Volume 9, Issue 3, March – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR598

Voice to Text Conversion using Deep Learning

[1]R. Azhagusundaram
(Assistant Professor)
[1]School of Computing, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India- 600073

[2]Ravipati Naveen; [3]Ravipati Ganesh
[4]Rambha Sivani; [5]Pragya Kumari Jha
(Student)
[2,3,4,5]School of Computing, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India- 600073.

Abstract:- Speech recognition is one of the quick developing engineering innovation. It has numerous applications in different areas, and offers numerous potential benefits. Numerous individuals might not communicate due to the dialect obstruction. Our objective is to diminish this boundary with our program planned and created to get to the framework in particular cases, giving crucial help in empowering individuals to share data by working the framework utilizing voice input May. This venture takes that calculate under consideration and endeavors to guarantee that our program recognizes discourse and changes over the input sound to content; This empowers the client to perform record operations such as spare, open, or select out of voice-only input. We plan a framework that recognizes human voices and sound clips and interprets between English and English. The yield is in content arrange and we offer choices to change over the sound from one dialect to another. Following, we trust to include a work that gives word reference implications for English words. Neural machine interpretation is the essential strategy utilized to perform machine interpretation within the industry. This work on discourse acknowledgment starts with an presentation to the innovation and its applications in different areas. Portion of the report is based on computer program enhancements in speech recognition.

Keywords:- *Speech Recognition, Communicate, Input, Text, Language, Neural Machine Translation.*

I. INTRODUCTION

The ability of a computer or program to recognize words and phrases in spoken language and translate them into a format that is readable by machines is known as speech recognition. There are numerous speech recognition applications available today, including speech-to-text, voice dialing, and basic data entry. Numerous distinct components from a wide range of fields, including statistical pattern recognition, communication theory, signal processing, combinatorial mathematics, and linguistics, are used in automatic voice recognition systems. An alternative to conventional computer interaction techniques, such as text entry via a keyboard, is speech recognition. Attempts to develop an effective system that could replace or reduce the reliability of standard keyboard input, automatic speech recognition (ASR) systems, were first attempted in the 1950s. These early speech recognition systems attempted to use a set of grammatical and syntactic rules for speech recognition. If the spoken words conform to a certain rule, the computer can recognize the words. However, human language has many exceptions to its rules. The way words and phrases are spoken can be substantially changed by accent, dialects, and customs. Therefore, we use algorithms to obtain ASR. In present day civilized societies, the foremost common strategy of communication between people is speech. Various thoughts shaped within the mind of the speaker are communicated through discourse within the shape of words, expressions and sentences by applying certain redress linguistic rules. Speech is the essential medium of communication between people, and discourse is the foremost characteristic and proficient frame of communication between people. By classifying speech by sounds, sounds, and silences (VAS/S), we can consider the basic acoustic segmentation required for speech. By following individual sounds called phonemes, this technique closely resembles the sounds of each letter of the alphabet that make up the structure of human speech. The main objective of speech recognition is to generate a set of words from a sound signal received from a microphone or telephone. Computers must be used to extract and determine the linguistic information communicated by speech waves.

- *Problem Statement*

Our project's primary goals are to promote the usage of our original tongue and assist those who are illiterate in typing text more easily. The idea is based on voice recognition using a microphone. By placing a noise filter cap over the microphone, you may lower the background noise. A combination of feature extraction and artificial intelligence is used to extract the words from the input speech. We translate the expression using NLTK, and word tokenizer is used to identify these words in the spoken voice. Data analysis is then used to compare the extracted words with the pre-trained data set.

II. LITERATURE SURVEY

The most important step in the software development process is the literature survey. Prior to developing the device, the time factor, economics, and company strength must be ascertained. The next stage is to decide which operating system and language can be used for tool development once these requirements are met. The programmers require a great deal of outside assistance once

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR598

they begin developing the tool. You can get this support from websites, books, and senior programmers. The aforementioned factors are taken into account before constructing the suggested system. A significant portion of the project development industry takes into account and thoroughly examines every demand that is necessary to produce the project. The most crucial step in the software development process is the literature review for each project. It is vital to ascertain and assess the time factor, resource demand, workforce, economy, and organizations strength prior to generating the tools and the related designs. The next step is to ascertain the software specifications in the corresponding system, such as what kind of operating system the project would require and what are all the necessary software are needed to proceed with the next step, such as developing the tools and the associated operations, once these items have been satisfied and thoroughly surveyed.

- *Real Time Text-to Speech Conversion System for Spanish*

The goal of the research is to create a text-to-speech converter (TSC) for Spanish that can process up to 250 words per minute of continuous alphanumeric input and output authentic, high-quality Spanish. This work considers four sets of problems: the linguistic processing rules, the parametrization of the Spanish language matched to a TSC, the sophisticated control software required to manage the orthographic input and linguistic programs, and the hardware structure chosen for real-time operation. The difficulties of fitting a general hardware framework into a particular language are highlighted.

- *Hindi-English Speech-to-Speech Translation System for Travel Expressions*

Speech signals in source language A can be translated into target language B using a speech-to-speech translation system. The capacity of a speech-to-speech translation (S2ST) system to preserve the original speech input's meaning and fluency is a defining characteristic. The suggested effort aims to provide translation between Hindi and English using an S2ST system. For this approach, a preliminary dataset focused on fundamental travel terms in both the languages under consideration is used. Three subsystems are needed to create a good S2ST system: text-to-speech synthesis (TTS), machine translation (MT), and automatic speech recognition (ASR).For both languages, an ASR system based on hidden Markov models is created, and word error rate (WER) is used to examine the performances of the languages. The statistical machine translation (SMT) technique is employed by the MT subsystem to translate the text between the two languages. To facilitate accurate translation, the SMT uses IBM alignment models and language models. Based on the translation table analysis and translated edit rate (TER), MT performance is evaluated. The translated text is synthesized using an HMM-based speech synthesis system (HTS). Based on a group of listeners' mean opinion score (MOS), the synthesizer's performance is examined.

- *Acoustic Modeling Problem for an Automatic Speech Recognition Systems*

The voice signal is recorded, parameterized, and assessed at the front end of automated speech recognition (ASR) systems utilizing the hidden Markov model (HMM) statistical framework. These systems' performance is highly dependent on the models that are employed as well as the signal analysis techniques that are chosen. To get over these restrictions, researchers have suggested a number of additions and changes to HMM-based acoustic models. We have compiled the majority of the HMM-ASR research conducted over the past three decades in this study. We organize all of these techniques into three groups: traditional techniques, HMM improvements, and refinements. The review is divided into two sections, or papers: (i) A synopsis of standard techniques for phonetic acoustic modeling; (ii) Improvements and developments in acoustic models. The construction and operation of the typical HMM are examined in Part I along with some of its drawbacks. It also covers decoders, language models, and various modeling units. Part II reviews the developments and improvements made to the traditional HMM algorithms as well as the difficulties and performance problems that ASR is currently facing.

- *Speech Recognition from English to Indonesian Language Translator using Hidden Markov Model*

Transmission of knowledge is a process that requires communication. Humans communicate with one another through language. However, there are several languages spoken throughout the world, thus not everyone can communicate in the same language. This study presents a language translation system. Based on speech recognition using feature extraction using Mel Frequency Cepstral Coefficients (MFCC) and the Hidden Markov Model (HMM) classification algorithm, the translated language is English to Indonesian.

- *Automatic Speech-Speech Translation form of English Language and Translate into Tamil Language*

In order to avoid the challenges that arise when people converse with someone who do not understand their language or another language, most people nowadays record their interactions and manually translate them into another language utilizing existing systems and techniques. This research aims to mechanically translate English speech into Tamil using speech recognition technology. This gadget is isolated into three segments: the discourse acknowledgment gadget, the English to Tamil machine interpretation, and the Tamil discourse generation. The English speech is first recognized by the speech reorganization device, which then displays the English text on the screen. Next, the Tamil text is translated and displayed, and finally the Tamil speech is generated, which should be audible on the other end of the device.

- *Multilingual Speech Recognition Speech-to-Speech Translation System for Mobile Consumer Devices*

Voice-to-voice translation technology is no longer a research topic because it has become widely used by users due to the development of machine translation and speech

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR598

recognition technologies as well as the rapid spread of mobile devices. However, in addition to enhancing the fundamental features within the experimental setting, the system must take into account a variety of utterance characteristics made by the users who will actually use the speech-to-speech translation system in order to be developed into a widely usable tool. After mobilizing a large number of people based on the survey on user requests, this study has produced a big language and voice database closest to the environment where speech-to-speech translation devices are actually being used. This work allowed for the achievement of outstanding baseline performance in a setting more akin to a speech-to-speech translation environment than merely an experimental one. Furthermore, a user-friendly user interface (UI) has been created for speech-to-speech translation. Additionally, errors were minimized during the translation process by implementing several techniques aimed at improving user satisfaction. Following the usage of the genuine administrations, a sifting method was utilized to apply the expansive database assembled through the benefit to the framework in arrange to get the most excellent potential strength toward the environment and points of interest of the users' articulations. By utilizing these measurements, this study aims to uncover the forms included within the improvement of a fruitful multilingual speech-to-speech framework for portable gadgets.

➢ *SOPC-Based Speech-to-Text Conversion*

Developers have been processing speech for a wide range of applications, from automatic reading machines to mobile communications, over the past few decades. The overhead resulting from using different communication channels is decreased via speech recognition. Due to the extensive range and complexity of speech signals and sounds, speech has not been exploited extensively in the fields of electronics and computers. However, we can quickly process audio signals and detect the text thanks to modern procedures, algorithms, and methodologies.

➢ *Text-to-Speech Algorithms based on FFT Synthesis*

Provide FFT synthesis techniques for a diaphone concatenation-based French text-to-speech system. FFT synthesis methods can produce natural voice prosodic modifications of excellent quality. A number of strategies are developed to lessen the distortions brought on by diaphone concatenation.

➢ *Exploring Speech-to-Text(STT) Conversion using SAPI for Bangla Language*

The achieved performance is encouraging for research on STT, but they also found a number of components that might improve accuracy and ensure that the study's subject would be beneficial for Speech-to-Text translation and comparable activities in other languages.

➢ *Objective*

The project's main goals are to support illiterate individuals in typing text more easily and to promote the use of our local tongue. The idea entails using a microphone to recognize voice. Putting a noise filter cap over the microphone helps to cut down on background noise. Artificial intelligence and feature extraction are used to extract the words from the input voice. Word tokenizer is used to identify these terms in the spoken speech after we transform the word using NLTK. After the words are extracted, data analysis is used to compare them with the pre-trained data set.

➢ *Existing System*

The Java Speech API defines a standard, cross platform software interface to state-of-the-art speech technology. The Java Speech API supports speech recognition and speech synthesis, two fundamental speech technologies. With the use of speech recognition, computers can now hear spoken language and comprehend what has been spoken. Put otherwise, it converts speech-containing audio input to text in order to process it. The open development process was used to create the Java Speech API. With the active involvement of leading speech technology companies, with input from application developers and with months of public review and comment, the specification has achieved a high degree of technical excellence. As a determination for a quickly advancing innovation, Sun will bolster and improve the Java Speech API to preserve its driving capabilities. The Java Discourse API is an expansion to the Java stage. Expansions are packages of classes composed within the Java programming dialect (and any related local code) that application designers can utilize to expand the usefulness of the center portion of the Java stage. But the main disadvantage is it recognize only the some reserved words only.

- *Disadvantages:*

✓ Some solutions are difficult to combine with the current HER.
✓ There is limited support for speech recognition on various devices and operating systems (namely Macs).
✓ Not all systems offer a lot of customizing choices.

➢ *Proposed System*

Text translation from spoken language is aided by speech recognition. Our solution is a spoken Recognition model that translates the user-provided spoken data into a written format in the language of his choice. This model is created by incorporating Multilingual functionalities into the current Google Speech Recognition model, which is founded on certain principles of natural language processing. The objective of this study is to develop a voice recognition model that can even make it simple for an illiterate individual to speak with a computer system in his native tongue.

- *Advantages:*

✓ Solves Inefficiencies and Reduces Wasted Time
✓ Clinics and Hospitals Can Save Money

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR598

➢ *Goals*

- Give users access to an expressive, ready-to-use visual modelling language so they can create and share valuable models.
- Offer methods for specialization and extendibility to expand the fundamental ideas.
- Be unaffected by specific development processes and programming languages.
- Offer an official foundation for comprehending the modelling language.
- Encourage the use of higher level development ideas like components, frameworks, partnerships, and patterns.
- Include industry best practices.

➢ *System* Architecture

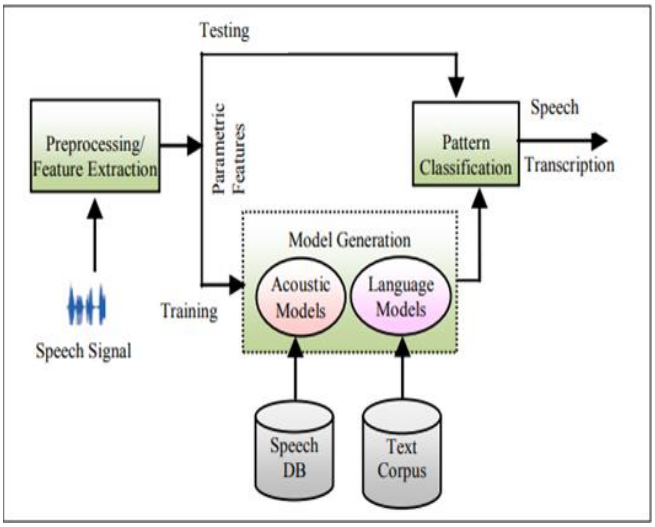

Fig 1 System Architecture

➢ *System Requirements*

- *Hardware Requirements*

✓ System: Pentium Dual Core.
✓ Hard Disk: 120 GB.
✓ Monitor : 15'' LED
✓ Input Devices : Keyboard, Mouse
✓ Ram: 4 GB.

- *Software Requirements*

✓ Operating system: Windows 7/10.
✓ Coding Language :Python

➢ *Modules*

- Speech Analysis Module
- Feature Extraction Module
- Speech To Text Module

➢ *Module Descriptions*

- *Speech Analysis Module*

Speech data for speech analysis techniques includes a variety of information that reveals the identity of the speaker. This comprises details unique to the speaker because of their vocal tract, source of excitation, and behavioral characteristic. Every speaker has a different vocal tract's dimensions and physical makeup, as well as a different excitation source. During speech creation, this distinctiveness is ingrained in the voice signal and can be utilized for speaker recognition.

- *Feature Extraction Module*

Since feature extraction is crucial to distinguishing one speech from another, it is the most significant component of speech recognition. because each speech contains unique qualities that are ingrained in each utterance. These qualities can be gleaned from a variety of feature extraction methods that have been put forth and effectively applied to speech recognition tasks. However, when handling the speech signal, the extracted feature needs to satisfy certain requirements, like:

✓ Measurement-friendly extracted speech features.
✓ It needs to be impervious to imitation.
✓ It must show the least amount of change between speaking circumstances.
✓ It needs to remain steady throughout time.
✓ It needs to come up regularly and organically in speech.

- *Speech to Text Module*

A real-time speech-to-text conversion system accurately translates spoken words into text by mimicking the user's pronunciation. We developed a real-time speech recognition system and tested it in a noisy real-time setting. The goal of this project was to present a novel speech recognition system that is more noise-resistant and computationally straightforward than the HMM-based system.

III. RESULT AND DISCUSSION

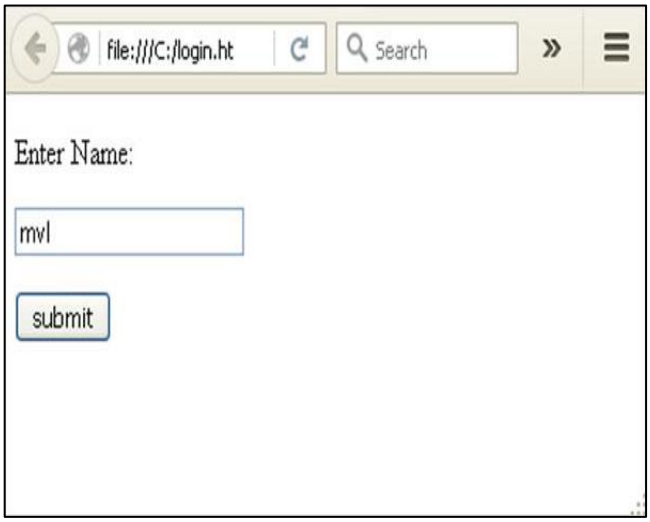

Fig 2 Local Host Login Page

Form data is POSTed to the URL in action clause of form tag.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR598

http://localhost/login is mapped to the login() work. Since the server has gotten information by POST strategy, esteem of 'nm' parameter gotten from the frame information is gotten by

```
client = request.form['nm']
```

It is passed to '/success' URL as variable portion. The browser shows a welcome message within the window.

Alter the strategy parameter to 'GET' in login.html and open it once more within the browser. The information gotten on server is by the GET strategy. The esteem of 'nm' parameter is presently gotten by-

Fig 3 Welcome Message in a Window

```
client = request.args.get('nm')
```

Here, args is lexicon protest containing a list of sets of shape parameter and its comparing esteem. The esteem comparing to 'nm' parameter is passed on to '/success' URL as some time recently.

➢ *The Following are the Main Objectives of the UML Design:*

- Give users access to an expressive, ready-to-use visual modelling language so they can create and share valuable models.
- To expand the fundamental ideas, offer tools for specialization and extendibility.
- Be unaffected by specific development processes or programming languages.
- Offer an official foundation for comprehending the modelling language.
- Encourage the use of higher level development ideas like components, frameworks, partnerships, and patterns.
- Combine the finest techniques.

IV. CONCLUSION

By implementing this model, we learned how speech recognition packages can be used to build speech translation models. The more we use these types of packages, the more flexibility we get in code display and output. This model can be used for any purpose of speech to text translation. This model has many advantages, one of them is that you can live in unknown places where you do not know the speaking language, but with the help of this model you can translate that regional speech into text and in areas like Can also use it. Telecommunication. and multimedia. Additionally, this model is also useful for providing effective communication between man and machine.

A. Screenshots

➢ *Home Page*

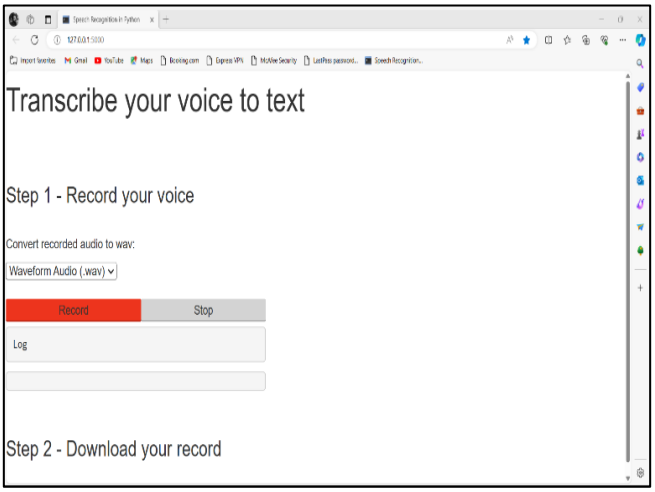

Fig 4 Home Page with Step 1&2

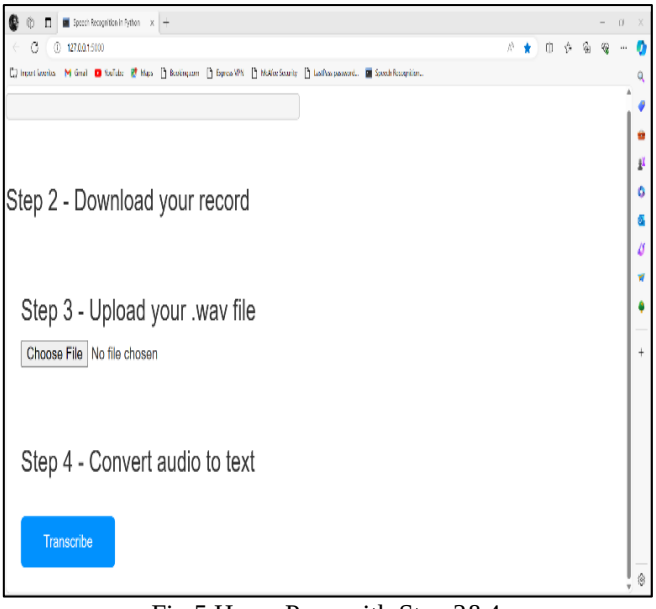

Fig 5 Home Page with Step 3&4

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR598

➢ *Speech to Text Transcribtion*

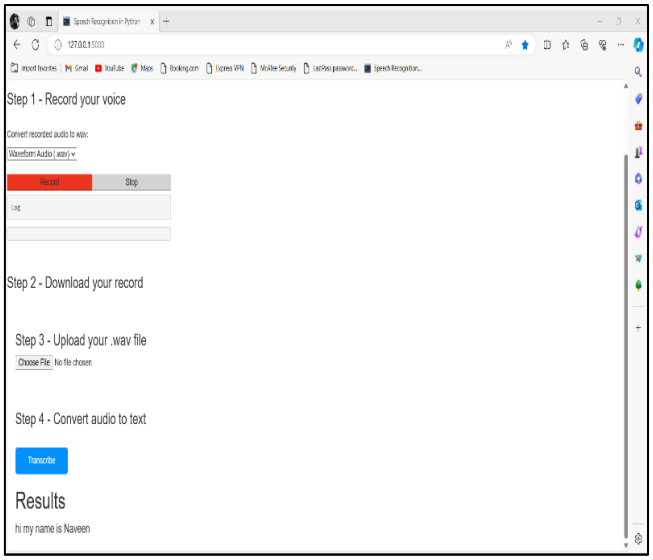

Fig 6 Transcribing Voice

REFERENCES

[1]. Mrinalini Ket al: Hindi-English Speech-to-Speech Translation System for Travel Expressions, 2015 International Conference on Computation of Power, Energy, Information And Communication.

[2]. Development and Application of Multilingual Speech Translation Satoshi Nakamura, Spoken Language Communication Research Group Project, National Institute of Information and Communications Technology , Japan.

[3]. Speech-to-Speech Translation: A Review, Mahak Dureja Department of CSE The NorthCap University, Gurgaon Sumanlata Gautam Department of CSE The NorthCap University, Gurgaon. International Journal of Computer Applications (0975 – 8887) Volume 129 – No.13, November2015.

[4]. Sequence-to-Sequence Models for Emphasis Speech Translation. Quoc Truong Do,Skriani Sakti; Sakriani Sakti; Satoshi Nakamura, 2018 IEEE/ACM..

[5]. Olabe, J. C.; Santos, A.; Martinez, R.; Munoz, E.; Martinez, M.; Quilis, A.; Bernstein, J., “Real time text-to speech conversion system for spanish," Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84. , vol.9, no., pp.85,87, Mar 1984.

[6]. Kavaler, R. et al., “A Dynamic Time Warp Integrated Circuit for a 1000-Word Recognition System”, IEEE Journal of Solid-State Circuits, vol SC-22, NO 1, February 1987, pp 3-14.

[7]. Aggarwal, R. K. and Dave, M., “Acoustic modeling problem for automatic speech recognition system: advances and refinements (Part II)”, International Journal of Speech Technology (2011) 14:309–320.

[8]. Ostendorf, M., Digalakis, V., & Kimball, O. A. (1996). “From HMM’s to segment models: a unified view of stochastic modeling for speech recognition”. IEEE Transactions on Speech and Audio Processing, 4(5), 360– 378.

[9]. Yasuhisa Fujii, Y., Yamamoto, K., Nakagawa, S., “AUTOMATIC SPEECH RECOGNITION USING HIDDEN CONDITIONAL NEURAL FIELDS”, ICASSP 2011: P-5036-5039.

[10]. Mohamed, A. R., Dahl, G. E., and Hinton, G., “Acoustic Modelling using Deep Belief Networks”, submitted to IEEE TRANS. On audio, speech, and language processing, 2010.