CHAPTER 3

Formant-Level Assessment

3.1 Formant and Formant Frequency

Since the glottal wave is periodic, consisting of fundamental frequency and a number of harmonics, it can be analyzed as a num of sine waves. The resonances of the vocal tract are excited by the glottal energy. For simplicity, we regard the vocal tract as a straight tube of uniform cross-sectional area, closed at the glottal end, open at the lips. When the shape of the vocal tract changes, the resonances change also. Harmonics near the resonances are emphasized, and, the resonances of the cavities that are typical of particular articulator configurations are called formant. Formant Frequencies are the resonances in the vocal tract, and they convey the differences between different sounds. Expert spectrogram readers are able to recognize speech by looking at a spectrogram, particularly at the formants. It has been argued that they are very useful features for speech recognition, but they haven't been widely used because of the difficultly in estimating them. In our System, to estimate the formant frequency, we use ESPS software, which is a

well-know tool developed by Entropic Research Laboratory in 1997.

3.2 Relation between Articulation and Formant

From the articulatory's point of view, vowels can be described with three articulatory elements：

- The aperture of the mouth opening
- The position of the tongue
- The amount of the lip rounding

Formant frequency is known to be related with these three elements. Specifically, the first formant frequency has relation with the aperture of mouth rounding, the second formant frequency with the position of the tongue, and the higher formants with the amount of lip rounding.

Figure 3.1 shows the articulatory position of each phoneme. Briefly speaking, a longer cavity of the oral portion leads to lower F1, while a lower F2 is determined by the bigger size of the forward portion.

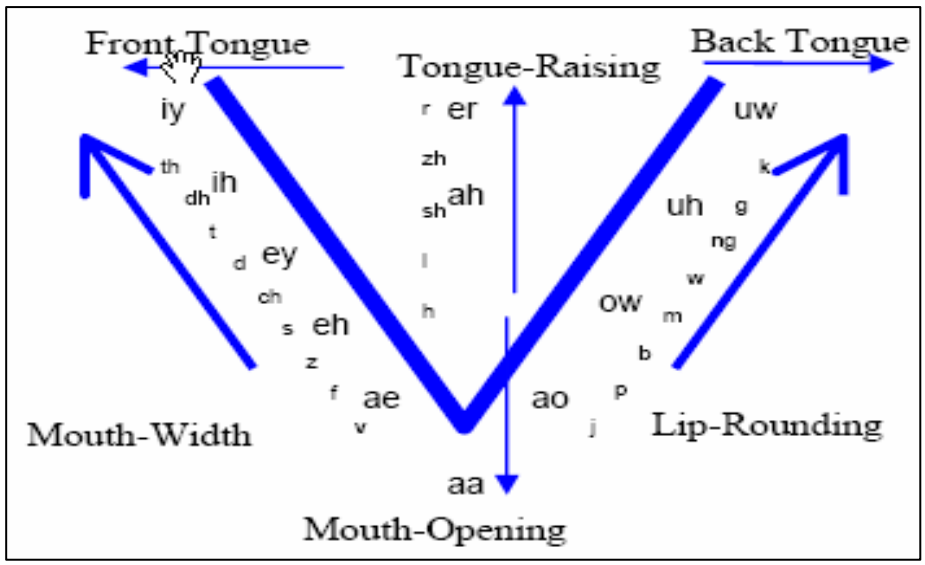


Figure 3.1 Mouth / lip shapes and tongue positions when pronounces the different vowels and consonants.

3.3 Formant Normalization

We use the ESPS software to extract formant coefficients. However, formant frequency depends on not only the articulator of the speaker but also his /her native language and sometimes different speakers' formant structure [16] may overlap. Some studies shows that the relative position of vowels is always kept the same between persons. Hence we need to normalize the formant frequencies before further processing. In particular, all the formant coefficients of a given speaker are normalized to the range between 0 and 1, as shown in Figure 3.2. For instance, the normalized F1 of the phone model "er" of a speaker can be calculated as :

$$normailze(F1) = \frac{F1_{er} - Min_{F1}}{Max_{F1} - Min_{F1}} \quad (3\text{-}1)$$

Before using the system, a test speaker is requested to pronounce several sentences embedded with five vowels, including "aa", "eh", "iy", "ow" and "uw". Then the maximum and minimum of F1 and F2 can then be found and applied in the normalized formant coefficients.

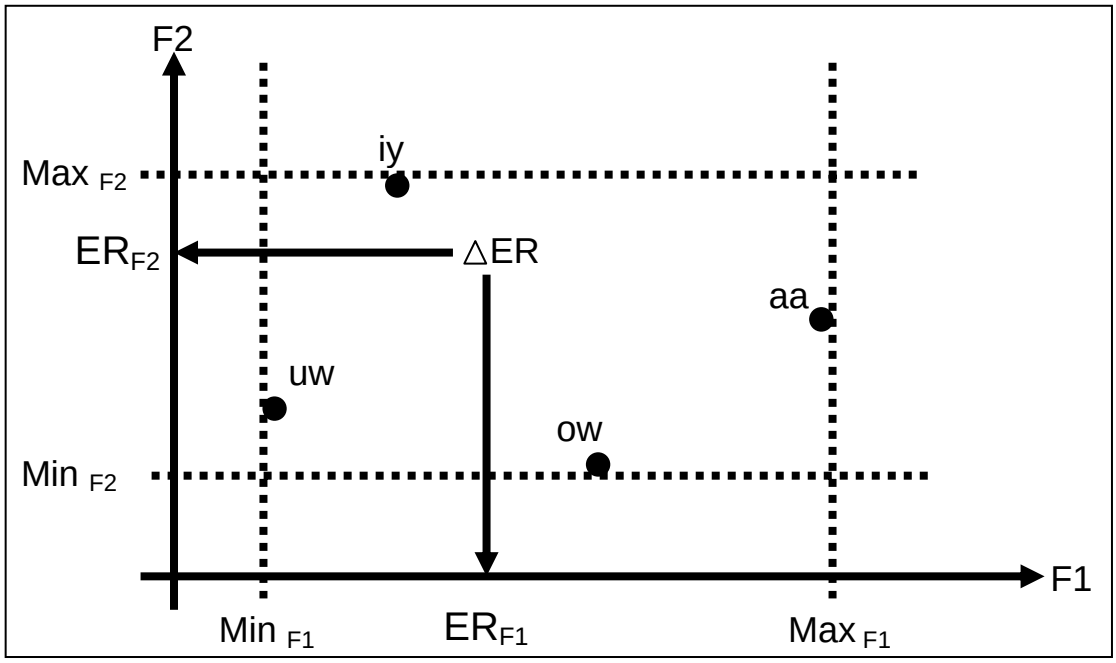


Figure 3.2 Speaker-dependent normalization for F1 and F2

3.4 Formant-Based HMM

MFCC-based HMM is widely used for speech recognition and segmentation. However, MFCC may not be able to embed the information of formants due to the feature reduction process of triangular filter bank. Figure 3.3 is the result of forced alignment using MFCC-based HMM. As we can see, it does not generate the correct boundaries. At the begging of four frames, the contour of F1 drastically changes and the contour of formant frequency of vowels are known to be smooth in the phonology. To improve the segmentation performance of vowels, we use the formant coefficients as a new feature set in the acoustic models training. This is achieved by appending a set of the formant coefficients to the original MFCC feature vector, including the original formant, the delta of the formant and double delta of the formant. In our implementation, the formant coefficients (plus their

delta and delta-delta versions) are added as a new stream in the HMM training using HTK tool. The stream configuration is shown in Figure3.4.

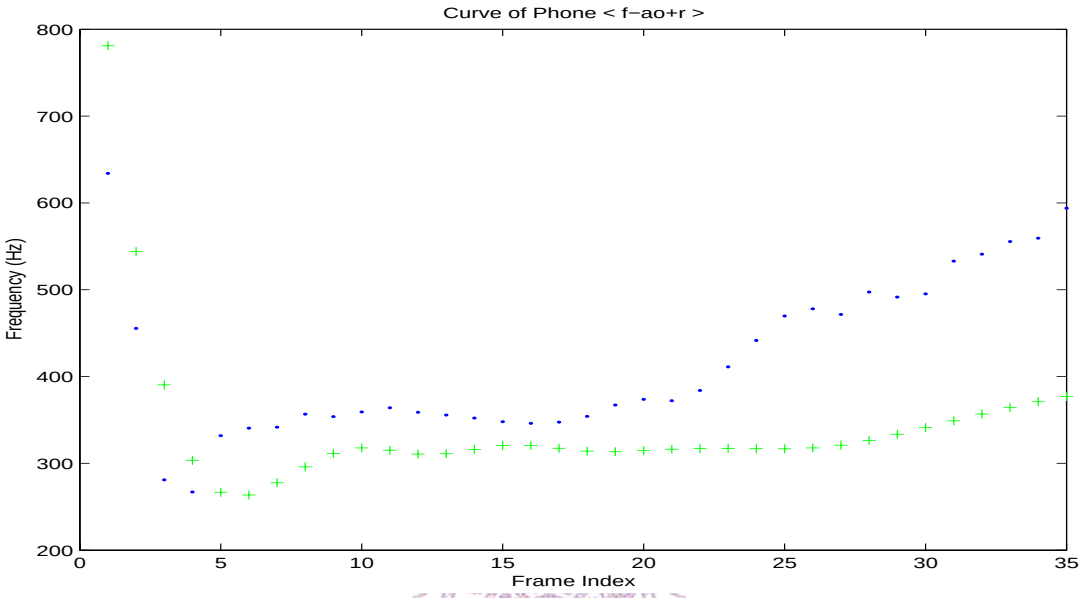


Figure 3.3 The formant contour of the phone model “f-ao+r” uttered by different native speakers. x-axis represents the frame index, while y-axis represents the first formant frequency

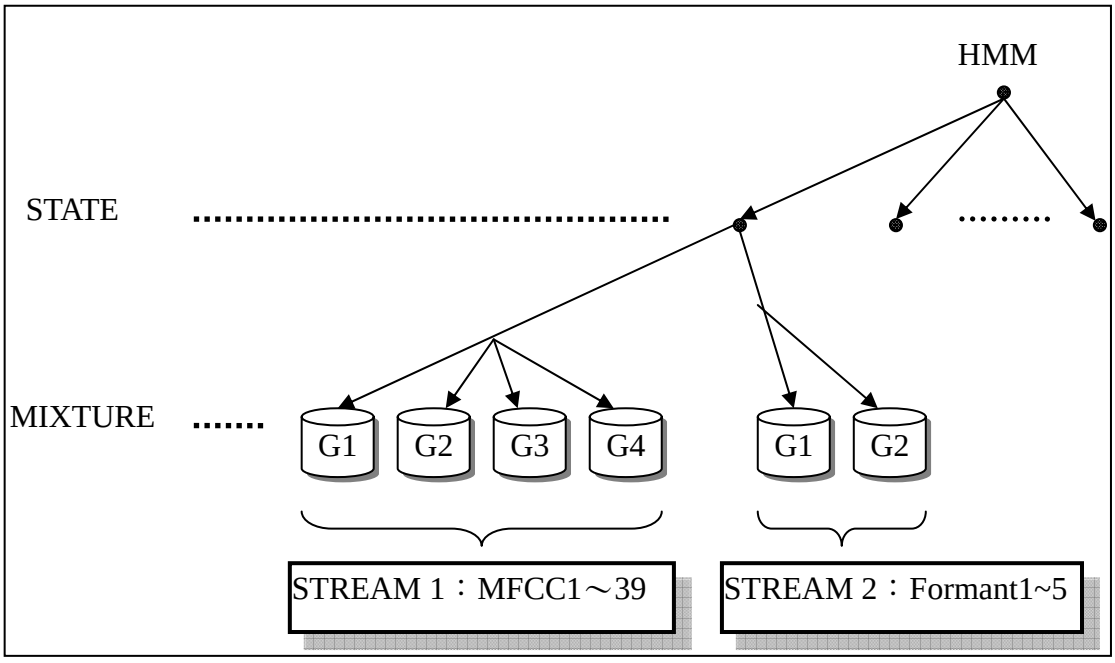
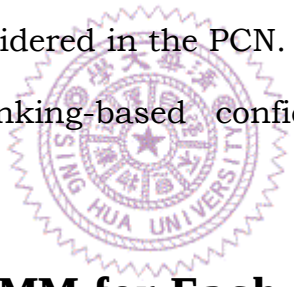


Figure 3.4 Stream configuration of Formant-Based HMM

3.5 Formant-Level Assessment

The PCN approach can predict the common pronunciation errors that might occur. However, for some errors which are not expected in the knowledge base, such like the low unknown quality of pronunciation of the word. For example, if a error is expected in one of the PCN path, there are two possibility the system detect the error pronunciation, first, the quality of pronunciation is less than the standard one, second, the pronunciation of the phone is not the expected one indeed. However, in the second situation, the recognizer may fail to find this kind of errors, because the unknown error pronunciation is not considered in the PCN. Considering this problem, a refinement called ranking-based confident measure (RCM) is suggested here.



3.5.1 Derived a GMM for Each Phone Models

As shown in Figure 3.3, a formant-based HMM can be used to better align the phoneme boundaries of a context dependent biphone model (CDBM). The contour of formants in the segment would be an important piece of information to assess the pronunciation. To this end, the first step of RCM is to derive a Gaussian probability density function of two variables for F1 and F2 based on maximum likelihood estimate for each CDBM. The flowchart of this step is shown in Figure3.5.

3.5.2 Ranking process

In the ranking process, we define a similar group for a CDBM or

CDTM (Context-Dependent Triphone Model). Table3.1 shows a example for similar groups definition.

Model Type	Example	Similar group
Mono-phone	ao	*
Bi-phone	ao+r	*+r
Tri-phone	f-ao+r	f-*+r

Table 3.1 Similar group definition (* means wildcard)

"*+r" means any phone model in the HMM mode set with model "r" on its right hand side. In CMU dictionary, we have 39 different monophone models, the number of context-dependent models in group "*+r" is less than or equal to 40, because some models of type "*+r" may not appear in the corpus.

For a formant contour in the phoneme segmentation, we calculate the GMM log probabilities of all CDBM (or CDTM) in the similar group of the phoneme. Then sorting these probabilities in descending order, finally we can determine the rank of this phoneme in the similar group.

3.5.3 Rank-Based Confidence Measure

In this section, we will discuss how to measure the confidence of a phoneme segment. For instance, for a formant contour in the context-dependent triphone model c = "f-ao+r". When the rank and GMM log probability of c is estimated, we use equation below to measure the confidence of this phoneme segment：

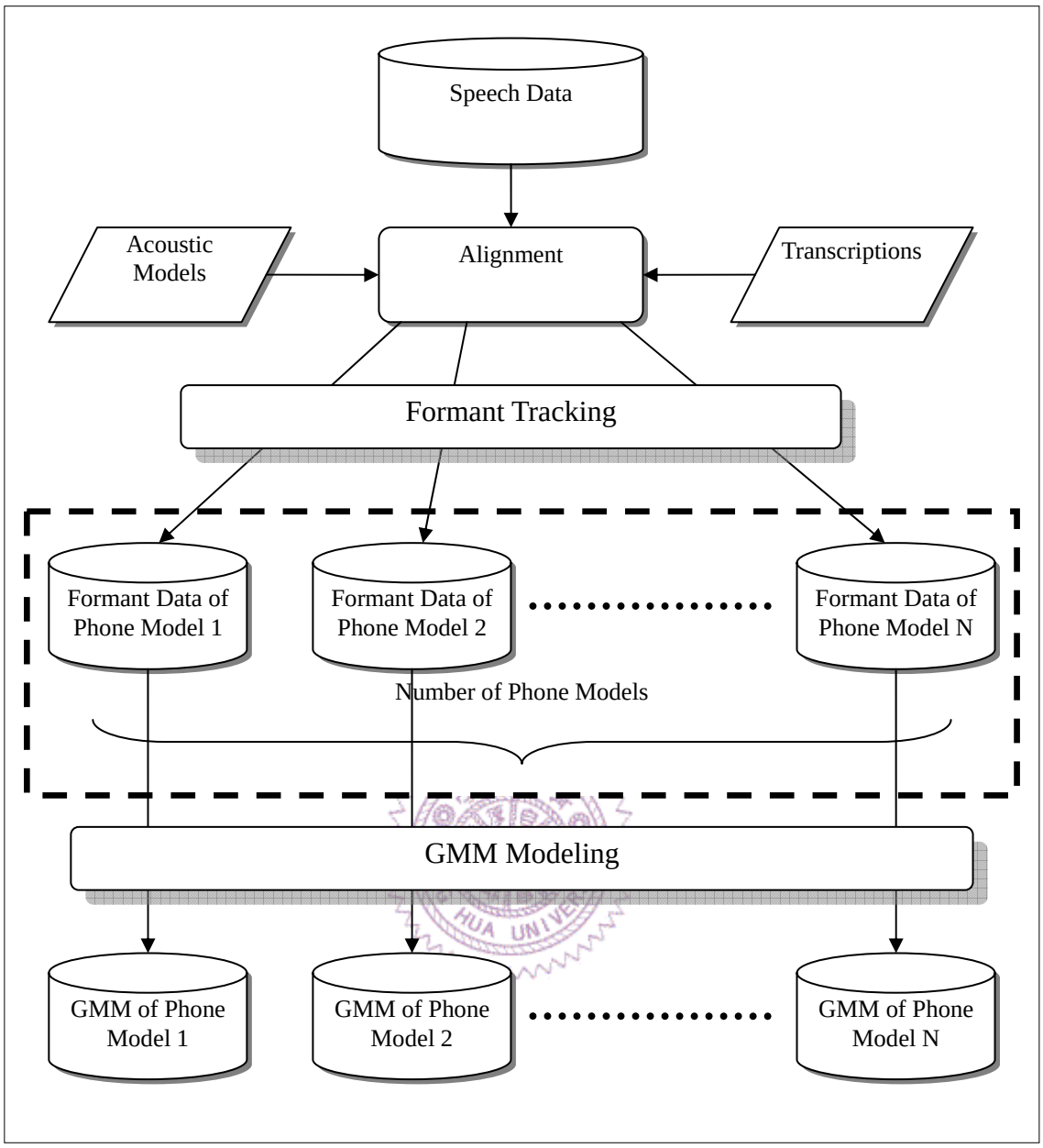


Figure 3.5 Flowchart of First Step of RCM

$$Conf_C = \frac{2}{1+\exp(\alpha \times (Rank_c - 1) \times \frac{\Pr ob(Rank_c)}{\Pr ob(Rank_1)})} \quad (4\text{-}1)$$

Where $Rank_{1,}$, $Rank_c$ indicate the first rank and the rank of c in the sorted order. $Prob(Rank_1)$ and $Prob(Rank_c)$ indicate the GMM log probabilities of $Rank_1$ and , $Rank_c$ respectively. α is a constant, and it decides the relation between rank and the overall confidence. In our case, α is fixed to 0.09, so a lower ranking number leads to higher confidence level. Figure3.6 shows different α with different contour according to the rank and the confidence.

3.6 Feedback Generation

In the last phase of the CAPT system, and also the most meaningful step, is the feedback presentation. This phase consisting in presenting the information obtained during the above phase. This phase of our system is to generate the articulator instruction according to the confidence measure and phonology. For a phoneme with lower confidence than threshold, a properly suggestion can be generated according to the map in Figure3.1. For example, if the normalized first formant estimated from student's utterance is too low for the CDTM "f-ao+r", the instruction will be "retroflex the tongue from low to high".

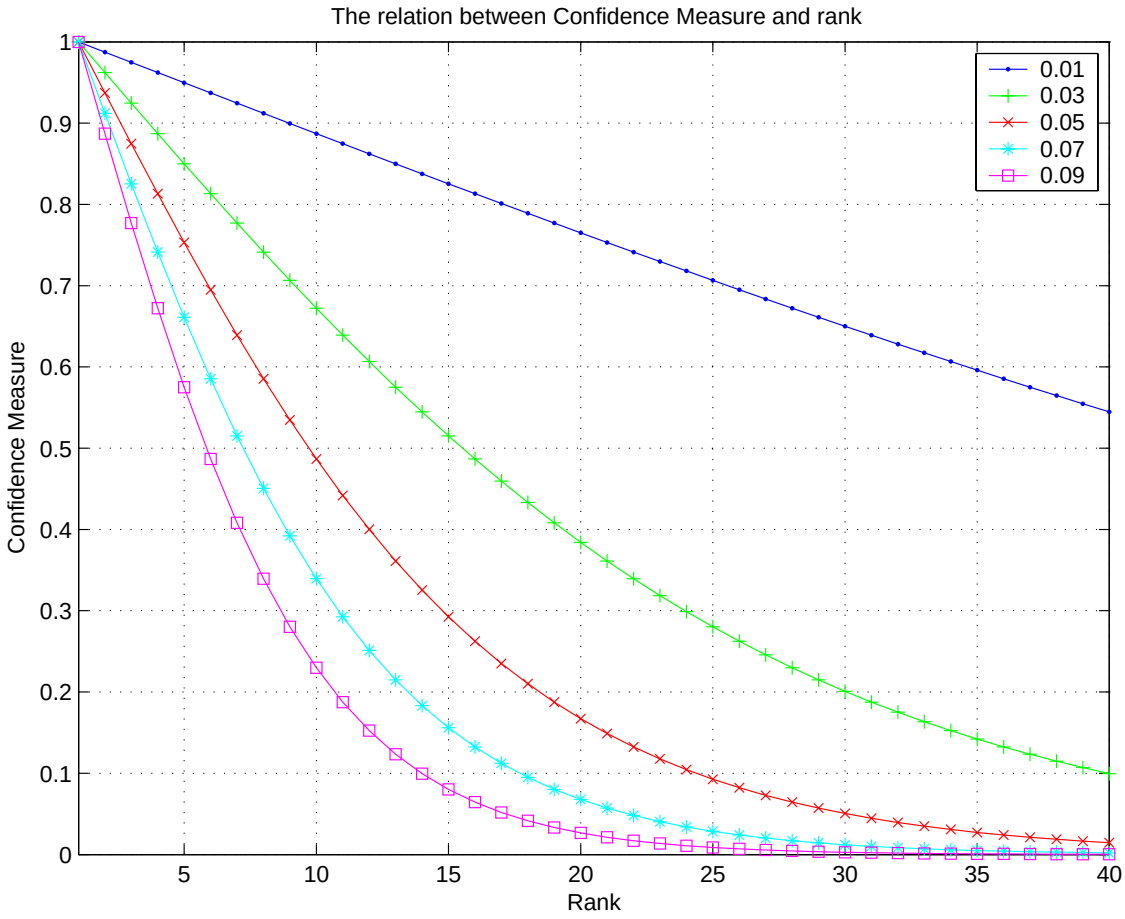


Figure 3.6 The relation between Confidence Measure and rank