An HMM System for Recognizing Articulation Features for Arabic Phones

Hosam Hammady[4], Sherif Abdou[13], Mostafa Shahin[12], Mohsen Rashwan[12] and Ossama Badawy[4]
[1]Research & Development International (RDI®), Giza, Egypt.
{sabdou, mostafa_shahin ,mrashwan}@rdi-eg.com
[2]Department of Electronics and Communication Engineering, Cairo University. Giza, Egypt.
[3]Department of IT, Faculty of Computers and Information, Cairo University. Giza, Egypt.
[4]College of Computing and Information Technology, Arab Academy for Science and Technology and Maritime Transport, Alexandria, Egypt.
hosam@hammady.net, obadawy@aast.edu

Abstract

In this paper, we introduce a Hidden Markov Model (HMM) recognition system for the articulation features of Arabic phones. The low-level features are described by Mel- Frequency Cepstral Coefficients (MFCCs). The created HMMs directly model certain articulation features. Articulation features are either place or manner features, here 10 basic manner features are used and arranged in pairs (Adhesion/Separation – Elevation/Lowering – Fluency/Desisting – Plosiveness/Fricativeness – Voicing/Unvoicing). Classification is done on these features regardless of the phone itself. The model has been created successfully and tested on reference speech data. The error rate is very low for many phones and acceptable for most of them. Finally, the system output is used as a confidence measure applied to other existing speech recognizers.

I. Introduction

Speech is the principal and the most convenient mode of communication among humans. Thus, technologies such as automatic speech recognition and text-to-speech have been under development since the early days of computer technology. [1]

Speech, a stream of utterances, produces time varying sound pressure waves of different frequencies and amplitudes. Speech recognition occurs when a corresponding sequence of discrete units (i.e. phones, words, or sentences) are derived from sound waves or acoustical waveforms. [2]

Several approaches are used to solve the problem of speech recognition; among which, Hidden Markov Models (HMMs) method is widely used.

Modern general-purpose speech recognition systems are generally based on HMMs. One possible reason why HMMs are used in speech recognition is that a speech signal could be viewed as a piecewise stationary signal or a short-time stationary signal. That is, one could assume in a short-time in the range of 10 milliseconds, speech could be approximated as a stationary process. Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. [3]

Normally, HMMs are able to model phones, words or sentences as classification units. However, as number of classification units increase, the classifier's task becomes more challenging. For example, In the Arabic language number of alphabet letters is 28. Number of phones is nearly double this number due to variance of pronunciation of each letter according to context. It has been suggested in other research [4] – [5] to group phones according to their phonological structure and performs the classification on such groups.

In linguistics, a distinctive feature is the most basic unit of phonological structure. Distinctive features are grouped into categories according to the natural classes of segments they describe: manner features, and place features. The place of articulation of a consonant is the point of contact, where an obstruction occurs in the vocal tract between an active (moving) articulator (typically some part of the tongue) and a passive (stationary) articulator (typically some part of the roof of the mouth). The manner of articulation describes how the tongue, lips, and other speech organs are involved in making a sound make contact. For any place of articulation, there may be several manners. [5]

Articulation features are more discriminative than MFCCs, which are the standard features for modern Automatic Speech Recognition (ASR) systems. A problem of MFCCs is that two phones could be very close in their acoustic models so they cannot be discriminated robustly. However, they could have different articulation features. An example is /@/ (همزة) as in /@/a/n/t/ (أنت) and /h/ (هاء) as in /h/a/z/a (هذا). They have very close acoustic features and can be classified as the same class by mistake. However, each one belongs to a different articulation feature class. The latter is fricative while the former is Plosive.

In this paper we introduce a speech recognition system based on HMM and articulation features. It is inspired by HAFSS© [8]–[9] which is a speech-enabled Computer Aided Pronunciation Learning (CAPL) System.

This CAPL system was developed for teaching Arabic pronunciations to non-native speakers. A challenging application is teaching the correct recitation of the Holy Quran[1]. The system uses a speech recognizer to detect errors in user recitation. To increase accuracy of the speech recognizer, only probable pronunciation variants, that cover all common types of recitation errors, are examined by the speech decoder. The decision reached by the recognizer is accompanied by a confidence score to reduce effect of misleading system feedbacks to unpredictable speech inputs.

Teaching the recitation of the Holy Quran is an artistic human-guided ancient2 task. Due to this fact, Tajweed [11] science was originally formalized. Tajweed is an ancient science that is well-known in Islamic literature. It is a set of rules studied by reciters to properly pronounce Quran. Such rules are typical mappings to linguistics and articulatory phonology. Each Arabic phone has a completely determined manner and place of articulation which varies according to its context among speech. To recite Quran, one should learn such articulation features to spell it the exact way. This highlights the importance of articulation features in Tajweed, which is much older than the science of phonetics. The resulting pronunciation is more sophisticated than the normal Arabic utterance. This is why distinct phones count to 50 although Arabic alphabet consists of only 28 letters.

In the implementation of a second language pronunciation teaching system [12], Witt argued that there exists no absolutely 'correct' pronunciation. A wide variety of pronunciations can be accepted by native speakers as being correct. However the Holy Quran is the word of Allah. And the purpose of Tajweed in essence is to make the reciter recite it the way it was originally received in the Classical Arabic dialect. [13] Finding a Tajweed teacher is not feasible all the time. This highlights the importance of CAPL systems in teaching Tajweed. Being an automated system motivates the improvement of score confidence to approach human accuracy and experience. The problem of the CAPL system mentioned above is its weak confidence measure. [8] Confidence measures (CM) are generally used to evaluate reliability of recognition results. A good confidence measure can largely benefit speech recognition systems in many practical applications. [14]

The initial target for this work was to create a confidence measure to evaluate the recognition results for this system, which could be generalized to be applicable on other existing ASR systems as well. The confidence score would be given to the recognized phones according to the matching of articulation features with the features classified by the system, along with system performance measure. In this paper we focus on implementing classifiers for several articulation features and merging them in a single score.

Section II describes the system design and basic units. Section III shows experimental results. Conclusion and possible future work are discussed in Sect. IV.

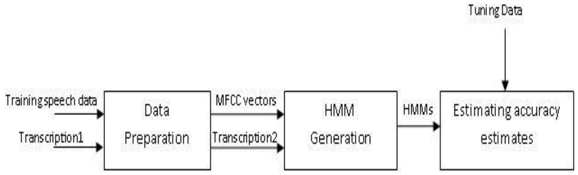


Figure 1: System design overview

II. SYSTEM DESIGN

The main objective is to build a model capable of classifying utterances according to articulation features. The articulation features grouped into sets each set consists of the articulation feature and it's opposite. The model is then trained for each set to classify into only 2 classes, either the feature or its opposite one. As shown in Figure 1, the system basically consists of 3 blocks. The first one is the data preparation, where feature extraction of speech data and transformation of transcriptions take place. This transformation is needed to map the original phones to either of the articulation classes. The second block describes generating and training the HMM models. The final block in the offline training is estimating accuracy estimates using the trained HMM models and tuning data (Sect. II-C). Following sections describe in detail each phase.

A. Data Preparation

Speech data is divided into two types; one type is the speech samples and the other is transcriptions for the speech samples. Transcriptions are phone labels corresponding to speech samples. Figures 2 and 3 illustrate preparation steps for both types, respectively.

Speech data is acquired from professional linguists. A portion of the data is hand-labeled by the linguists and the rest are unlabeled. This means that the labeled portion identifies exactly the phone classes for the corresponding speech samples. This is used to train the model and for testing the recognizer accuracy. Thus, the first step is to divide the labeled data into two portions, about 90% for training the model and the rest for testing it.

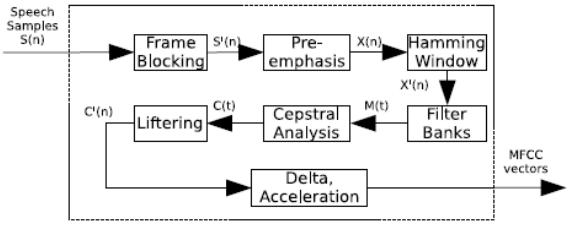


Figure 2: Speech data samples preparation

C_1	C_2	$\dots$	C_N	C_0
dC_1	dC_2	$\dots$	dC_N	dC_0
DC_1	DC_2	$\dots$	DC_N	DC_0

Table 1: FEATURE VECTOR.

Feature extraction is then performed on all speech samples. The target is MFCC features. Figure 2 summarizes such extraction.

The input to the first block is raw speech data. Frame-blocking, pre-emphasis and application of a Hamming-window are done as preprocessing of the digital speech samples. Filterbank method is used to perform frequency analysis using the Fourier transform. Cepstral analysis is then applied on the Mel-scale filterbank magnitudes. Liftering is then applied followed by delta analysis after which the feature vector becomes ready. It has the form depicted in Table I. *dCi* denotes delta coefficients where *DCi denotes* double-delta (or acceleration) coefficients. Note that *C*0 denotes the 0'th order cepstral coefficient representing the energy component. More details on MFCC generation can be found in [15] and [16].

The final step in data preparation is to transform the labeled data into new transcriptions where only two labels are used. Original phones are pre-classified as having certain feature classes. Transformation is simply done by renaming labels of phones to the group labels they belong to (Figure 3). This results in a transcription where labels of similar groups can appear successively. Such runs are merged as single labels.

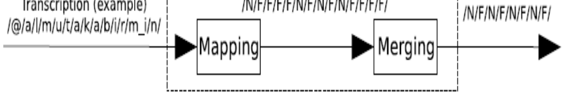

Transcription (example)
/@/a/l/m/u/t/a/k/a/b/i/r/m_i/n/
Mapping
/N/F/F/F/F/N/F/N/F/N/F/F/F/F/
Merging
/N/F/N/F/N/F/N/F/

Figure 3: Transcriptions preparation: an example. Input transcription contains labels from both classes *F*ricatives and *N*on-fricatives (plosive). *Mapping* is done to rename labels to their corresponding labels. After that *Merging* is done to merge similar consecutive labels.

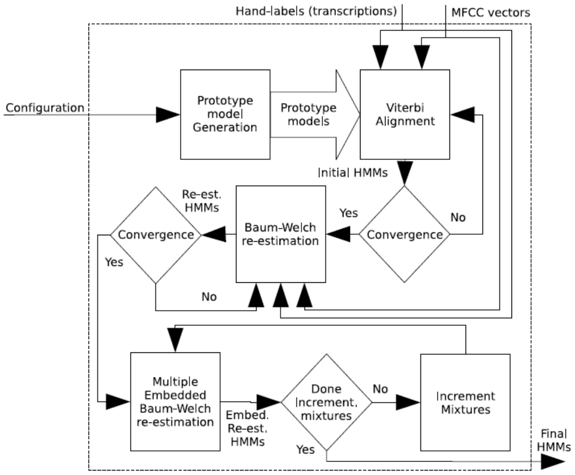


Figure 4: HMM generation details

B. Generating and Training the HMMs

The desired HMMs are simple models corresponding to each of the articulation feature classes. In our case there are only two classes; the feature and its opposite. A single HMM in its simplest form has three states, of which only one is emitting while other states are entry and exit states. Such auxiliary states are required for connecting models to do connected-word recognition afterwards. The emitting states of each model are initialized with single Gaussian mixture components for each feature vector component. The model parameters are repeatedly re-estimated using Viterbi alignment followed by Baum-Welch re-estimations. Rabiner [17] has written a very good tutorial on such algorithms (Viterbi and Baum-Welch) and how HMMs are trained formally. A detailed training procedure is illustrated in Figure 4 the conversion from single Gaussian HMMs to multiple mixture component HMMs is usually one of the final steps in building a system. This is done to cover diversity of feature vector components representing same classes which makes it difficult to fit a single Gaussian model on each class. It is usually a good idea to increment mixture components in stages, for example, by incrementing by 10 or 20 then re-estimating, then incrementing by 10 or 20 again and re-estimating, and so on until the required number of components is obtained. [18] Due to the non-uniformity of the distribution of phones over the articulation features, a different number of mixture components is applied to each phone class.

C. Estimating accuracy estimate

As shown in Figure 5, the trained acoustic models for each feature model are tested against reference speech samples resulting in hit-ratios grouped by phone.

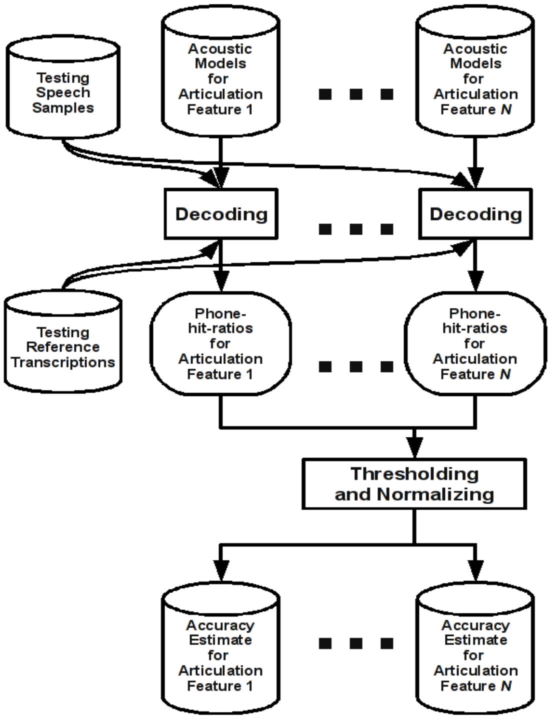

Testing Speech Samples
Acoustic Models for Articulation Feature 1
Acoustic Models for Articulation Feature N
Decoding
Decoding
Testing Reference Transcriptions
Phone-hit-ratios for Articulation Feature 1
Phone-hit-ratios for Articulation Feature N
Thresholding and Normalizing
Accuracy Estimate for Articulation Feature 1
Accuracy Estimate for Articulation Feature N

Figure 5: Estimating accuracy estimates

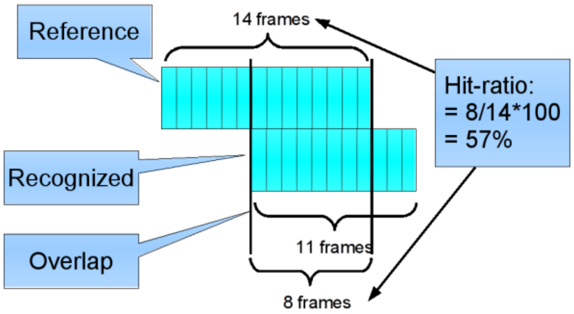


Figure 6: Hit-ratio calculation example

Figure 6 explain the meaning of hit ratio which can be formulated with the following formula:

$$hitratio = \frac{matc\ hing\ frames}{total\ \ frames} \qquad \ldots\text{Eq[II.1]}$$

Let's denote these hit-ratios as phone-hit-ratios these phone-hit-ratios represent how accurate given phones are classified using given feature models.

Not all phone-hit-ratios are expected to be high enough. Some phones may perform poorly for some feature models. Thresholding is thus done to discard results totally from such poor feature models for given phones. For example, if the threshold value is given as 85%, then phones resulting in phone-hit-ratios below this threshold on a given feature model will be discarded. This prevents poor feature models from affecting final accuracy estimate for corresponding phones. As a result, for each phone there exist zero or more effective feature models. They are normalized so that they sum up to unity. These values are the so-called accuracy estimates. They will be used later to estimate how accurate a phone is expected to be correctly classified given a feature model. Other poor feature models are given zeros for their accuracy estimates. In the case of a phone performing poorly in all feature models, the accuracy estimate is taken as the highest phone-hit-ratio among all feature models. This is done to protect accuracy estimates to be all zeros as this will lead to a confidence score always having a value of 0 for the given phone. This will result in always rejecting the given phone thus rendering a useless score.

The algorithm for thresholding and normalizing is illustrated in the following formulae:

$$\textbf{Thresholding: } \begin{cases} \boldsymbol{r_{mp} = \overline{r_{mp}}} & \boldsymbol{if\ \overline{r_{mp}} \geq T} \\ \boldsymbol{rmp = 0} & \boldsymbol{Otherwise} \end{cases} \quad ..\text{Eq[II.2]}$$

Where $\overline{\mathrm{r}_{\mathrm{mp}}}$ denotes phone-hit-ratio for phone p given a feature model m, $\boldsymbol{r_{mp}}$denotes effective phone-hit-ratio for phone p given a feature model m, and T represents the acceptance threshold.

$$\textbf{Normalizing: } \boldsymbol{a_{mp}} = \frac{\boldsymbol{r_{mp}}}{\boldsymbol{\sum_{j=1}^{M} r_{jp}}} \qquad \ldots\text{Eq[II.3]}$$

Where a_{mp} represents accuracy estimate for phone p given a feature model m and M denotes total number of feature models used. Viterbi decoding is then done on speech samples followed by the word-spotting technique which compares testing reference transcriptions to recognized feature classes. The main difference is that hit-ratios are calculated for the matching frames on phone basis instead of label basis, and are therefore called phone-hit-ratios. This can be further illustrated by the following formula:

$$\overline{r_{mp}} = \frac{matc\ hing\ frames}{total\ \ frames} \qquad \ldots\text{Eq[II.4]}$$

III. EXPERIMENTAL RESULTS

In this section the system design described earlier is going to be projected on experimental values and procedures. The data set used is going to be described in details. Articulation features listings are going to be illustrated.

A. Data Set Description

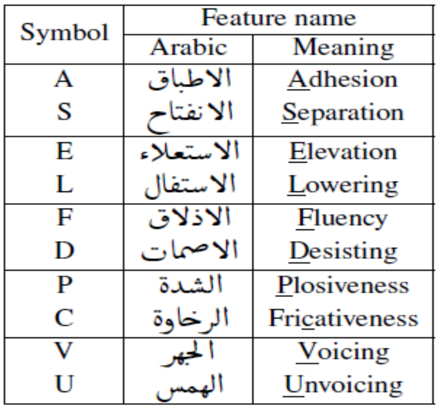
The data set used consists of verses from the Arabic Quran. It consists of two types. The first one is speech samples which are sound recordings acquired from professional linguists. The second type is transcriptions corresponding to the speech samples. Transcriptions are the representation of speech samples by labels describing phones contained in the samples by exact time boundaries. One problem in creating such transcriptions is that it is a very tedious task that consumes a very long time. Another problem is that it should be done by professional linguists who can accurately identify phone boundaries manually. Due to these problems, only a small portion of the speech samples is manually labeled as transcriptions. Thus speech samples having the corresponding manual transcriptions are referred to as handlabeled data.

The rest of the speech samples are unlabeled data. Automatic segmentation techniques can be applied to such data to obtain near-accurate transcriptions.

Collectively, the data set is divided into three parts:

- Training data. This represents the majority of the data. Part of it is handlabeled and the rest is automatically labeled. This part of the data is used to train the HMMs in the offline training phase.
- Tuning data for building the accuracy estimates. This constitues about 4.4% of all speech samples (40% of handlabeled data). It is used to test the HMMs after they have been created in order to build the accuracy estimates discussed in the previous chapter.
- Testing data for evaluating the accuracy of the confidence measure. This is also about 4.4% of all speech samples. It is used to verify that the trained models give high confidence scores to correct reference data.

Training data size is 9 hrs of which 1 hour is accurately hand-labeled by the linguists. The rest is auto-labeled. Testing data lasts for 24 minutes. This makes about 4.4% of the training data. All speech samples are encoded as raw data with mono-channel 16000 Hz and 16 bits-per-sample.



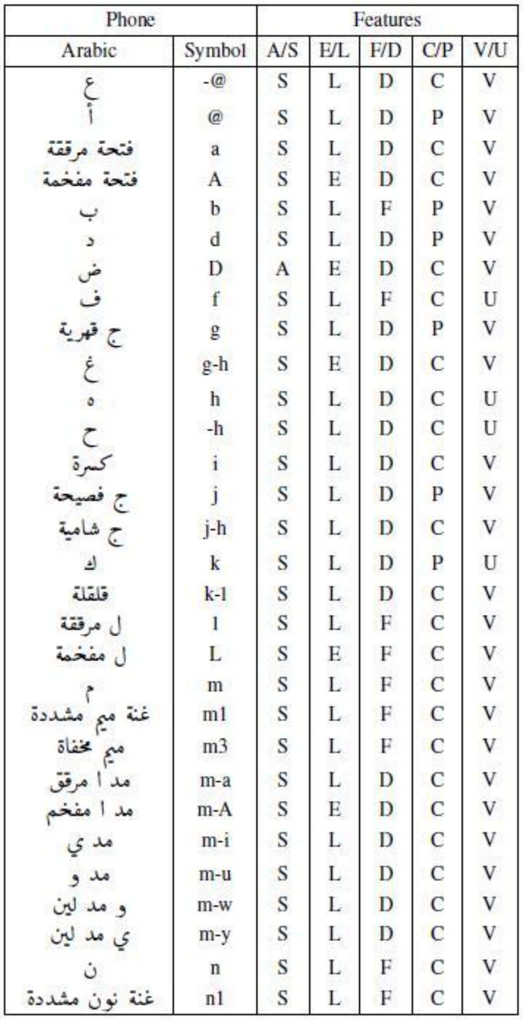
Symbol	Feature name	
	Arabic	Meaning
A S	الاطباق الانفتاح	Adhesion Separation
E L	الاستعلاء الاستفال	Elevation Lowering
F D	الاذلاق الاصمات	Fluency Desisting
P C	الشدة الرخاوة	Plosiveness Fricativeness
V U	الجهر الهمس	Voicing Unvoicing

Table 2: Manner features legend.

B. Articulation Features

As described earlier, the system generates the confidence scores by recognizing several articulation features then combining them in a single score. Here, the main manner features used are explained along with the different phones belonging to each feature.

The articulation features and its corresponding symbols used listed in Table 2. All phones with their manner features are listed in Table 3-1 and Table 3-2



Phone		Features				
Arabic	Symbol	A/S	E/L	F/D	C/P	V/U
ع	-@	S	L	D	C	V
أ	@	S	L	D	P	V
فتحة مرققة	a	S	L	D	C	V
فتحة مفخمة	A	S	E	D	C	V
ب	b	S	L	F	P	V
د	d	S	L	D	P	V
ض	D	A	E	D	C	V
ف	f	S	L	F	C	U
ج قهرية	g	S	L	D	P	V
غ	g-h	S	E	D	C	V
ه	h	S	L	D	C	U
ح	-h	S	L	D	C	U
كسرة	i	S	L	D	C	V
ج فصيحة	j	S	L	D	P	V
ج شامية	j-h	S	L	D	C	V
ك	k	S	L	D	P	U
قلقلة	k-1	S	L	D	C	V
ل مرققة	1	S	L	F	C	V
ل مفخمة	L	S	E	F	C	V
م	m	S	L	F	C	V
غنة ميم مشددة	m1	S	L	F	C	V
ميم مخفاة	m3	S	L	F	C	V
مد ا مرقق	m-a	S	L	D	C	V
مد ا مفخم	m-A	S	E	D	C	V
مد ي	m-i	S	L	D	C	V
مد و	m-u	S	L	D	C	V
و مد لين	m-w	S	L	D	C	V
ي مد لين	m-y	S	L	D	C	V
ن	n	S	L	F	C	V
غنة نون مشددة	n1	S	L	F	C	V

Table 3-1: Manner features used along with their phones

C. OPTIMIZING THE HMM MODEL PARAMETERS

The offline training phase involves training the acoustic models represented by HMM networks. The conversion from single Gaussian HMMs to multiple mixture component HMMs is done as a final step in such training. The mixtures are incremented in steps to inspect the best point in the sense of phone-hit-ratio. Figure 7 summarizes phone-hit-ratios for optimal HMMs.

Phone		Features				
Arabic	Symbol	A/S	E/L	F/D	C/P	V/U
غنة نون مخفاه	n3	S	L	F	C	V
ق	q	S	E	D	P	V
ر مرققة	r	S	L	F	C	V
ر مفخمة	R	S	E	F	C	V
س	s	S	L	D	C	U
ص	S	A	E	D	C	U
ش	s-h	S	L	D	C	U
ت	t	S	L	D	P	U
ط	T	A	E	D	P	V
ث	t-h	S	L	D	C	U
ضمّة	u	S	L	D	C	V
و	w	S	L	D	C	V
نون مدغمة قبل و	w1	S	L	D	C	V
خ	x	S	E	D	C	U
ي	y	S	L	D	C	V
نون مدغمة قبل ي	y1	S	L	D	C	V
ز	z	S	L	D	C	V
ذ	-z	S	L	D	C	V
ظ عامية	Z	A	E	D	C	V
ظ	-Z	A	E	D	C	V

Table 3-2: Manner features used along with their phones(cont.)

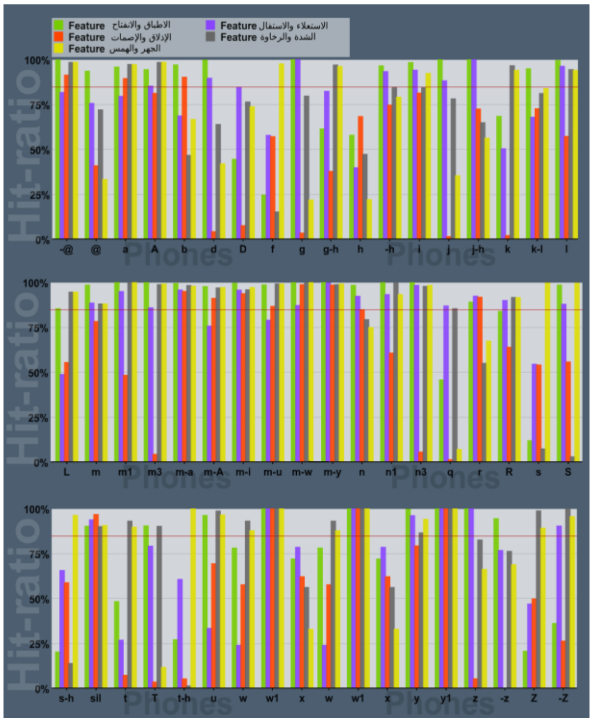


Figure 7: Summary of phone-hit-ratios for all feature models

D. EVALUATING THE CONFIDENCE MEASURE PERFORMANCE

The confidence measure system performance is tested against some reference testing data. Figure 8 shows average phone-hit-ratios resulting from this experiment. These figures, although not the real output of the system in the real-time phase, indicate how accurate reference data are given confidence scores.

It can be noticed that most phones achieve high performance. This is an indicator on confidence accuracy offered by the system. However, some phones still have poor performance. Refering to the analysis of the previous experiment, such phones have poor performance for all their articulation features. This is the reason why the combination of features rendered these results. It should be noted also that these results are statistically real indicators for phones that occur in speech with a large percent. However, for rare phones in speech, the results are not statistically reliable. As an average, the average confidence score is given by 92.58%. For the external ASR system, defining a confidence threshold below this value should render a reliable confidence measure.

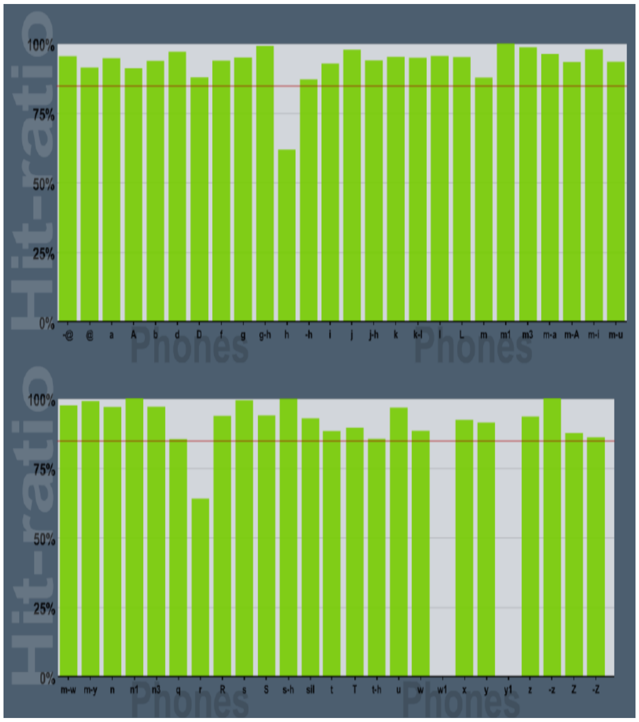


Figure 8: Average confidence measure results

E. INTEGRATING THE NEW CONFIDENCE INTO HAFSS

Testing utterances from the 29 speakers were applied to the models (before and after adaptation). Resulting confidence scores had to be feedback to Hafss CAPL system. Unfortunately, Hafss system was not available in the experimental environment. This was solved by acquiring recognition results from Hafss that was instructed to run in an offline batch mode. By this way, we simulated the realtime running of the confidence system along with Hafss. It is desired to give confidence scores to each phone segmented and recognized by Hafss. These scores are used with a threshold to determine if each of the recognized phones should be accepted, rejected or the user get prompted of a repeat request. This repeat request reflects the inability of the recognition system to determine – with an appropriate confidence – the correctness of the recognition decision. In order to compare the new confidence system results with the classical Hafss confidence measure, it was necessary to apply the confidence scores in the same manner the old confidence measure was applied. This manner is straightforward and can be summarized as follows:

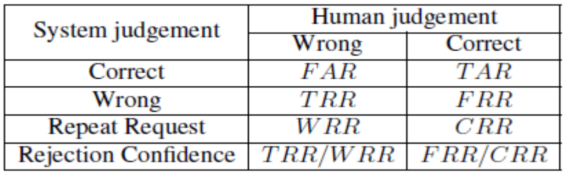
- If the recognized phone matches the corresponding reference phone that the user was requested to recite, then Hafss accepts the phone as correct and bypasses the confidence score.
- 2. Otherwise, Hafss uses the confidence score given to that phone as follows:
    - o If it is greater than a certain predefined threshold (confidence threshold), then Hafss is confident about its recognition and reject that phone prompting the user of his/her error.
    - o Otherwise, Hafss is not sure if its recognition was correct (low confidence), so it prompts the user to repeat his/her recitation because the phone was not clear.

The utterances used in this experiment contain a percent of error in the recitation. This was manually identified by human experts. It is required from Hafss to correctly identify errors and accept most of the correct parts of speech. The original work of Hafss [8] used utterances from 38 different speakers having 6.6% errors identified by human experts. It was not possible to acquire the same exact test utterances. Only 29 speakers were used having 3.8% error percent judged by human experts.

For the new confidence measure to be robust and useful, repeat requests should be minimized as much as possible for the errornous parts of speech. Because it is already known that the speakers made an error, it is desired that the system prompts for such errors. As for the correct parts of speech, it is desired to give low confidence for corresponding phones in order not to reject the correct phones, but to prompt the user to repeat such phone as they were not clear. As for wrong parts of speech that are identified as correct (the recognized phones match the reference phones), the confidence scores are not used, so the phones go undetected. The following points summarize various performance measures to be calculated from the system:

- True Acceptance Ratio (TAR): percent of correct speech that Hafss recognizes as correct.
- True Rejection Ratio (TRR): percent of wrong speech that Hafss recognizes as wrong.
- False Acceptance Ratio (FAR): percent of wrong speech that Hafss recognizes as correct.
- False Rejection Ratio (FRR): percent of correct speech that Hafss recognizes as wrong.
- Wrong Repeat Request Ratio (WRR): percent of wrong speech that Hafss prompts for repeat request.
- Correct Repeat Request Ratio (CRR): percent of correct speech that Hafss prompts for repeat request.
- Rejection confidence for wrong speech: ratio of TRR to WRR. It represents how much the system rejection is confident for wrong speech. It is desired to maximize this measure.
- Rejection confidence for correct speech: ratio of FRR to CRR. It represents how much the system rejection is confident for correct speech. It is desired to minimize this measure.

Table 4 illustrates the meaning of various performance measure.



System judgement	Human judgement	
	Wrong	Correct
Correct	FAR	TAR
Wrong	TRR	FRR
Repeat Request	WRR	CRR
Rejection Confidence	TRR/WRR	FRR/CRR

Table 4: Relation between human judgement and system judgement

These measures are optimized by varying the confidence threshold and recalculating them. However, TAR and FAR remain constant because they do not depend on the threshold where the confidence score is bypassed.

The results are illustrated graphically in Figure 9 as stacked vertical bars. The X-axis denotes the threshold value while the Y-axis denotes the percent of speech. TAR was omitted from the graph because it occupies most of the speech and it is constant, so no need to display it.

IV. CONCLUSIONS

Articulation features proved to be useful in phone recognition. A confidence measure system was built upon articulation features for the Arabic language. Using several articulation features proved to be useful in increasing the reliability of the confidence. This was achieved because articulation features appear with different strengths throughout the different phones. This enabled the system to compensate weak features for certain phones with strong features for the same phones. A combination method was constructed that takes into account the accuracy of each feature for the corresponding phones. Applying test reference data to estimate the accuracy of the confidence showed that the average confidence given was 92.58%. Accuracy of different features ranged from 80.34% to 93.01% for the original speaker.

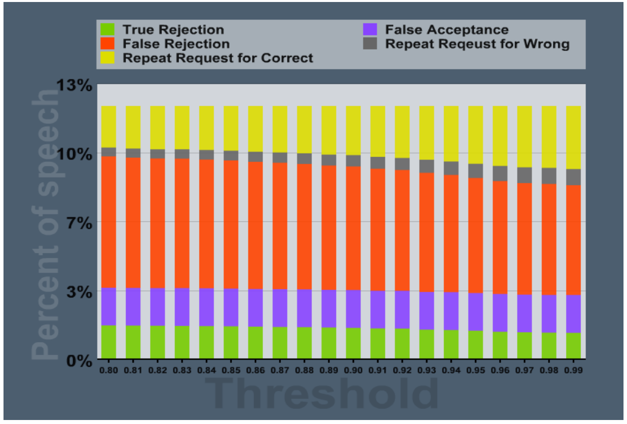


Figure 9: Hafss recognition results with the new confidence measure

V. FUTURE WORK

All articulation features, including manner and place features should be tried out to reach the best discriminative features. Discriminative training and analysis should be tried out. These methods take into account the discrimination power of the different classes. There is further tuning that could be done on the acoustic models to achieve higher accuracy in recognizing features.

References

[1] Schafer, R.W., “Scientific bases of human-machine communication by voice”. In proceedings of the National Academy of Science, Vol 92. Pp 9914–9920, 1995

[2] Moore, D.W., “Automatic speech recognition for electronic warfare verbal reports”, Unpublished master’s thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA. 1994.

[3] Rabiner, L. and Juang, B. 1993, “Fundamentals of Speech Recognition”. Prentice-Hall, Inc.

[4] R. Jakobson, G. Fant, and M. Halle. “Preliminaries to speech analysis”. Technical Report 13, MIT Acoustics Laboratory, 1952.

[5] G. A. Miller and P. E. Nicely. “Analysis of perceptual confusions among some English consonants”. Journal of the Acoustical Society of America, 27:338–352, 1955.

[6] Wang, Yu / Fosler-Lussier, Eric (2006): “Integrating phonetic boundary discrimination explicitly into HMM systems”, In INTERSPEECH-2006, paper 1820-Wed1BuP.5

[7] Jarifi, Safaa / Pastor, Dominique / Rosec, Olivier (2006): “Cooperation between global and local methods for the automatic segmentation of speech synthesis corpora”, In INTERSPEECH-2006, paper 1160-Wed1FoP.4

[8] Sherif. Abdou, S. Hamid , M. Rashwan, A. Samir, O. Abd-Elhamid, M. Shahin, W. Nazih “Computer Aided Pronunciation Learning SystemUsing Speech Recognition Techniques”, InterSpeech 2006, Pittsburgh, USA.

[9] S. Hamid “Computer Aided Pronunciation Learning System Using Statistical Based Automatic Speech Recognition Techniques”, PhD Thesis, Cairo University, Faculty of Engineering, Department of Electronics and Communication, Egypt, 2005.

[10] Islamic guide, http://www.islamicity.com

[11] About Tajweed, http://www.abouttajweed.com and Tajweed books, http://www.tajweedbooks.com

[12] Witt, S. M., “Use of speech recognition in computer-assisted language learning”, PhD thesis, Cambridge University Engineering Department, Cambridge, UK, 1999.

[13] Barkatulla, F., “The Importance of Tajweed”. Article number 934, http://www.islamicawakening.com/viewarticle.php?articleID=934

[14] Hui Jiang “Confidence measures for speech recognition: A survey”. Speech Communication, Volume 45, Issue 4, April 2005, Pages 455–470

[15] “Speech and Audio Signal Processing: Processing and Perception of Speech and Music”, B Gold, N Morgan - 1999 - John Wiley & Sons, Inc. New York, NY, USA

[16] Rabiner, L. R., and Schafer, R. M., “Digital Processing of Speech Signals” Prentice-Hall, Englewood Cliffs, N. J., 1978

[17] Rabiner, L.R. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. In Proceedings of the IEEE, Volume 77, Issue 2, Pages 257–286, 1989

[18] HMM ToolKit, http://htk.eng.cam.ac.uk/

[19] Hasegawa-Johnson, M., 2005. “Landmark-based speech recognition: Report of the 2004 Johns Hopkins Summer Workshop”. In: Proc. ICASSP, Vol. 1, pp. 213–216, http://people.csail.mit.edu/klivescu/papers/ws04ldmk_final.pdf