THE UNIVERSITY OF JORDAN

KING ABDULLAH II SCHOOL FOR INFORMATION TECHNOLOGY

AUTOMATIC SPEECH RECOGNITION (ASR) TECHNOLOGY

PRESENTED BY:

DR. MOHAMMAD A. M. ABUSHARIAH

Chapter Outline

- Lecture 1:
  - Introduction to ASR
- Lecture 2:
  - Written and Spoken Language Resources
- Lecture 3:
  - Implementation of ASR System

Lecture 1:
Introduction to ASR

Presentation Outline

Lecture 1

- Introduction
- What is ASR?
- Why is ASR Difficult?
- Problems Description
  - ASR Technology Related Problems
  - Language Related Problems
  - Speaker/Human Related Problems
- Overall Solution Approach
- Classifications of ASR Research Efforts
- Techniques Used for ASR Systems
- Software and Tools
- ASR Requirements
- References

Introduction

- Automatic Speech Recognition (ASR) is gaining its importance due to the vast growth generally in relevant technology and computing in specific.

- From industrial perspective, computers, laptops, and mobile devices nowadays have the ASR support embedded into the operating systems.

- From academia on the other hand, there are many research efforts being conducted addressing this technology in order to contribute to its state-of-the-art.

Introduction

- **Examples:**
  - Hands-free operating and control
  - Automatic query answering
  - Interactive voice response
  - Automatic dictation
  - Automatic speech translation
  - Pronunciation scoring
  - …etc.
- **Question:**
- How can ASR technology become beneficial to human?
  - ASR applications are of not much benefit to human unless they provide support to human natural languages worldwide including Arabic, Malay, English, Spanish, Mandarin, Dutch and various others.

Introduction

- Unlike English, many languages such as Arabic, Malay, and many others still need more research to achieve matured ASR technology and highly performing applications.

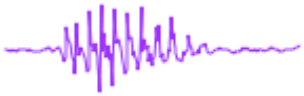
- Research interests have grown significantly in the past few years for ASR research for languages other than English.

- **For Example:** It is noticed that Arabic ASR research is not only conducted and investigated by researchers in the Arab world, but also by many others located in different parts of the world especially the western countries.

Introduction

- **Question:** Why research interests have grown significantly in the past few years for languages other than English?
- **1)** Effect of globalization.
- **2)** Number of speaking population (native and non native).
- **3)** Number of countries whereby it is the official language.
- **4)** Being among the 6 official languages of the United Nations.
- **5)** Politics and security.

What is ASR?



Speech Signal

Speech Recognition → Words
"How are you?"

Language Recognition → Language Name
English

Speaker Recognition → Speaker Name
James Wilson

WHAT IS ASR?

- Researchers and scientists defined speech recognition technology and ASR systems according to the way they use them in their research works.

- Generally, ASR systems aim at automatically extracting the string of spoken words from input speech signals.

- Forsberg (2003) defined automatic speech recognition (ASR) as the process of interpreting human speech in a computer.

- Jurafsky and Martin (2000) defined ASR more technically as the building of system for mapping acoustic signals to a string of words.

Why is ASR Difficult?

- **Speech is:**
  - Time-varying signal.
  - Well-structured communication process.
  - Depends on known physical movements.
  - Composed of known distinct units (phonemes).
- ⇒ ***Should Be Easy.***

Why is ASR Difficult?

- However, speech:
  - Is different for every speaker.
  - May be fast, slow, or varying in speed.
  - May have high pitch, low pitch, or be whispered.
  - Has widely-varying types of environmental noise.
  - Changes depending on sequence of phonemes.
  - Changes depending on speaking style ("read" vs. "conv.").
  - May not have distinct boundaries between units (phonemes).
  - Boundaries may be more or less distinct depending on speaker style and phoneme class.
  - Changes depending on the semantics of the utterance.
  - Has an unlimited number of words.

Problems Description

- In order to explicitly address problems and issues pertaining to ASR research, it is important to divide them into three major categories, which are:
- **1) ASR Technology Related Problems:**
- **Examples:** Isolated Words or Continuous Speech, Speaker-Dependent or Speaker-Independent, Vocabulary Size, ..etc.
- **2) Language Related Problems:**
- **Examples:** Different forms of the language, Lack of proper spoken resources, diacritical representations, ...etc.
- **3) Human/Speaker Related Problems:**
- **Examples:** The impact of participating speakers' characteristics such as gender, age, and region on the systems performance.

Problems Description

ASR Technology Related problems

- ASR is defined by certain issues that may affect its use, which are as follows (Rabiner and Juang, 1999):

- **1)** The manner used to speak to the machine, which includes isolated word, connected word, or continuous speech modes, whereby the latter is the hardest.

- **2)** The size of the recognition vocabulary, whereby ASR systems can be **small vocabulary** that can recognize up to 100 words, **medium vocabulary** that can recognize a range from 100 to 1000 words, and **large vocabulary** that have the capability of recognizing over 1000 words.

Problems Description

ASR Technology Related problems

- **3)** The knowledge of the user's speech patterns whereby ASR systems are classified into **speaker dependent** that are designed for each individual talker; **speaker independent** systems that work on broad population of talkers, and **speaker adaptive** that customize their knowledge to each individual user over time while the system is in use.
  - According to Huang and Lee (1993), speaker independent systems are more desirable to many applications and believed to be the ideal systems.
  - However, they can be outperformed by well-trained speaker dependent systems by a factor of two to three.
  - Therefore, high performance for speaker independent systems is harder to achieve compared to other systems.

Problems Description

ASR Technology Related problems

- **4)** Issues regarding the sources of the variability of speech, which include transducer, transmission system, and speaking environment variabilities should be highlighted.
  - Hence, each of these variabilities would have its effects on the production of speech.

Problems Description

Language Related problems

- Deficiencies of ASR research can also be caused by characteristics of the language itself.

- Similar to other languages, Arabic has its own set of issues that make the ASR task even more complex.

- **1)** Having different forms of Arabic language such as CA, MSA, and DA is an important issue, where each form has substantial differences with others especially in their script representation (Kirchhoff et al. 2003; and Elmahdy et al. 2009b).
  - As a result, a mismatch between written and spoken Arabic.

Problems Description

Language Related problems

- **Consequently:**
- Arabic ASR training materials become limited and potentially unsuitable to train the acoustic and language models for the recognizer,.

- Each word in the text may have a larger set of linguistic contexts and pronunciations with possibly different meaning, which leads to less predictive language models and a loss in recognition accuracy (Kirchhoff et al. 2003; Vergyri and Kirchhoff 2004; Alotaibi and Hussain 2010; Diehl et al. 2008).

Problems Description

Language Related problems

- **2)** Being a morphologically rich and productive language, words in Arabic can be concatenated using certain conjunctions, prepositions, articles, and pronouns through inserting prefixes and suffixes to the word stem, which leads to a huge list of potential word forms (Kirchhoff et al., 2003; Alghamdi et al., 2009).

- **3)** Lack of spoken and written training data is one of the main issues encountered by ASR researchers.
  - Majority of the available spoken and written language resources are not readily available to the public and many of them can only be obtained by purchasing from the Linguistic Data Consortium (LDC), the European Language Resource Association (ELRA), or other external vendors.

Problems Description

Language Related problems

- **4)** For many languages, the available spoken corpora are mainly collected from broadcast news (radios and televisions), and telephone conversations (Cieri et al., 2006.

- Broadcast news corpora are widely used in many recent ASR research efforts not only for its central interest and broad vocabulary coverage, but also for its abundant availability.
  - Systems developed using broadcast news corpora may lack generality, because this kind of data may not provide adequate variability among speakers and broadcast conditions.

Problems Description

Language Related problems

- **4) …. (Continue)**

- The spread of telephones helped in conversational corpora collection from samples (not necessarily local) in the population.
  - Variability among speakers is somewhat improved.
  - However, the telephone-based collection of data is a limited solution, because of its quality and variation characteristics of telephone networks and handsets.

Problems Description

Human/Speaker Related problems

- The human/speaker can certainly affect the performance of the ASR systems, because the he/she provides the data for training and testing the speech recognizer. Among these effects are:

- **1)** Sampling of subjects is among the major risks for linguistic data collection.
  - Language resources need to cover important categories related to gender, age, region, education, occupation, and others in order to provide an adequate representation of the subjects, which is necessary in order to develop ASR systems that serve a wide range of users and become speaker-independent.

Problems Description

Human/Speaker Related problems

- **2)** Native speakers could live in different regions or states, whereby each region or state has its own characteristics especially the dialects, accents, and speaking styles used, which in fact are the major differences between them.
  - In order to develop ASR systems that can cater for these differences, an acceptable sampling of the speakers is required.
  - It is unfair to claim that the developed ASR system can be used by any native speaker regardless of the region or state unless a representative data from each region or state is collected.

Problems Description

Human/Speaker Related problems

- **3)** The gender and its effect on speech acoustic parameters (Biemans , 2000; Eunjin, 2011).
  - Females have higher fundamental frequencies and formant frequency values than males.
  - Females also use wider vowel spaces and more frequently produce glottal stops more than males.
  - Females are found to have a more breathy and less harsh voice than males.
  - Females also speak in higher pitch and lower loudness than males.
  - The speaking rate of males is faster than that of females.

Problems Description

Human/Speaker Related problems

- **4)** Another human/speaker related factor to be taken into account is the age.
  - Major vocal characteristics for old voices are increased breathiness, lower pitch, increased harshness, possibility of higher voice breaks, and reduced loudness (Linville, 2002).
  - Adults and children speech are compared in Benzeghiba et al. (2007) showing that children speech is worse than adults speech on average Word Error Rates (WER) of two to five times higher.
  - The effect of ageing on ASR is investigated by Vipperla et al. (2010) and found that the WER is 10% higher in older voices compared to adult voices.

Overall Solution Approach

- It is important to address the ASR technology, the language, and human/speaker related problems in order to eliminate their impact on causing mismatches between the training and testing data and ultimately improve the performance of the ASR systems.

- From ASR technology perspective, major issues such as the manner used to speak to the machine, the size of the recognition vocabulary, the knowledge of the user's speech patterns, and the sources of the variability of speech should be addressed.

Overall Solution Approach

- A lot of emphasis on language and human/speaker related issues must be given especially in form of providing language resources that are neither broadcast news, nor telephone conversations, by taking into account a proper sampling and representation of speakers.

- It is believed that a lot of impact can be due to human/speaker variability such as the gender, age, education, region, and others.

- Therefore, such variabilities have to be taken into serious consideration when collecting the data.

Classifications of ASR Research Efforts

Case Study on Arabic Language

- From literature investigation, ASR research efforts on Arabic language are classified as follows:
- **1) Isolated Arabic part of word {consonants, vowels, syllables, phonemes, and phones} recognition systems**
  - Recognize small units and segments of speech such as alphabets, phones, vowels, and syllables.
- **2) Isolated Arabic words recognition systems**
  - Recognize small units such as isolated words, command and control, and Arabic digits.
- **3) Continuous Arabic speech recognition systems**
  - Largest and most difficult units such as The Holy Qur'an verses, questions, proverbs, broadcast news, broadcast conversations, and telephone conversations.
- **NOTE:** The larger the unit of speech to be recognized, the harder is the ASR task.

Techniques Used for ASR Systems

- For any ASR system, features extraction and classification techniques are applied.
- Features extraction techniques normally focus on extracting unique, discriminative, robust and computationally efficient characteristics from the input speech signals, which result in producing corresponding feature vectors.
- These feature vectors are then trained and classified into unique patterns or classes by means of features classification techniques (Ursin, 2002).

Techniques Used for ASR Systems

- **Feature Extraction:**
- Based on literature investigation, it is noticed that many researchers applied:
  - Mel-Frequency Cepstral Coefficient (MFCC)
  - Perceptual Linear Prediction (PLP)
  - Linear Predictive Coding (LPC)

- However, MFCC is the most dominant and prevalent technique for extracting spectral features, which leads to a performance that is slightly superior to PLP and LPC (Huang et al., 2001; Chetouani et al., 2002; Milner, 2002; Hönig et al., 2005).

Techniques Used for ASR Systems

- **Feature Classification:**
- Based on literature investigation, it is noticed that many researchers applied:
  - Vector Quantization (VQ)
  - Artificial Neural Networks (ANN)
  - Hidden Markov Models (HMM)

- Due to its powerful statistical tools, HMM is the most widely used technique for building acoustic models for all languages in general.

- Unlike VQ and ANN, HMM is very successful in managing time variant speech files (Marcos, 2005).

Software and Tools

- As far as ASR research efforts are concerned, Hidden Markov Model Toolkit (HTK) and Carnegie Mellon University (CMU) Sphinx engine are the most widely used open source ASR toolkits, and they are getting more and more popular as the ASR technology is applied into new languages.
- The HTK and CMU Sphinx contain ready-to-use downloadable tools, which are devoted for training the acoustic models due to their capabilities in implementing large vocabulary, speaker-independent, continuous speech recognition system in any language (Samudravijaya and Barot, 2003; Kacur and Rozinaj, 2008; Novak et al., 2010).

Software and Tools

- Although both HTK and CMU Sphinx have common goal to achieve, they have various differences.

- Samudravijaya and Barot (2003) believed that CMU Sphinx has more advanced features and its license is meant for unrestricted use as compared to HTK.

- Samudravijaya and Barot (2003) also experimented the use of HTK and CMU Sphinx and concluded that the CMU Sphinx is able to produce better quality acoustic models than that of the HTK.

Software and Tools

- **Major technical differences include:**
- **1)** HTK is more flexible in terms of allowing the users to specify the number of states for each unit, whereas CMU Sphinx has fixed the number of states to 5-state models.

- **2)** For language modeling, HTK supports the use of bi-gram models, whereas CMU Sphinx supports both bi-gram and tri-gram language models.

- **3)** HTK is more user-friendly than CMU Sphinx.

Software and Tools

- **Major technical differences include:**
- **4)** Overall, CMU Sphinx is believed to be better than HTK especially in terms of performance and accuracy rates.
  - Based on the above as well as the tables 1, 2, and 3 as presented earlier, it is noticed that many researchers utilized the CMU Sphinx tools especially for large vocabulary, speaker-independent, continuous speech recognition systems.

ASR Requirements

- In order to develop an ASR system, the following major requirements are needed:
- **1)** Written and Spoken Resources
- **2)** Phonetic Dictionary
- **3)** Feature Extraction
- **4)** Acoustic Model Training
- **5)** Language Model Training
- **6)** Decoder

REFERENCES

- Jurafsky, D., and Martin, J.H. (2000). Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, Upper Saddle River, NJ, USA.
- Forsberg, M. (2003). Why is Speech Recognition Difficult?. Department of Computing Science, Chalmers University of Technology, Gothenburg, Sweden.
- Rabiner, L. R., and Juang, B. H., (1999). Speech Recognition by Machine. In Madisetti, V. K., and Williams, D. B. (Eds.), *Digital Signal Processing Handbook* (pp. 987 – 1002). CRCnetBASE. CRC Press LLC.
- Abushariah, M. A. M., (2006). A Vector Quantization Approach to Isolated-Word Automatic Speech Recognition. Master Thesis, Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.
- Huang, X., and Lee, K. F., (1993). On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 2, pp. 150 – 157.
- Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R., and Vergyri, D., (2003). Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns-Hopkins Summer Workshop. *Proceedings of the IEEE International Conference on Acoust Speech, and Signal Processing*, Hong Kong, Vol. 1, pp. 344 – 347

References

- Vergyri, D., and Kirchhoff, K., (2004). Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, pp. 66 – 73.
- Alotaibi, Y. A., and Hussain, A., (2010). Comparative Analysis of Arabic Vowels using Formants and an Automatic Speech Recognition System. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol. 3, No. 2, pp. 11 – 22.
- Diehl, F., Gales, M. J. F., Tomalin, M., and Woodland, P. C., (2008). Phonetic Pronunciations for Arabic Speech-to-Text Systems. *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*. Las Vegas, USA, pp. 1573 – 1576.
- Alghamdi, M., Elshafei, M., and Al-Muhtaseb, H., (2009). Arabic Broadcast News Transcription System. *International Journal of Speech Technology*, Springer, pp. 183 – 195.
- Cieri, C., Liberman, M., Arranz, V., and Choukri, K., (2006). Linguistic Data Resources. In Schultz, T., and Kirchhoff, K. (Eds.), *Multilingual Speech Processing* (pp. 33 – 70). USA, Academic Press, Elsevier.
- Eunjin, O., (2011). Effects of Speaker Gender on Voice on Set Time in Kore Stops. *Journal of Phonetics*. Vol. 39, pp. 59 – 67.

References

- Biemans, M., (2000). *Gender Variation in Voice Quality*. LOT Publisher, The Netherlands.
- Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C., (2007). Automatic Speech Recognition and Speech Variability: A Review. *Speech Communication*, Vol. 49, pp. 763 – 786.
- Linville, S. E., (2002). Source Characteristics of Aged Voice Assessed from Long-Term Average Spectra. *Journal of Voice*. Vol. 16, No. 4, pp. 472 – 479.
- Vipperla, R., Renals, S., and Frankel, J., (2010). Ageing Voices: The Effect of Changes in Voice Parameters on ASR Performance. *EURASIP Journal on Audio, Speech, and Music Processing*. Vol. 2010, Article ID 525783, 10 pages, doi:10.1155/2010/525783
- Ursin, M., (2002). Triphone Clustering in Finnish Continuous Speech Recognition. Master Thesis, Department of Computer Science, Helsinki University of Technology, Finland
- Huang, X., Acero, A., and Hon, H. W., (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall, Upper Saddle River, NJ, USA.
- Hönig, F., Stemmer, G., Hacker, C. and Brugnara, F., (2005). Revising Perceptual Linear Prediction (PLP). *INTERSPEECH'05*. Lisbon, Portugal, pp. 997–300

References

- Milner, B., (2002). A Comparison of Front-End Configurations for Robust Speech Recognition. *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*. Vol. 1, pp. I797 – I800.
- Chetouani, M., Gas, B., Zarader, J. L., and Chavy, C., (2002). Neural Predictive Coding for Speech Discriminant Feature Extraction: The DFE-NPC. *Proceedings of the European Symposium on Artificial Neural Networks*, Bruges, Belgium, pp. 275 – 280.
- Marcos, F. Z., and Enric, M. M., (2005). State-of-the-Art in Speaker Recognition, *IEEE Transaction on Aerospace and Electronic Magazine*, Vol. 20, No.5, pp.7 – 12.
- Samudravijaya, K., and Barot, M., (2003). A Comparison of Public Domain Software Tools for Speech Recognition. *Workshop on Spoken Language Processing*. India.
- Novak, J. R., Dixon, P. R., Furui, S., (2010). An Empirical Comparison of the T^3, Juicer, HDecode and Sphinx3 Decoders. *INTERSPEECH'10*. Japan, pp. 1890 – 1893.
- Kacur, J., and Rozinaj, G., (2008). Practical Issues of Building Robust HMM Models Using HTK and SPHINX Systems. In Mihelič, F., and Žibert, J. (Eds.), *Speech Recognition, Technologies and Applications* (pp. 171 – 192). Austria, I-Tech Education and Publishing.

End of Lecture 1

Thank You Very Much for your Concentration!!!

Questions and Answers Session

Lecture 2:

Written and Spoken Language Resources

Presentation Outline

Lecture 2

- Introduction
- Language Forms
- Written and Spoken Resources
- Speech Corpus Participants
- Speech Corpus Recording Set-up and Equipment
- Speech Corpus Preparation and Pre-processing Requirements
- Distribution of the Speech Corpus for ASR Implementation and Evaluation
- References

Introduction

- According to Elmahdy et al. (2009a):
  - Arabic language is the largest Semitic language which is still in existence.
  - It is one of the six official languages of the United Nations (UN).
  - The number of first language speakers of Arabic exceeds 250 million, whereas the number of second language speakers can reach four times the number of first language speakers.
  - It is the official language in 21 countries situated in Levant, Gulf, and Africa.
  - It is ranked as fourth after Mandarin, Spanish and English in terms of the number of first language speakers.

Language Forms

- Arabic language consists of three main forms, each of which has distinct characteristics.
- These forms are **1)** Classical Arabic (CA), **2)** Modern Standard Arabic (MSA), and **3)** Colloquial or Dialectal Arabic (DA).

- Al-Sulaiti and Atwell (2006) believed that there is another form of Arabic language referred to as Educated Spoken Arabic (ESA), which is considered as a hybrid form that derives its features from both the standard and dialectal forms, and is mainly used by educated speakers.

Written and Spoken Resources

- Written and spoken resources are closely related and very necessary to exist in order to develop any ASR system.

- Written and spoken corpora are examples of linguistic resources for a language, which normally consist of large sets of machine readable data that are used for developing, improving, and evaluating natural language, and speech algorithms and systems.

- Advancements in these technologies elevated the need by many communities for written and spoken resources in large volumes with relatively different types of data and variety of languages (Godfrey and Zampolli, 1997; Ejerhed and Church, 1997; Lamel and Cole, 1997; Cieri et al., 2006).

Written and Spoken Resources

- Depending on the type of data to be collected and the application to be developed, the written corpus can be produced prior to the spoken corpus or vice-versa.

- Spoken corpora contain signals that correspond to the pronunciation of utterances by various speakers, which are used to develop the acoustic models in ASR systems.

- Written corpora contain texts that correspond to the utterances pronounced in the spoken corpora, which are used to develop the language model in ASR systems.

Written and Spoken Resources

- For instance, the written corpora must be prepared prior to the spoken corpora in read speech, whereas in conversational speech the spoken corpora are normally produced first and the written corpora are transcribed either manually or using semi-automatic approaches (Mariani, 1995; Ejerhed and Church, 1997; Lamel and Cole, 1997).

- In ASR systems, the spoken data type is given more emphasis especially to the various styles of the spoken corpora, because the written corpora can be transcribed manually or using semi-automatic approaches.

- As a result, the type and contents of the written corpora are dependent on and determined by the type and contents of the spoken corpora.

Written and Spoken Resources

- The relationship between the written and spoken forms of the language resources is essential to be addressed since both forms are required for various applications especially for ASR research.

- Many of the available Arabic spoken resources are collected prior to having the written form.
  - In such resources, the written form is produced as a result to what has been collected in the spoken form.

- From the investigation of linguistic characterization of speech and writing (Parkinson and Farwaneh, 2003), writing is more structurally complex and elaborate, more explicit, and more organized and planned compared to speech.

Written and Spoken Resources

- Due to these differences, the written form of the corpora needs to be created prior to producing and recording the spoken form for more comprehensive data.
  - Therefore, linguists and phoneticians carefully produce written corpora before handing them to speech recording specialists for recording purposes.

- In the past few years, efforts have been devoted to the design and development of speech corpora for different languages.

- These efforts have addressed the relationship between the written and spoken forms of the corpora, and gave more emphasis to designing quality written form that embeds the language's phonetic knowledge prior to collecting the spoken form.

Written and Spoken Resources

- Speakers would have their own speaking style; however, their speech of the same language has the same phonological structure.
  - The phonological level of the language is selected to design phonetically rich and balanced text and speech corpora for many languages (Uraga and Gamboa, 2004).

- Creating phonetically rich and balanced text corpora requires selecting a set of phonetically rich words, which are combined together to produce sentences and phrases.

- These sentences and phrases are verified and checked for balanced phonetic distribution.

Written and Spoken Resources

- Some of these sentences and phrases might be deleted and/or replaced by others in order to achieve an adequate phonetic distribution (Pineda et al., 2004).

- Such text corpora are then recorded in order to produce phonetically rich and balanced speech corpora.

- This approach is highly adopted in languages such as English (Garofolo et al., 1993; Black and Tokuda, 2005; D'Arcy and Russell, 2008), Mandarin (Chou and Tseng, 1999; Liang et al., 2003), Spanish (Uraga and Gamboa, 2004), and Korean (Hong et al., 2008).

Written and Spoken Resources

- In order to produce a speaker-independent, continuous, and automatic speech recognizer, a set of speech recordings that are rich and balanced is required.
- The **rich characteristic** is in the sense that it must contain all phonemes of the target language.
- The **balanced characteristic** is in the sense that it must preserve the phonetic distribution of the target language.
- This set of speech recordings must be based on a proper written set of sentences and phrases created by experts.

Written and Spoken Resources

- Jorschick (2009) identified four main speech styles of the corpus that are determined by the task used to collect the data.
  - **1)** Read speech
  - **2)** Elicited experimental speech
  - **3)** Semi-spontaneous monologue speech
  - **4)** Conversational speech

- **Question:** What spoken resources do you need for Arabic language?

Written and Spoken Resources

- The need for Arabic spoken resources was surveyed by Nikkhou and Choukri (2004 and 2005).
  - Prepared and Read Speech for office environment.
  - In most cases, respondents did not show much interest in telephone and broadcast news spoken resources.
- Written and spoken resources are publically available to all communities through membership subscription to the Linguistic Data Consortium (LDC) and the European Language Resources Association (ELRA) online catalogs.

Written and Spoken Resources

- Based on the language resources catalogs provided by the LDC and the ELRA, there are:
  - 13 {4 from LDC, and 9 from ELRA} and 30 {22 from LDC, and 8 from ELRA} spoken corpora for MSA and DA forms, respectively.
  - 71 {64 from LDC, and 7 from ELRA} written corpora.
- This analysis indicates that the written corpora for Arabic language are available in large volumes.
- However, there is real lack of spoken corpora especially for MSA form.
- Therefore, more work need to empathize on providing written and spoken corpora for MSA form.

Speech Corpus Participants

- The phonetically rich and balanced Arabic speech corpus was initiated in March 2009.
- Although participants were generally recruited based on their interest to join this work, speakers were suitably and specifically selected based on predetermined characteristics as follows:
  - They have a fair distribution of gender and age.
  - Their current professions vary.
  - They have a mixture of educational backgrounds with a minimum of high school certification. This is important to secure an efficient reading ability of the participants.
  - They belong to various native Arabic speaking countries.
  - They belong to any of the three major regions where Arabic native speakers mostly live (Levant, Gulf, and Africa). This is important to produce a comprehensive speech corpus that can be used by all Arabic language research community.

Speech Corpus Recording Set-up and Equipment

- Recording sessions were conducted in a sound-attenuated studio located in the language center in the International Islamic University Malaysia (IIUM), whereby participants were asked to complete their recordings in one session.
- However, some participants exceeded one session and completed their recordings in 2 to 3 sessions due to scheduling reasons.
- Participants were asked to read the 415 sentences prepared for this task.
- Each sentence was recorded at least twice depending on the participant's reading ability and quality.
- Some participants had to utter sentences for 10 times due to pronunciation deficiencies and mistakes.

Speech Corpus Recording Set-up and Equipment

- During the recording sessions in the sound-attenuated studio room, speakers used the SHURE SM58 wired unidirectional dynamic microphone to utter the recordings.
- They also used the Beyerdynamic DT 231 Headphone in order to listen to instructions from the recording specialist.
- In addition, the YAMAHA 01V 96 Version 2 (Digital Audio Mixer) was used.
- In terms of software, Sony Sound Forge 8 was used on a normal Personal Computer (PC) located in the studio with Windows XP in order to record the utterances from the speakers.

Speech Corpus Preparation and Pre-processing Requirements

- In order to use the phonetically rich and balanced speech corpus for training and testing Arabic ASR systems, a number of MATLAB programs have to be developed in producing a ready-to-use speech corpus.
- These MATLAB programs intend to provide all necessary preparation and pre-processing requirements for the speech corpus.
    - **1)** Automatic Arabic speech segmentation.
    - **2)** Parameters conversion of speech data.
    - **3)** Directory structure and sound filenames convention.
    - **4)** Automatic generation of training and testing transcription files.
- Manual classification and validation of the correct speech data were conducted with great care and precision.
- **Run Example on MATLAB**

Distribution of the Speech Corpus for ASR Implementation and Evaluation

- During the fourth development phase, a total of 36,071 utterances were used resulting in about 38 hours of speech data collected from 36 (18 male and 18 female) Arabic native speakers from 11 different Arab countries.

- The leave-one-out cross validation and testing approach was applied, where every round speech data of 35 out of 36 speakers were trained and speech data of the 36th were tested.

References

- Elmahdy, M., Gruhn, R., Minker, W., and Abdennadher, S., (2009a). Survey on common Arabic language forms from a speech recognition point of view. *International Conference on Acoustics (NAG-DAGA)*, Rotterdam, Netherlands, pp. 63 – 66.
- Al-Sulaiti, L., and Atwell, E., (2006). The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, John Benjamins Publishing Company, pp. 1 – 36.
- Ejerhed, E., and Church, K., (1997). Language Resources: Written Language Corpora. In Cole, R., Mariani, J., and Uszkoreit, H. (Eds.), *Survey of the State of the Art in Human Language Technology* (pp. 359 – 363). Italy, Cambridge University Press and Giardin.
- Cieri, C., Liberman, M., Arranz, V., and Choukri, K., (2006). Linguistic Data Resources. In Schultz, T., and Kirchhoff, K. (Eds.), *Multilingual Speech Processing* (pp. 33 – 70). USA, Academic Press, Elsevier.
- Godfrey, J. J., and Zampolli, A., (1997). Language Resources: Overview. In Cole, R., Mariani, J., and Uszkoreit, H. (Eds.), *Survey of the State of the Art in Human Language Technology* (pp. 357 – 359). Italy, Cambridge University Press and Giardin.

References

- Lamel, L., and Cole, R., (1997). Language Resources: Spoken Language Corpora. In Cole, R., Mariani, J., and Uszkoreit, H. (Eds.), *Survey of the State of the Art in Human Language Technology* (pp. 363 – 367). Italy, Cambridge University Press and Giardin.
- Mariani, J., (1995). Tasks of a European Center for Spoken Language Resources (ECSLR). Technical Report, Mlap SPEECHDAT Project, Computer Sciences Laboratory for Mechanics and Engineering Sciences, National Center for Scientific Research, France.
- Parkinson, D. B., and Farwaneh, S., (Ed), (2003). *Perspectives on Arabic Linguistics XV*. John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 149 – 180.
- Uraga, E., and Gamboa, C., (2004). VOXMEX Speech Database: Design of a Phonetically Balanced Corpus. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Portugal, pp. 1471 – 1474.
- Pineda, L. V., Montes-y-Gómez, M., Vaufreydaz, D., and Serignat, J-F., (2004). Experiments on the Construction of a Phonetically Balanced Corpus from the Web. *5th International Conference on Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, Springer, Vol. 2945/2004, pp. 416 – 419, Korea.

References

- Black, A. W., and Tokuda, K., (2005). The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. *INTERSPEECH'05*, Portugal, pp. 77-80.
- Chou, F. C., and Tseng, C. Y., (1999). The Design of Prosodically Oriented Mandarin Speech Database. *ICPhS'99*, San Francisco, pp. 2375 – 2377.
- D'Arcy, S., and Russell, M., (2008). Experiments with the ABI (Accents of the British Isles) Speech Corpus. *INTERSPEECH'08*, Australia, pp. 293 – 296.
- Hong, H., Kim, S., and Chung, M., (2008). Effects of Allophones on the Performance of Korean Speech Recognition. *INTERSPEECH'08*, Australia, pp. 2410 – 2413.
- Liang, M. S., Lyu, R. Y., and Chiang, Y. C., (2003). An Efficient Algorithm to Select Phonetically Balanced Scripts for Constructing a Speech Corpus. *IEEE Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, China, pp. 433 – 437.
- Jorschick, A., (2009). Sound to Sense Corpus Manual. Technical Report, Faculty for Linguistics and Literary Sciences, University of Bielefeld, Germany.
- Nikkhou, M., and Choukri, K., (2004). Survey on Industrial needs for Language Resources. Technical Report, NEMLAR – Network for Euro-Mediterranean Language Resources.

References

- Nikkhou, M., and Choukri, K., (2005). Survey on Arabic Language Resources and Tools in the Mediterranean Countries. Technical Report, NEMLAR – Network for Euro-Mediterranean Language Resources.

End of Session 2

Thank You Very Much for your Concentration!!!

Questions and Answers Session

Lecture 3:
Implementation of ASR System

Presentation Outline

Lecture 3

- Introduction
- Implementation Requirements and Components
- Feature Extraction
- Phonetic Dictionary
- Acoustic Model Training
- Language Model Training
- The Decoder
- Performance Measures
- Experimental Results
- Conclusion and Future Work
- References

Introduction

- Development of speaker-independent automatic continuous speech recognition systems is a multi-disciplinary task, whereby the target language's phonetics, speech processing techniques and algorithms, and Natural Language Processing (NLP) are integrated, which result in improved and optimized performance of the developed systems.

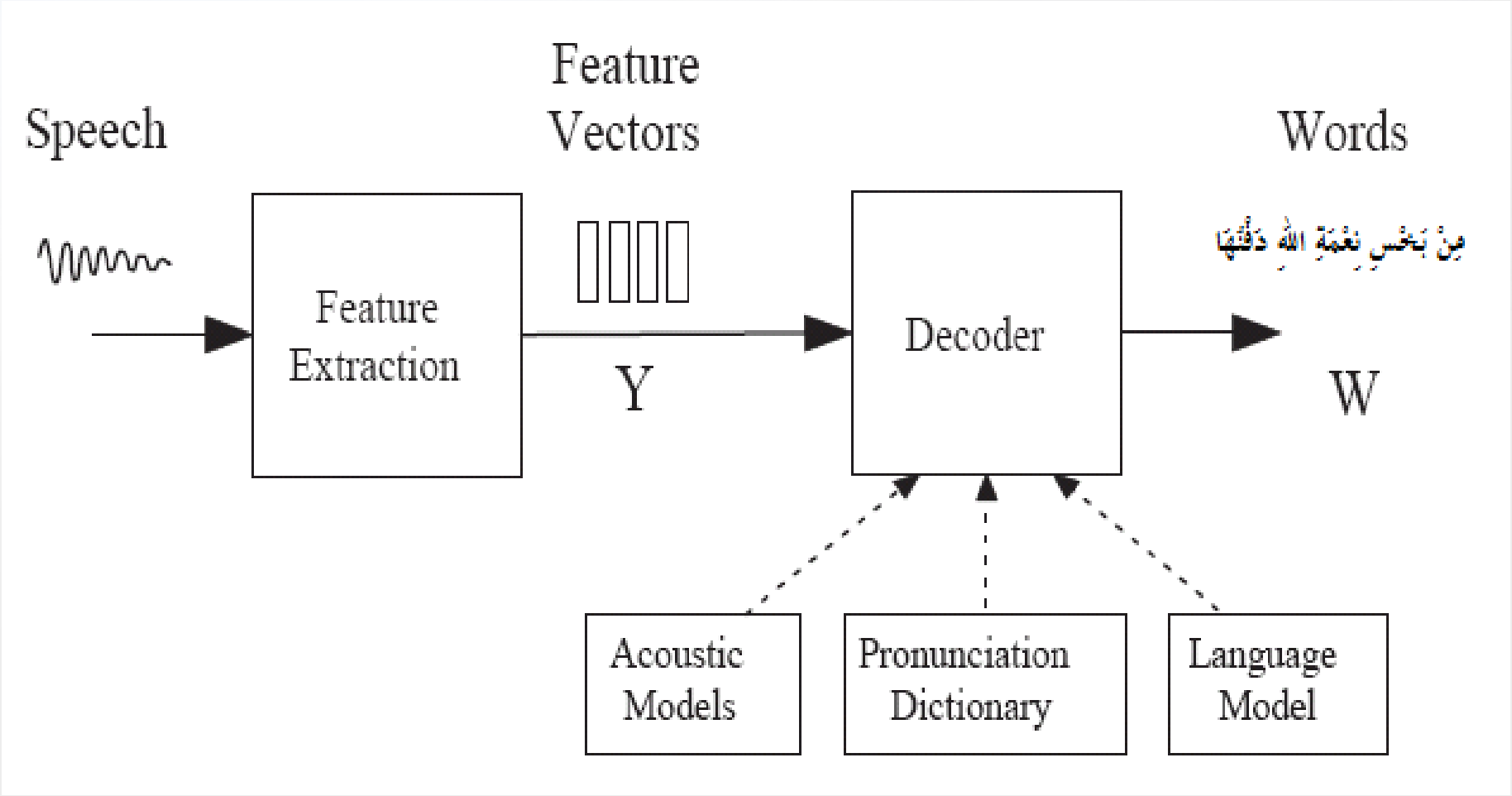
Implementation Requirements and Components

- The major implementation requirements and components for developing any speaker-independent automatic continuous speech recognition system consist of feature extraction, phonetic dictionary, the acoustic model training, and the statistical language model training, which are identified in the HMM-based architecture of the system as shown in Figure 2.

- This also complies with the architecture of the Carnegie Mellon University (CMU) Sphinx engine for ASR systems as shown in Figure 3.

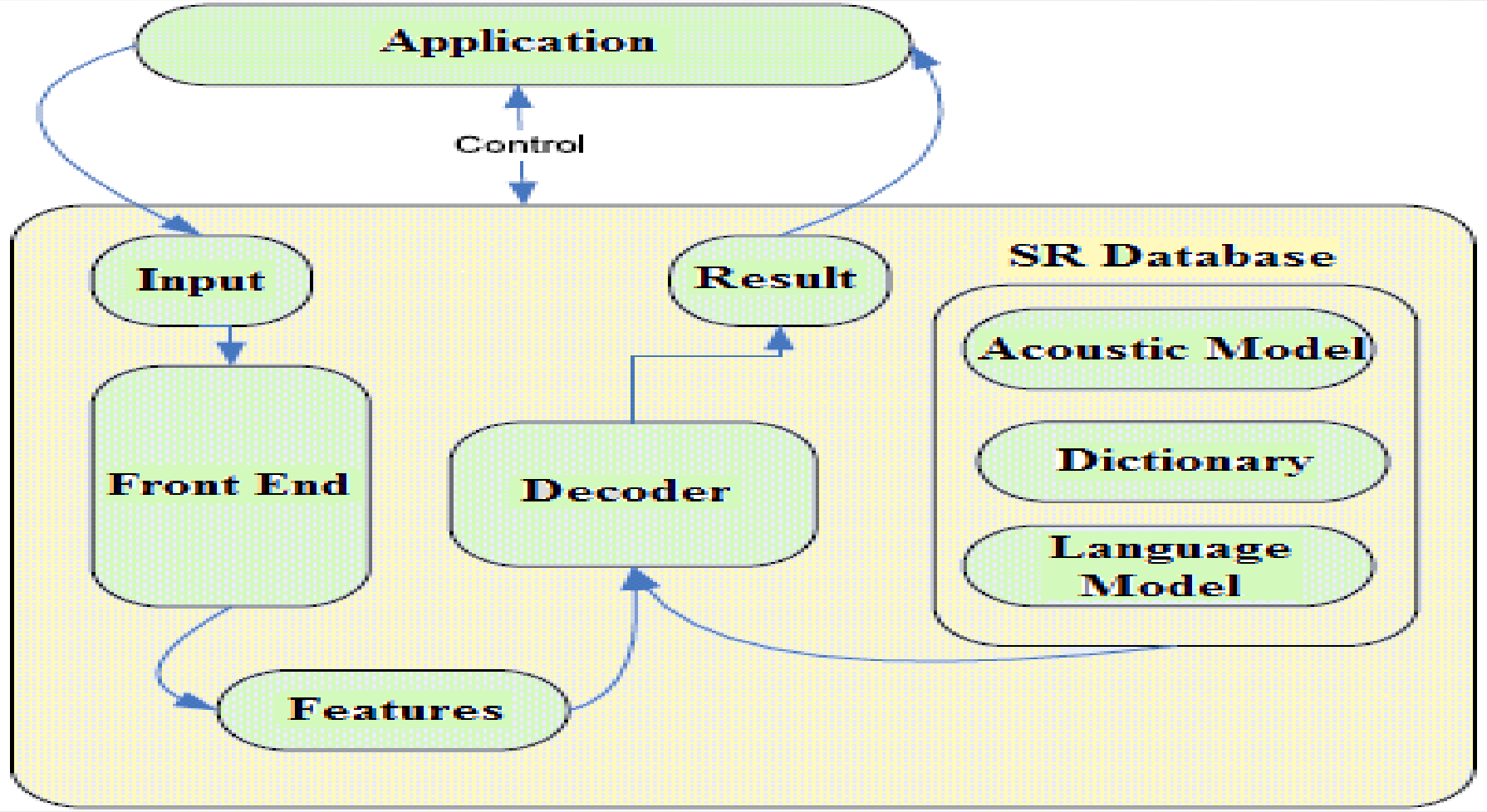
- The speech signal (input) shown in Figure 2 is represented by the phonetically rich and balanced Arabic speech corpus, which has been discussed in Session 2.

Implementation Requirements and Components



- **Figure 2**

Implementation Requirements and Components



- Figure 3

Implementation Requirements and Components

- The decoder is then used when all implementation requirements are created.

- It receives the new input features ***Y*** converted into a sequence of fixed size acoustic vectors at the feature extraction stage.

- It then attempts to identify the sequence of words ***W*** that is most likely to have generated ***Y***. Therefore, the decoder attempts to find (Gales and Young, 2008):

$$\hat{W} = \arg\max_{W} P(W \mid Y)$$

Implementation Requirements and Components

- The conditional probability P(W|Y) is difficult to be modeled directly, and therefore, Bayes' Rule is used in order to transform the equation (1) to an equivalent problem resulting in the following equation:

$$\hat{W} = \arg\max_{W} P(Y \mid W)P(W)$$

- The acoustic model is determined by the likelihood conditional probability P(Y|W) in order to observe a signal ***Y*** given a word sequence ***W*** was spoken, whereas the statistical language model is determined by the priori probability P(W) that word sequence ***W*** was spoken.

Implementation Requirements and Components

- ASR systems are expected to serve a large number of words; and therefore, each word has to be decomposed into a subword (phone) sequence.

- The acoustic model that corresponds to a given W is synthesized through concatenating the phone models in order to make words according to the way they are defined by the pronunciation dictionary.

- More details on these processes and components in line with the development of the Arabic speaker-independent automatic continuous speech recognition system are described in the following sections.

Feature Extraction

- Feature extraction, also referred to as belonging to front end component, is the initial stage of any ASR system that converts speech inputs into feature vectors in order to be used for training and testing the speech recognizer.
- The dominating feature extraction technique known as Mel-Frequency Cepstral Coefficients (MFCC) is used to extract features from the set of spoken utterances.
- The MFCC is also used in CMU Sphinx 3 tools (Chan et al., 2007) as the main feature extraction technique.
- As a result, a feature vector that represents unique characteristics of each recorded utterance is produced, which is considered as an input for training and testing the acoustic model.

Feature Extraction

- The main objective of feature extraction is to consider extracting characteristics from the speech signal that are unique, discriminative, robust and computationally efficient to each word which are then used to differentiate between different words (Ursin, 2002).

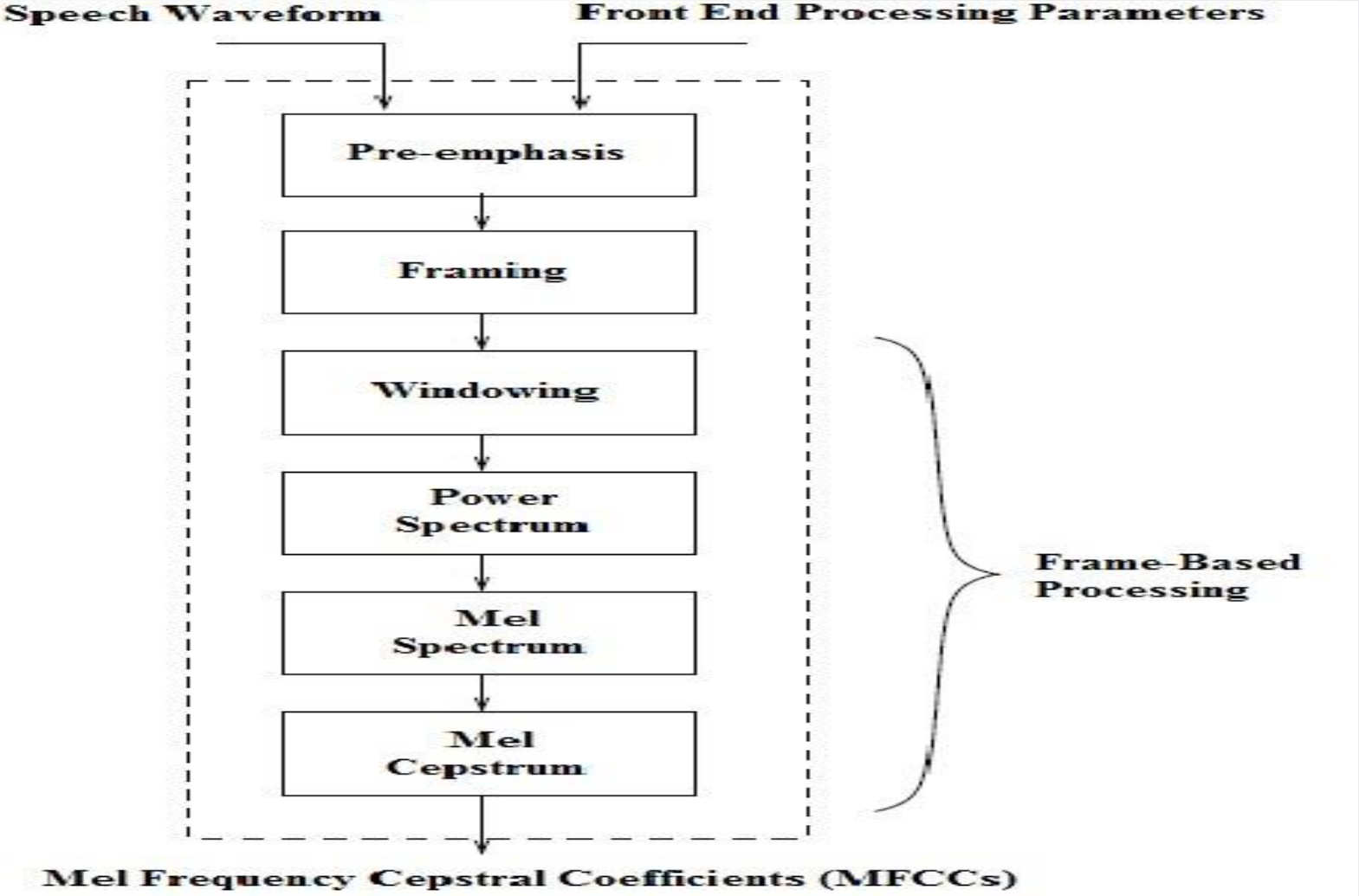
- **Table 1** summarizes the default parameters used for the computational process to perform the MFCC-based feature extraction algorithm as shown in Figure 4 (Chan et al., 2007).

Feature Extraction

Table 1: MFCC Feature Extraction Parameters

Parameter	Value
Sampling Rate	16,000 Hz
Frame Rate	100 Frames per Second
Window Length	0.0256 Second
Filter Bank Type	Mel-Frequency Filter Bank
Number of Cepstra	13
Number of Mel Filters	40
DFT Size	512
Lower Frequency	133.33 Hz
Upper Frequency	6855.5 Hz
Pre-emphasize	0.97
Dimension of the Basic MFCC Feature Vector	13
Dimension of the Overall Feature Vector	39

Feature Extraction



- Figure 4

Phonetic Dictionary

- The pronunciation or phonetic dictionary is one of the key components of the modern large vocabulary ASR systems, which serves as an intermediary link between the acoustic model and the language model in ASR systems.

- A rule-based approach to automatically generate the Arabic phonetic dictionary for large vocabulary ASR systems based on a given transcription is used.

- This tool uses the classic Arabic pronunciation rules, common pronunciation rules of MSA, and morphologically driven rules.

- Arabic pronunciation follows certain rules and patterns when the text is fully diacritized.

PHONETIC DICTIONARY

- According to Ali et al. (2008), this tool helps in developing the Arabic phonetic dictionary through choosing the correct phoneme combination based on the location of the letters and their neighbors, and providing multiple pronunciations for words that might be pronounced in different ways such as the word (التَّمْرِ) and some other words that have different ways of pronunciation as shown in Figure 5.

التَّمْرِ E T AE M R IX
2)التَّمْرِ(E L T AE M R IX
3)التَّمْرِ(T AE M R IX
4)التَّمْرِ(L T AE M R IX

- **Figure 5**

PHONETIC DICTIONARY

- A detailed description of these rules and patterns can be found in the work of Elshafei (1991). Description of the development of this Arabic phonetic dictionary tool can be found in the work of Ali et al. (2008).

- In this work, the transcription file contains 2,110 words and the vocabulary list contains 1,626 unique words.
- The number of pronunciations in the developed phonetic dictionary is 2,482 entries.
- Refer **Figure 6**

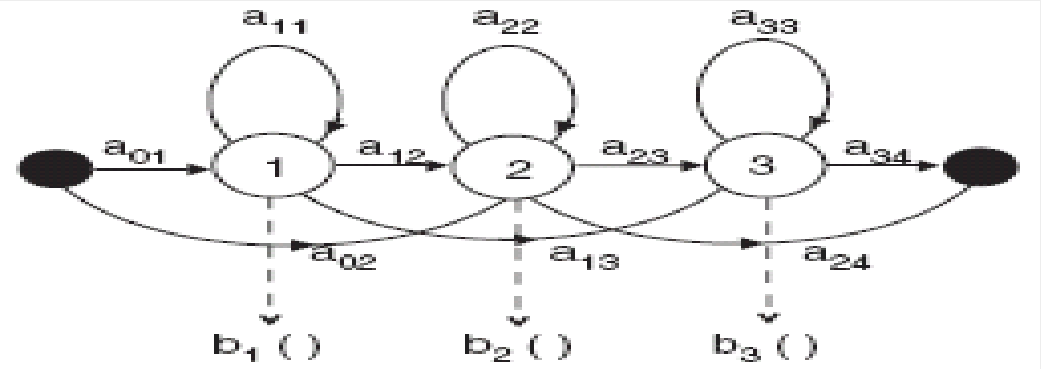
Phonetic Dictionary

آلَامُ E AE: L AE: M UH
آمِنٍ E AE: M IH N IH N
آيَاتُ E AE: Y AE: T UH
أَبَدَ E AE B AE D AE
أَبِي E AE B IY
أَبْجَلَنِي E AE B JH AE L AE N IY
أَبْطَأَ E AE B TT AH E AE
أَبْلَجُ E AE B L AE JH UH
أَتْهَمَ E AE T H AE M AE
أَثْجَاهُ E AE TH JH AE: H UH
أَثْخَنُوا E AE TH KH AA N UW

- **Figure 6**

Acoustic Model Training

- The acoustic model component provides the Hidden Markov Models (HMMs) of the Arabic tri-phones to be used in order to recognize speech.
- The basic HMM structure known as Bakis model as shown in **Figure 7**, has a fixed topology consisting of five states with three emitting states for tri-phone acoustic modeling (Rabiner, 1989; Bakis, 1976).

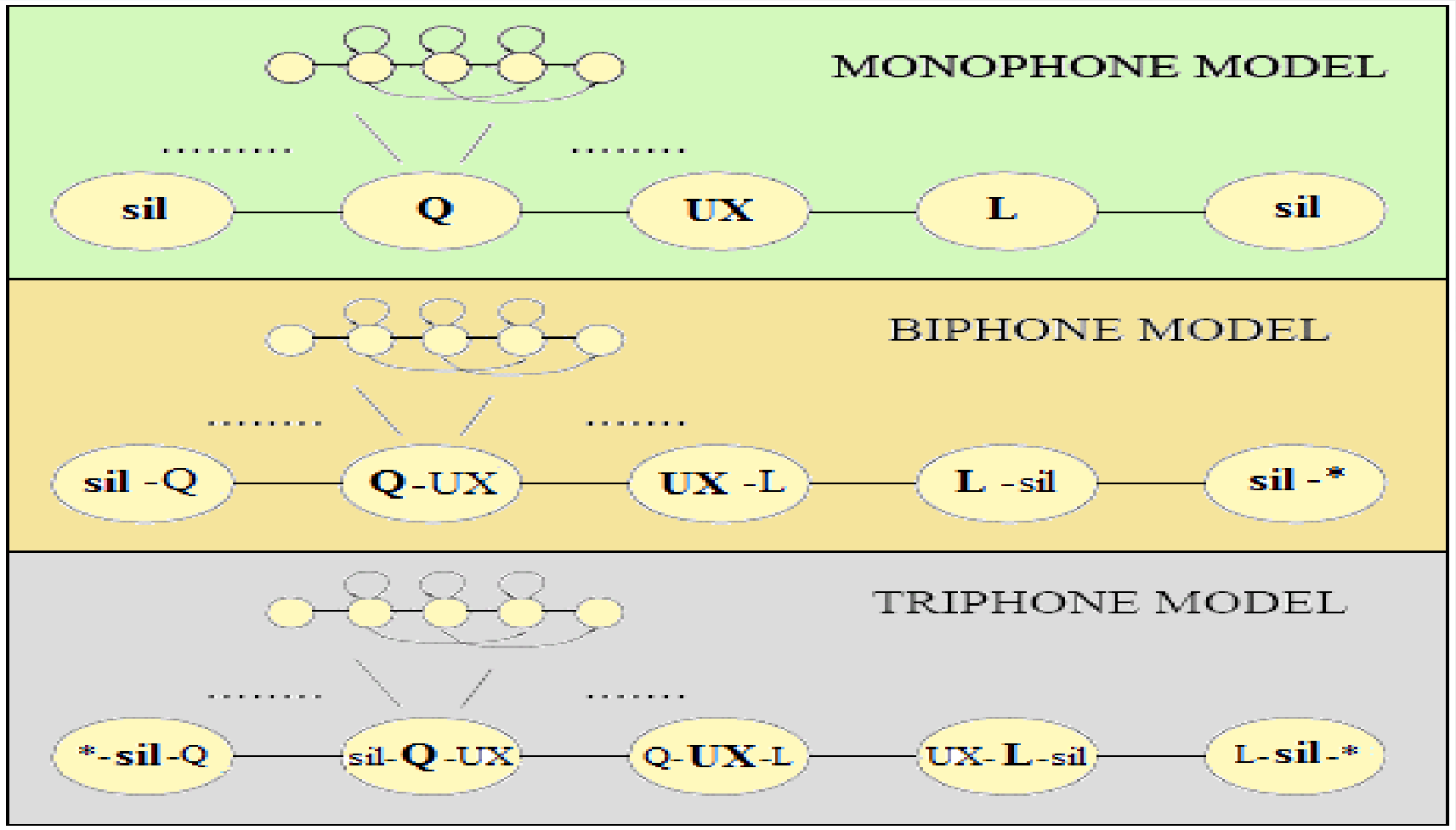


- **Figure 7**

Acoustic Model Training

- In order to build a better acoustic model, CMU Sphinx 3 (Placeway et al., 1997) uses tri-phone based acoustic modeling.
- A tri-phone not only models an individual phoneme, but it also captures distinct models from the surrounding left and right phones.
- **Figure 8** illustrates the concept of tri-phone acoustic modeling as compared to mono-phone and bi-phone acoustic modeling for the Arabic word (قُلْ) /Q/ /UX/ /L/, which means (Say) in English.

Acoustic Model Training

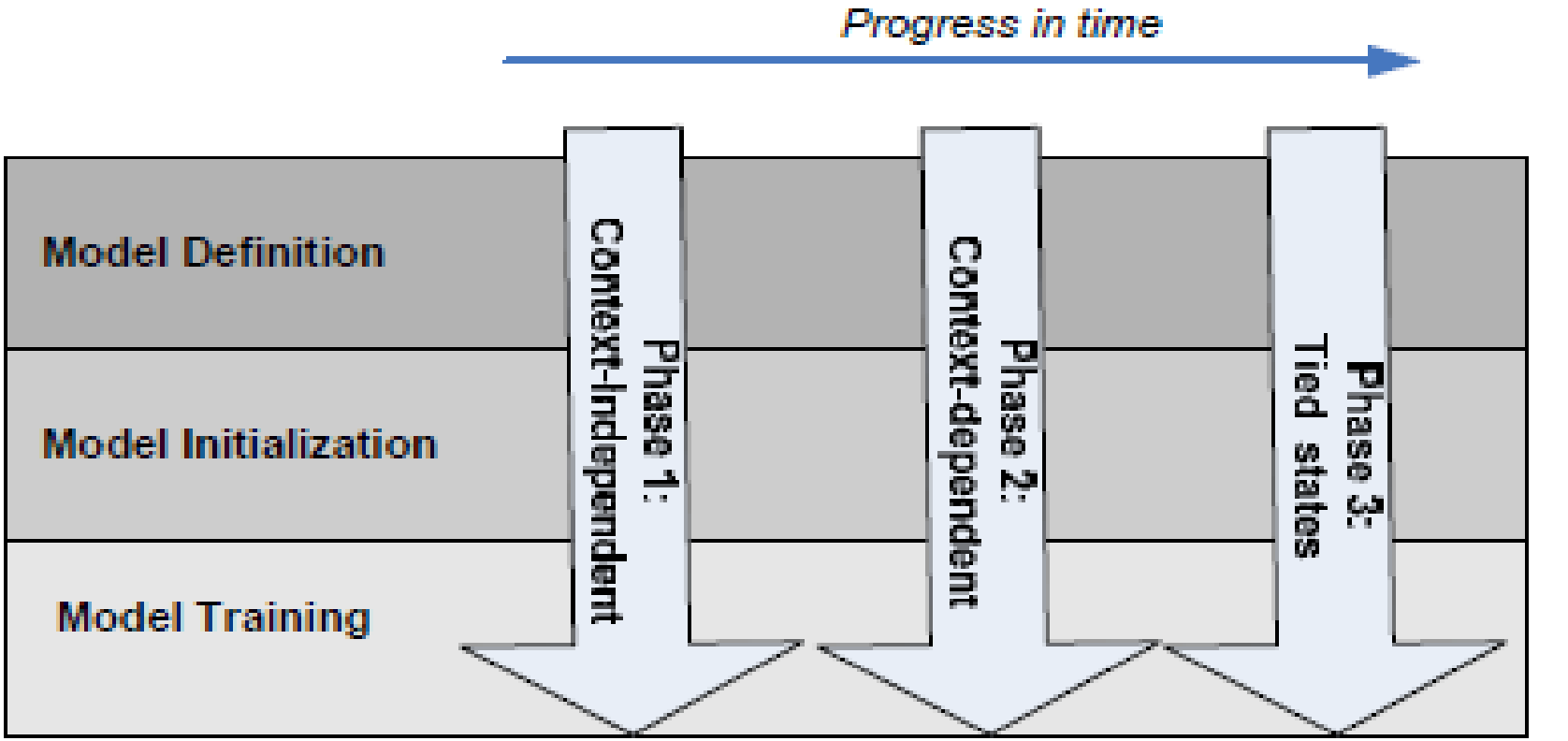


- **Figure 8**

Acoustic Model Training

- Continuous Hidden Markov Model (CHMM) technique is also supported in CMU Sphinx 3 for parametrizing the probability distributions of the state emission probabilities.
- Training the acoustic model using CMU Sphinx 3 tools requires successfully passing through three phases of Context-Independence (CI), Context-Dependence (CD), and Tied States, whereby each phase consists of three main steps, which are 1) model definition, 2) model initialization, and 3) model training, as shown in **Figure 9** (Rabiner, 1989; Alghamdi et al., 2009).

Acoustic Model Training



- Figure 9

Acoustic Model Training

- Baum-Welch re-estimation algorithm is used during **the first phase** in order to estimate the transition probabilities of the Context-Independent (CI) HMMs.
- Arabic basic sounds are classified into phonemes or phones as shown in Table 10.
- In this work, 44 (including silence) Arabic phonemes and phones are used.
- During **the second phase**, Arabic phonemes and phones are further refined into Context-Dependent (CD) tri-phones.
  - The HMM model is now built for each tri-phone, where it has a separate model for each left and right context for each phoneme and phone.
  - As a result of the second phase, tri-phones are added to the HMM set.

Acoustic Model Training

- In the **Tied-States phase**, the number of distributions is reduced through combining similar state distributions (Rabiner, 1989; Rabiner and Juang, 1993).

- In my work, there are four development phases of the Arabic ASR system, each of which differs in the size of the training and testing data.
- The acoustic model for each development phase has also undergone several training attempts aiming to identify the best combination of parameters in order to optimize the performance of the Arabic ASR system.
- Two main types of systems are developed using default and modified values of parameters, which are discussed in the following sub-sections.

Acoustic Model Training

- **Acoustic Model Training Using Parameters' Default Values:**
- The acoustic model is first trained using default values of major parameters as identified in Sphinx 3 configuration file, which are as follows:
  - **Number of Gaussian mixture distributions** is 8, which represents the maximum distribution that can be reached during the re-estimation of the senones.
  - **Number of senones is 1000**. A senone is also called a tied-state and is obviously shared across the tri-phones which contribute to it.

Acoustic Model Training

- **Acoustic Model Training Using Parameters' Modified Values:**
- The systems' performance using CMU Sphinx 3 default values may not necessarily be the best.
  - Therefore, different values must be examined in order to find the best combination that yields the best performance in terms of word recognition correctness rate (WRCR) as well as word error rate (WER).

- For each of the four development phases, the first experiment is always used to identify the best combination of values at training level.

- Such values are then applied for the rest of the experiments.

Acoustic Model Training

- **Acoustic Model Training Using Parameters' Modified Values:**
- In order to identify the best combination of Gaussian mixture distributions and senones at training level, 54 experiments were conducted.

- Gaussian mixture distributions range from 2 to 64, whereas senones range from 300 to 2500.

Language Model Training

- The language model component provides the grammar used in the ASR system.
- The grammar's complexity depends on the system to be developed.
- The language model computes the probability P(W) of a sequence of words $W = w_1, w_2, \ldots, w_L$.
- The probability P(W) can be expressed as follows:

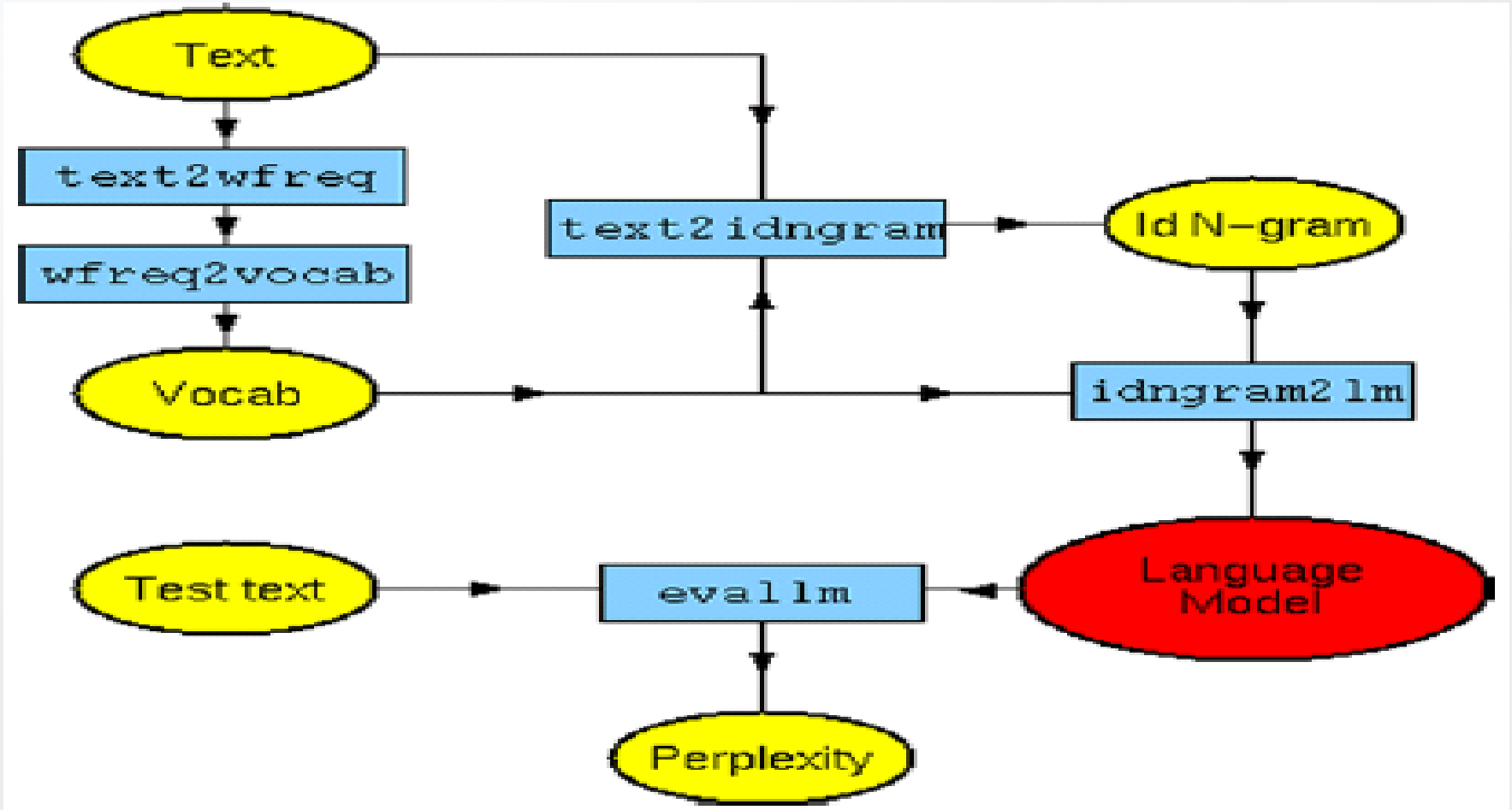
$$P(W) = P(w_1, w_2, \ldots, w_L) = \prod_{i=1}^{L} P(w_i \mid w_1, \ldots, w_{i-1})$$

- In my PhD work, the language model is built statistically using the CMU-Cambridge Statistical Language Modeling toolkit, which is based on modeling the uni-grams, bi-grams, and tri-grams of the language for the subject text to be recognized (Clarkson and Rosenfeld, 1997).

Language Model Training

- Creation of a language model consists of computing the word uni-gram counts, which are then converted into a task vocabulary with word frequencies, generating the bi-grams and tri-grams from the training text based on this vocabulary, and finally converting the N-grams into a binary format language model and standard ARPA format as illustrated in Figure 10 (Alghamdi et al., 2009).

Language Model Training

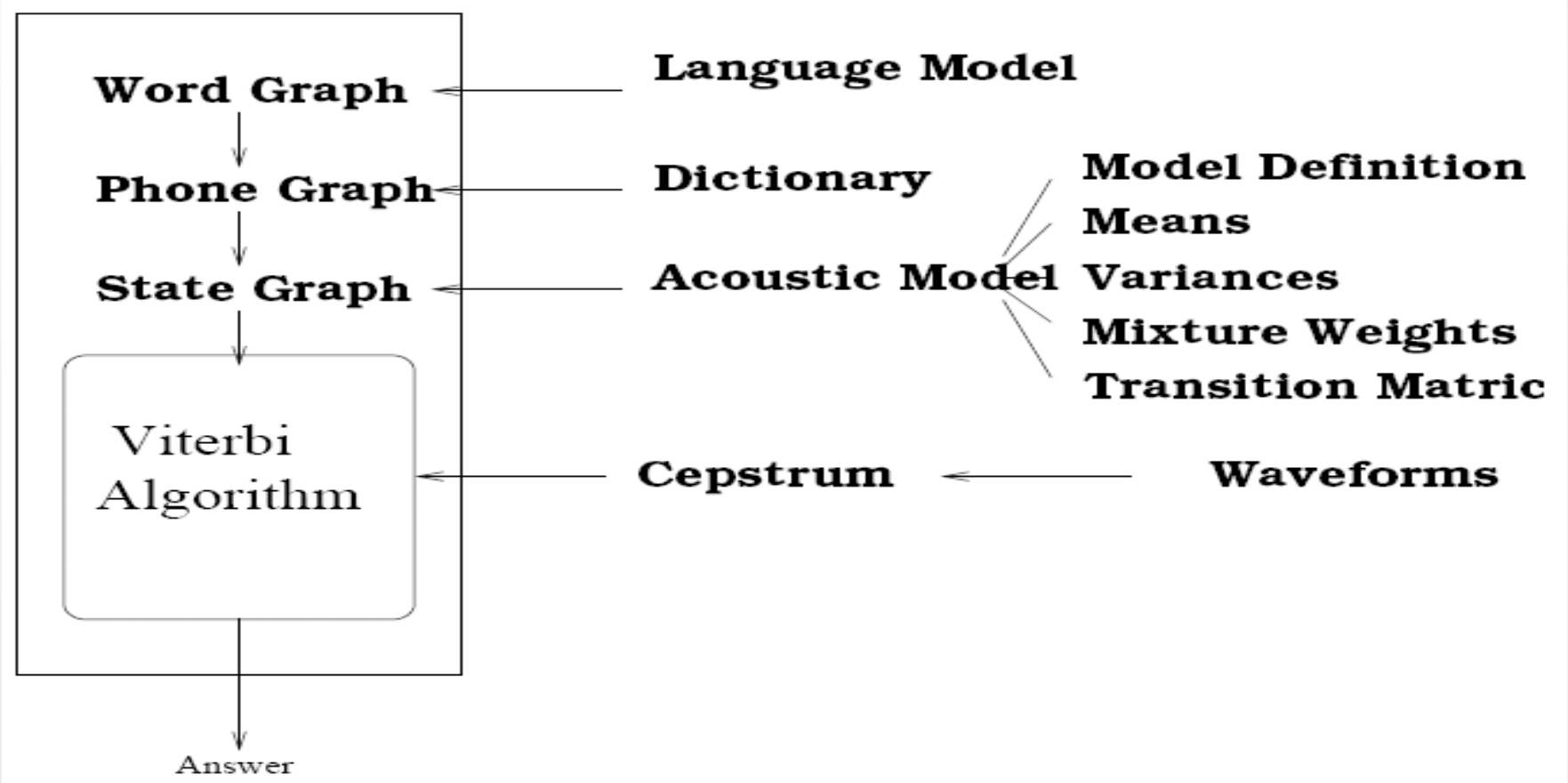


- **Figure 10**

The Decoder

- This work is based on the conventional Viterbi search algorithm and beam search heuristics, which are available in CMU Sphinx 3 decoder.
- It uses a lexical-tree search structure.
- The decoder requires certain inputs and resources such as the acoustic model, language model, phonetic dictionary, and feature vector of the unknown utterance as shown in Figure 11.
- The result is a recognition hypothesis, which is a single best recognition result for each utterance processed.
- It is a linear word sequence, with additional attributes such as their time segmentation and scores (Chan et al., 2007).

The Decoder



- **Figure 11**

The Decoder

- There are two (default and modified) versions of the decoder used in this work to examine possible combination of Word Insertion Penalty (WIP), Language Model Weight (LW), and Beam Pruning (BP) that take place at decoding (recognition) level.

- **Decoding Using Parameters' Default Values:**
- The Arabic ASR systems are first tested using default values of the CMU Sphinx 3 decoder, which are as follows:
  - Word Insertion Penalty (WIP) is **0.7**
  - Language Model Weight (LW) is **9.5**
  - Beam Pruning (BP) is **1.0e-35**

The Decoder

- **Decoding Using Parameters' Modified Values:**
- The Arabic ASR systems are first tested using default values of the CMU Sphinx 3 decoder.

- For performance optimization, a modified version of the decoder is used in order to identify possible combinations of Word Insertion Penalty (WIP) ranging between **0.2** and **0.7**, Language Model Weight (LW) ranging between **8** and **11**, and Beam Pruning (BP) ranging between **1.e-40** and **1.e-85** that yields a higher word recognition correctness rate and lower WER compared to what the standard decoder could achieve.

- As a result, **160** iterations of the decoder were required at this initial stage.

The Decoder

- **Decoding Using Parameters' Modified Values:**
- However, it was found that the ranges are too broad and some results are even less than what the standard decoder used to achieve.
- Therefore, the modified decoder has a WIP ranging between **0.4** and **0.7**, LW remains the same ranging between **8** and **11**, and BP is fixed to be **1.e-85**.

Performance Measures

- Experimental work conducted as part of this research is evaluated using two main performance measures known as word recognition correctness rate (WRCR) and the word error rate (WER), which are computed using the formulae listed below:

  - $\text{Word Recognition Correctness Rate (WRCR)} = \frac{N-D-S}{N} \times 100\%$
  - $\text{Percent Accuracy} = \frac{N-D-S-I}{N} \times 100\%$
  - $\text{Word Error Rate (WER)} = 100\% - \text{Percent Accuracy} = \frac{D+S+I}{N} \times 100\%$

- where {*N*} is the total number of words in the reference transcriptions, {*D*} is the number of deletion errors, {*I*} is the number of insertion errors, and {*S*} is the number of substitution errors, which are resulted when comparing the recognized word sequence with the reference (spoken) word sequence.

Experimental Results

- Based on the distribution of the speech corpus, the first experiment's data sets are used to identify the optimal combination of the modified values for the development phases.
- Based on the range values of the number of Gaussian mixture distributions (G) = {2, 4, 8, 16, 32, and 64}, and number of senones = {300, 350, 400, 450, 500, 1000, 1500, 2000, and 2500}, 54 different combinations are produced, each of which corresponds to a unique experiment.
- Experiments conducted here are initial and meant for identifying the best combination in order to apply it for other experimental data sets.

Experimental Results

Table 2: Systems' Performance Using CMU Sphinx 3 Default Values

Development Phase	Same Speakers with Different Sentences	
	WRCR (%)	WER (%)
1st Phase (4 Hours)	87.53	21.64
2nd Phase (8 Hours)	89.79	16.19
3rd Phase (11 Hours)	90.57	14.12
4th Phase (38 Hours)	91.53	12.57

Table 3: Average Best Performing Experiments for the Four Development Phases

Development Phase	Same Speakers with Different Sentences	
	WRCR (%)	WER (%)
1st Phase (4 Hours)	88.74	18.68
2nd Phase (8 Hours)	90.47	15.08
3rd Phase (11 Hours)	91.96	12.14
4th Phase (38 Hours)	92.40	11.35

Experimental Results

Table 4: Best Combinations of Gaussian Mixture Distributions and Senones

Development Phase	Gaussian Mixture Distributions	Senones	Same Speakers with Different Sentences	
			WRCR (%)	WER (%)
1st Phase (4 Hours)	16	400	91.23	14.37
2nd Phase (8 Hours)	16	500	93.24	10.73
3rd Phase (11 Hours)	16	300	93.63	8.89
4th Phase (38 Hours)	64	350	94.73	7.42

Experimental Results

Table 5: Summary of the Systems' Performance for the Four Development Phases Based on Modifications at Training Level

Development Phase	Same Speakers with Different Sentences		Different Speakers with Same Sentences		Different Speakers with Different Sentences	
	WRCR (%)	WER (%)	WRCR (%)	WER (%)	WRCR (%)	WER (%)
1st Phase (4 Hours)	91.23	14.37	89.42	16.71	80.83	25.88
2nd Phase (8 Hours)	92.67	11.27	95.92	5.78	89.08	15.59
3rd Phase (11 Hours)	93.38	9.38	97.57	3.37	92.54	10.40
4th Phase (38 Hours)	94.52	7.64	98.50	2.22	94.27	7.82

Experimental Results

Table 6: Summary of the Systems' Performance for the Four Development Phases Based on Modifications at Testing/Decoding Level

Development Phase	Same Speakers with Different Sentences		Different Speakers with Same Sentences		Different Speakers with Different Sentences	
	WRCR (%)	WER (%)	WRCR (%)	WER (%)	WRCR (%)	WER (%)
1st Phase (4 Hours)	92.10	12.54	93.04	13.37	84.35	22.84
2nd Phase (8 Hours)	93.86	9.70	96.91	4.58	91.48	12.39
3rd Phase (11 Hours)	94.32	8.10	98.10	2.67	93.73	8.75
4th Phase (38 Hours)	95.56	6.14	98.81	1.81	95.39	6.39

Experimental Results

Table 7: Comparison of Systems' Performance Based on Automatically and Manually Generated Linguistic Questions Sets

Development Phase	Automatically Generated Linguistic Questions Set		Manually Generated Linguistic Questions Set	
	WRCR (%)	WER (%)	WRCR (%)	WER (%)
1st Phase (4 Hours)	91.23	14.37	90.75	14.29
2nd Phase (8 Hours)	93.24	10.73	91.08	13.73
3rd Phase (11 Hours)	93.63	8.89	93.63	9.03
4th Phase (38 Hours)	94.73	7.42	94.71	7.39

Experimental Results

Table 8: Comparison of Male and Female Speakers Performance in the Fourth Development Phase Based on ٣8 Hours after Performance Optimization

Speaker's Gender	Experiment_ID	Different Speakers with Same Sentences		Different Speakers with Different Sentences	
		WRCR (%)	WER (%)	WRCR (%)	WER (%)
Male	Experiment3	97.38	4.25	96.23	8.06
	Experiment7	98.69	2.15	96.82	5.07
	Experiment10	98.53	2.60	95.56	7.66
	Experiment13	97.74	4.24	91.80	13.56
	Experiment14	99.27	1.07	94.96	6.39
	Experiment15	98.85	1.90	94.50	6.66
	Experiment16	99.46	0.63	96.59	4.24
	Experiment17	98.84	1.61	93.01	13.85
	Experiment19	98.89	1.42	95.25	6.17
	Experiment21	98.99	1.30	96.20	4.89
	Experiment22	97.43	5.91	92.87	11.63
	Experiment23	99.32	0.82	96.79	4.12
	Experiment24	99.47	0.62	96.32	4.37
	Experiment25	99.16	0.86	97.68	3.21
	Experiment26	97.59	3.89	93.05	10.66
	Experiment33	98.58	2.00	93.63	8.99
	Experiment34	97.89	3.40	92.68	10.89
	Experiment35	98.02	2.61	96.15	4.15
	Average Result	**98.56**	**2.29**	**95.01**	**7.48**

Experimental Results

Table 8: Comparison of Male and Female Speakers Performance in the Fourth Development Phase Based on 38 Hours after Performance Optimization

Female	**Experiment1**	99.06	1.24	97.62	2.66
	Experiment2	97.83	2.91	91.44	9.40
	Experiment4	99.59	0.50	97.88	2.40
	Experiment5	99.28	0.85	95.05	6.11
	Experiment6	99.34	0.74	96.37	3.80
	Experiment8	99.54	0.59	96.70	4.76
	Experiment9	99.16	1.14	97.53	3.21
	Experiment11	99.48	0.64	95.60	5.24
	Experiment12	99.85	0.23	96.54	4.65
	Experiment18	99.13	0.94	96.33	4.83
	Experiment20	99.37	0.72	93.43	7.64
	Experiment27	99.37	0.69	97.75	2.47
	Experiment28	97.27	4.92	93.69	8.65
	Experiment29	99.04	2.14	95.42	6.87
	Experiment30	98.95	1.44	95.36	7.02
	Experiment31	98.91	1.63	94.07	8.09
	Experiment32	99.03	1.18	95.72	4.41
	Experiment36	98.91	1.47	97.58	3.09
	Average Result	**99.06**	**1.33**	**95.78**	**5.29**

Experimental Results

Table 9: Effect of Speakers' Country and Region on the Overall Systems' Performance after Performance Optimization

Region	Country	Based on Country				Based on Region			
		WRCR (%)		WER (%)		WRCR (%)		WER (%)	
		Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.
Gulf	**Iraq**	96.81	0.89	4.58	1.73	**97.37**	**0.79**	**3.38**	**0.99**
	Saudi Arabia	97.87	0.46	2.69	0.35				
	Yemen	96.55	1.79	3.97	1.89				
	Oman	98.25	0.00	2.28	0.00				
Africa	**Sudan**	97.00	1.78	4.92	3.52	**96.77**	**0.81**	**4.98**	**1.50**
	Algeria	96.90	1.46	4.35	2.46				
	Egypt	97.05	0.00	5.13	0.00				
	Morocco	96.11	0.00	5.50	0.00				
Levant	**Jordan**	97.10	1.11	4.07	2.03	**97.71**	**0.40**	**3.00**	**0.69**
	Palestine	97.96	0.10	2.47	0.05				
	Syria	98.06	0.00	2.47	0.00				

Experimental Results

Table 10: Effect of Speakers' Age on the Overall Systems' Performance after Performance Optimization

Age Category	Male Speakers				Female Speakers				Ratio (Age Category)			
	WRCR (%)		WER (%)		WRCR (%)		WER (%)		WRCR (%)		WER (%)	
	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.
Less Than 30 Years	97.47	0.72	3.14	0.98	97.60	1.22	2.87	1.40	**97.54**	**0.97**	**3.01**	**1.19**
30 Years and Above	96.44	1.31	5.76	2.51	97.07	1.08	4.20	1.85	**96.76**	**1.20**	**4.98**	**2.18**

Experimental Results

- The experimental results show that training data play a crucial role in enhancing and improving the performance of speech recognition systems as they are considered the major contributor to improved systems' performance.

- It is found that modified systems for the four development phases perform better than the base systems using standard and default CMU Sphinx 3 setup.

- Therefore, it is advisable to try different combinations of parameters in order to identify the best combination that is more suitable to the data used in order to obtain better performance.

Experimental Results

- The modified decoder used at testing/decoding level using different combination of Word Insertion Penalty (WIP), Language Model Weight (LW), and Beam Pruning (BP), achieved better performance than the standard CMU Sphinx 3 decoder.

- Therefore, it is important to look for the best combination of such key parameters in order to enhance the performance of the decoder and obtain better performance compared with the standard version based on default values of the parameters.

Experimental Results

- It is noticed that utterances collected from female speakers achieve better performance than that of the male speakers.
  - This is due to the fact that male and female speakers obviously differ in features and characteristics of the voice.

- In addition, the speakers' age was also examined in this work. It is found that speakers that are less than 30 years old outperformed speakers that are 30 years old and above.
  - This is due to vocal characteristics, whereby as the speaker grows older the vocal characteristics change and that obviously affects the speech recognition systems' performance. It is noticed that younger speakers have better vocal characteristics than the older speakers.

Experimental Results

- The effects of speakers' country and region have also been examined in this research work.
- Speakers living in the Levant region outperform the speakers living in Gulf and Africa regions although all of them have recorded in the MSA.
- Speakers from Africa region are influenced by their dialects even though they were asked to record in MSA, but the dialect is really influential especially for speakers from Sudan and Egypt.
- Therefore, the region where the speaker is located can affect the speech recognition systems' performance.

Experimental Results

- Speaker-independence is highly achieved in the fourth development phase.

- If we refer to Table 6, we can see that for the same speakers with different sentences, the systems obtain an average WRCR of 95.56% and an average WER of 6.14%, whereas for different speakers with different sentences they obtain an average WRCR of 95.39% and an average WER of 6.39%.

- This is important due to the fact that speech recognition systems must adhere to the differences between speakers. Obviously not all potential users can be used in training, therefore, the systems must be able to adapt to users who are not being used in training the systems

Conclusion and Future Work

- In conclusion, modified parameters have generally shown the capability of increasing the word recognition correctness rates and reducing the WERs compared to default values of parameters used in the CMU Sphinx 3 tools.
- It is important to highlight that the framework used in this work, whereby the 3-emitting state Continuous Density Hidden Markov Model (CDHMM) for tri-phone based acoustic models is adopted, produces efficient speaker-independent, automatic, and continuous Arabic ASR systems.
- As a summary, this research has introduced and contributed unique language resources for MSA based on phonetically rich and balanced approaches gathered from speakers with different variabilities.

Conclusion and Future Work

- This study has also shown a complete modeling and development of all required components for a speaker-independent, automatic, and continuous based and modified ASR systems for MSA.
- In addition, a knowledge-based linguistic questions set is developed and compared against the CMU Sphinx 3 automatically generated questions set. Currently, there is not much difference in the performance of the two linguistic questions sets.
- This study has conducted a detailed examination on the effects of speakers' characteristics including the gender, age, country, and region on the overall systems' performance.

Conclusion and Future Work

- Finally, based on the experimental results of this work, it is found that the fourth development phase is the best phase that also narrows down the gender, country and region, and age differences between speakers for the modified version of the CMU Sphinx 3 decoder.

- Therefore, the fourth development phase can be considered as the best output of this research work as it is capable of successfully recognizing speech from different speakers with different variabilities.

Conclusion and Future Work

- Currently the speech corpus contains recordings of 36 native speakers from 11 Arab countries. This research can be enhanced further to include speakers from all the 21 Arab countries and possibly non-native Arabic speakers from any country in the world.
- The knowledge-based set can be improved and customized further in order to be used for training the acoustic models in Arabic ASR systems.
- For performance optimization, there could be other ways to optimize the systems' performance such as modifying the feature extraction parameters or even using a technique other than the MFCC, and making use of hybrid approaches to train the acoustic model.

References

- Elmahdy, M., Gruhn, R., Minker, W., and Abdennadher, S. (2009a). Survey on common Arabic language forms from a speech recognition point of view. *International Conference on Acoustics (NAG-DAGA)*, Rotterdam, Netherlands, pp. 63 – 66.
- Elmahdy, M., Gruhn, R., Minker, W., and Abdennadher, S. (2009b). Modern Standard Arabic Based Multilingual Approach for Dialectal Arabic Speech Recognition. *IEEE Proceedings of the Eighth International Symposium on Natural Language Processing*, Bangkok, Thailand, pp. 169 – 174.
- Cieri, C., Liberman, M., Arranz, V., and Choukri, K., (2006). Linguistic Data Resources. In Schultz, T., and Kirchhoff, K. (Eds.), *Multilingual Speech Processing* (pp. 33 – 70). USA, Academic Press, Elsevier.
- Uraga, E., and Gamboa, C. (2004). VOXMEX Speech Database: Design of a Phonetically Balanced Corpus. Proceedings of the 4th International Conference on Language Resources and Evaluation, Portugal, pp. 1471 – 1474.
- Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. University Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

References

- Black, A. W., and Tokuda, K. (2005). The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. *INTERSPEECH'05*, Portugal, pp. 77-80.
- D'Arcy, S., and Russell, M. (2008). Experiments with the ABI (Accents of the British Isles) Speech Corpus. *INTERSPEECH'08*, Australia, pp. 293 – 296.
- Chou, F. C., and Tseng, C. Y. (1999). The Design of Prosodically Oriented Mandarin Speech Database. *ICPhS'99*, San Francisco, pp. 2375 – 2377.
- Liang, M. S., Lyu, R. Y., and Chiang, Y. C. (2003). An Efficient Algorithm to Select Phonetically Balanced Scripts for Constructing a Speech Corpus. IEEE Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, China, pp. 433 – 437.
- Hong, H., Kim, S., and Chung, M. (2008). Effects of Allophones on the Performance of Korean Speech Recognition. *INTERSPEECH'08*, Australia, pp. 2410 – 2413.
- Rabiner, L. R., and Juang, B. H., (1999). Speech Recognition by Machine. In Madisetti, V. K., and Williams, D. B. (Eds.), *Digital Signal Processing Handbook* (pp. 987 – 1002). CRCnetBASE. CRC Press LLC.

References

- Huang, X., and Lee, K. F., (1993). On Speaker-Independent, Speaker-Dependent, and Speaker-Adaptive Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 2, pp. 150 – 157.
- Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R., and Vergyri, D., (2003). Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns-Hopkins Summer Workshop. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, Vol. 1, pp. 344 – 347.
- Soltau, H., Saon, G., Kingsbury, B., Kuo, H. K. J., Mangu, L., Povey, D., and Emami, A., (2009). Advances in Arabic Speech Transcription at IBM Under the DARPA GALE Program. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 5, pp. 884 – 894.
- Vergyri, D., and Kirchhoff, K., (2004). Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, pp. 66 – 73.

References

- Alotaibi, Y. A., and Hussain, A., (2010). Comparative Analysis of Arabic Vowels using Formants and an Automatic Speech Recognition System. *International Journal of Signal Processing, Image Processing, and Pattern Recognition*, Vol. 3, No. 2, pp. 11 – 22.
- Diehl, F., Gales, M. J. F., Tomalin, M., and Woodland, P. C., (2008). Phonetic Pronunciations for Arabic Speech-to-Text Systems. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, USA, pp. 1573 – 1576.
- Alghamdi M., Elshafei M., and Al-Muhtaseb H., (2009). Arabic Broadcast News Transcription System. *International Computer Journal of Speech Technology*, Vol. 10, No. 4, pp. 183 – 195.
- Nikkhou, M., and Choukri, K., "Survey on Industrial needs for Language Resources", Technical Report, NEMLAR – Network for Euro-Mediterranean Language Resources, 2004.
- Nikkhou, M., and Choukri, K., "Survey on Arabic Language Resources and Tools in the Mediterranean Countries", Technical Report, NEMLAR – Network for

References

- Al-Sulaiti, L., and Atwell, E. (2006). The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, John Benjamins Publishing Company, pp. 1 – 36.
- Elmahdy, M., Gruhn, R., Minker, W., and Abdennadher, S. (2009a). Survey on common Arabic language forms from a speech recognition point of view. *International Conference on Acoustics (NAG-DAGA)*, Rotterdam, Netherlands, pp. 63 – 66.
- Elmahdy, M., Gruhn, R., Minker, W., and Abdennadher, S. (2009b). Modern Standard Arabic Based Multilingual Approach for Dialectal Arabic Speech Recognition. *IEEE Proceedings of the Eighth International Symposium on Natural Language Processing*, Bangkok, Thailand, pp. 169 – 174.
- Alotaibi, Y. A., and Meftah, A. H. (2010). Comparative Evaluation of Two Arabic Speech Corpora. *IEEE Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China.
- Elshafei, A. M. (1991). Toward an Arabic Text-to-Speech System. *The Arabian Journal for Science and Engineering*, Vol. 16, No. 4B, pp.565 – 583.

References

- Alansary, S., Nagi, M., and Adly, N., (2007). Building an International Corpus of Arabic (ICA): Progress of Compilation Stage. *8th International Conference on Language Engineering*, Egypt.
- Alghamdi M., Alhamid A. H., and Aldasuqi M. M., (2003). Database of Arabic Sounds: Sentences. *Technical Report*, King Abdulaziz City of Science and Technology, Saudi Arabia. (In Arabic).
- Alghamdi M., Basalamah M., Seeni M., and Husain A., (1997). Database of Arabic Sounds: Words. *Proceedings of the 15th National Computer Conference*, pp. 797 – 815, Saudi Arabia. (In Arabic).
- Ali, M., Elshafei, M., Alghamdi, M., Almuhtaseb, H., and Al-Najjar, A., (2008). Generation of Arabic Phonetic Dictionaries for Speech Recognition. *IEEE Proceedings of the International Conference on Innovations in Information Technology*, UAE, pp. 59 – 63.
- Billa, J., Noamany, M., Srivastava A., Makhoul, J., and Kubala, F., (2002). Arabic Speech and Text in TIDES OnTAP. *Proceedings of the Second International Conference on Human Language Technology Research*, California, USA, pp. 7 – 12

References

- Chan, A., Gouvˆea, E., Singh, R., Ravishankar, M., Rosenfeld, R., Sun, Y., Huggins-Daines, D., Seltzer, M., (2007). The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Related Resources. http://www-2.cs.cmu.edu/~archan/documentation/sphinxDocDraft3.pdf , accessed 15 September 2010.
- Hyassat, H., and Abu Zitar, R., (2008). Arabic speech recognition using SPHINX engine. *International Journal of Speech Technology*, Springer, pp. 133 – 150.
- Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R., and Vergyri, D. (2003). Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop. *ICASSP'03*, Hong Kong, vol.1, pp. 344 – 347.
- Messaoudi, A., Gauvain, J. L., and Lamel, L., (2006). ARABIC BROADCAST NEWS TRANSCRIPTION USING A ONE MILLIONWORD VOCALIZED VOCABULARY. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, pp. 1093 – 1096.
- Mourtaga, E., Sharieh, A., and Abdallah, M., (2007). Speaker Independent Quranic Recognizer Based on Maximum Likelihood Linear Regression. *Proceedings of World Academy of Science, Engineering and Technology*, V. 36, p[illegible]61 [illegible]7, B[illegible]il.

References

- Parkinson, D. B., and Farwaneh, S. (Ed) (2003). *Perspectives on Arabic Linguistics XV*. John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 149 – 180.
- Pineda, L. V., Montes-y-Gómez, M., Vaufreydaz, D., and Serignat, J-F., (2004). Experiments on the Construction of a Phonetically Balanced Corpus from the Web. *5th International Conference on Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, Springer, Vol. 2945/2004, pp. 416 – 419, Korea.
- Satori, H., Harti, M., and Chenfour, N., (2007). Arabic Speech Recognition System Based on CMUSphinx. *IEEE Proceedings of ISCIII'07*, Morocco, pp. 31 – 35.
- Solatu, H., Saon, G., Kingsbury, B., Kuo, J., Mangu, L., Povey, D., and Zweig, G., (2007). THE IBM 2006 GALE ARABIC ASR SYSTEM. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, Hawaii, USA, pp. 349 – 352.

End of Lecture 3

Thank You Very Much for your Concentration!!!

Questions and Answers Session