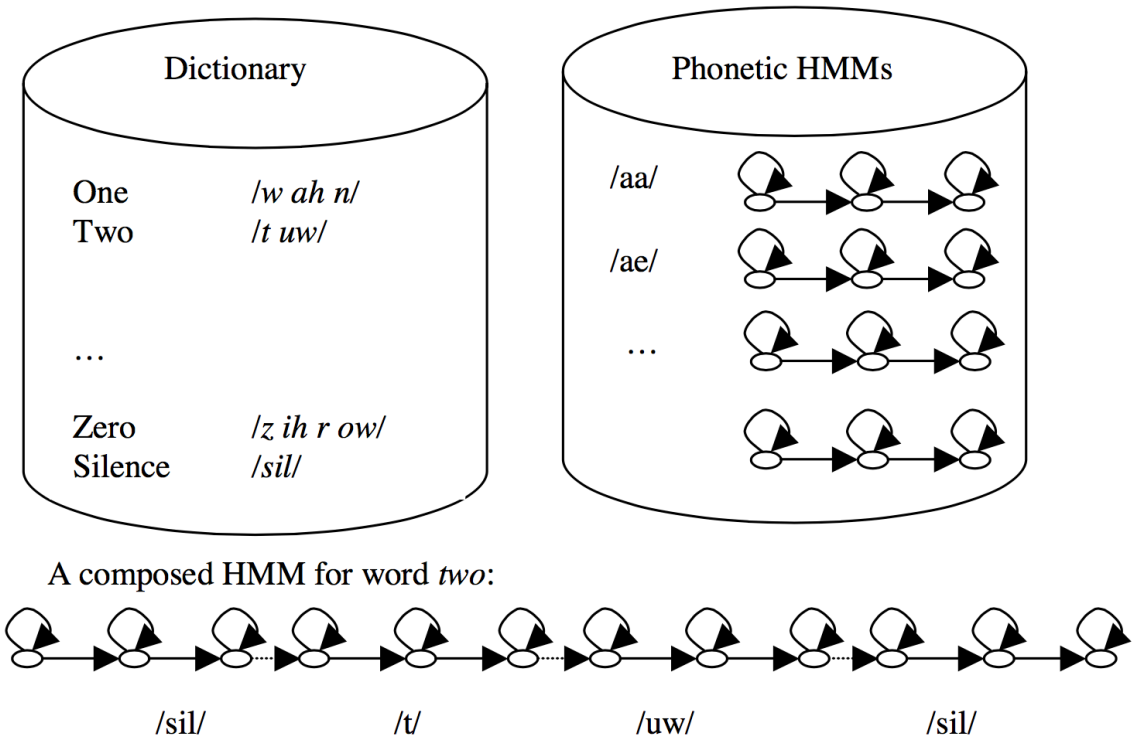


Figure 9.9 The construction of an isolated word model by concatenating multiple phonetic models based on the pronunciation dictionary.

Acoustic Modeling - Scoring Acoustic Features

2) Isolated vs Continuous Speech Training

- The ability to automatically align each individual HMM to the corresponding unsegmented speech observation sequence is one of the most powerful features in the forward-backward algorithm.
- When the HMM concatenation method is used for continuous speech, you need to compose multiple words to form a sentence HMM based on the transcription of the utterance.
- In the same manner, the forward-backward algorithm absorbs a range of possible word boundary information of models automatically.
- No need to have a precise segmentation of the continuous speech.
- If there is a need to modify interword pronunciations due to interword pronunciation change
  - such as want you, you can add a different optional phonetic sequence for t-y in the concatenated sentence HMM.

Adaptive Techniques - Minimizing Mismatches

- The mismatch between the model and operating conditions always exists.
- One of the most important factors in making a speech system usable is to minimize the possible mismatch dynamically with a small amount of calibration data.
- **Adaptive techniques** can be used to modify system parameters to better match **variations in microphone, transmission channel, environment noise, speaker, style, and application contexts.**
- As a concrete example, speaker-dependent systems can provide a significant word error-rate reduction in comparison to speaker-independent systems if a large amount of speaker-dependent training data exists

Adaptive Techniques - Minimizing Mismatches

- Since the use of recognition results may be imperfect, there is a possibility of divergence if the recognition error rate is high.
- If the error rate is low, the adaptation results may still not be as good as supervised adaptation in which the correct transcription is provided for the user to read, a process referred to as the enrollment process.
- In this process you can check a wide range of parameters as follows:
  - Check the back ground noise by asking the user not to speak.
  - Adjust the microphone gain by asking the user to speak normally.
  - Adapt the acoustic parameters by asking the user to read several sentences.
  - Change the decoder parameters for the best speed with no loss of accuracy.
  - Compose dynamically new enrollment sentences based on the user-specific error patterns.
- Adaptation : Maximum a Posteriori (MAP) / Maximum Likelihood Linear Regression (MLLR)
  - 참고 : http://darkpgmr.tistory.com/62