Automatic Speech Segmentation for Automatic Speech Translation

Piotr Kłosowski and Adam Dustor

Silesian University of Technology, Akademicka Str. 16,
44-100 Gliwice, Poland
{Piotr.Klosowski,Adam.Dustor}@polsl.pl

Abstract. The article presents selected, effective speech signal processing algorithms and their use in order to improve the automatic speech translation. Automatic speech translation uses natural language processing techniques implemented using algorithms of automatic speech recognition, speaker recognition, automatic text translation and text-to-speech synthesis. It is very possible to improve the process of automatic speech translation by using effective algorithms for automatic segmentation of speech signals based on speaker recognition and language recognition.

Keywords: speech recognition, speech translation, speech synthesis.

1 Introduction

Division of Telecommunication, a part of the Institute of Electronics and Faculty o Automatic Control, Electronics and Computer Science Silesian University of Technology, for many years has been specializing in advanced fields of telecommunication engineering [1–5]. One of them is speech signal processing [6–8]. The one of many research areas aims to gain new knowledge in the field of the basic phenomena of perception and processing of human speech such as understanding and translation of speech made by a person. The main scientific objective of this research area is development of selected, effective speech signal processing algorithms and their use in order to improve the automatic speech translation. Automatic speech translation system uses natural language processing techniques implemented using algorithms of automatic speech recognition, speaker recognition, automatic translation of text and text-to-speech synthesis [9–11]. Research hypothesis can be formulated as follows: It is possible to improve the process of automatic speech translation by using efficient algorithms for automatic segmentation of speech signals coming from different speakers.

2 Automatic Speech-to-Speech Machine Translation

Field of automatic speech translation (called SSMT – Speech-to-Speech Machine Translation) is part of a long-established area of research on speech processing

A. Kwiecień, P. Gaj, and P. Stera (Eds.): CN 2013, CCIS 370, pp. 466–475, 2013.

and natural language [12]. This is an area of great importance, which is associated with high hopes, because it relates to the basic problems and needs of the modern information society, such as communication between people and access to information in different languages, which is essential in today's globalized world [9]. The catalog of languages lists over 7000 living languages [13, 14]. Although currently available technology provides many ways of global communication, it is the variety of different language speakers that can be a serious barrier to communication. Basic research in the field of digital signal processing of speech can make a significant contribution to solving basic technical problems in the field of automatic speech translation, which is one of the main priorities for the development of the information society. The importance of the research field of automatic speech-to-speech translation is mirrored by a multitude of recent or ongoing large-scale research projects [15–17].

Automatic speech-to-speech translation systems can play a critical role in empowering people to communicate with speakers of a different language and to access or present information in a cross-lingual way. Speech translation is the process by which conversational spoken phrases are instantly translated and spoken aloud in a second language. A speech translation system would typically integrate the following three software technologies: automatic speech recognition (ASR), automatic machine translation (AMT) and voice synthesis (TTS). Tasks of typical automatic speech-to-speech translation is presented in Fig. 1. The tasks are as follows:

- **Automatic Speech Recognition** – translation of spoken words into text,
- **Automatic Machine Translation** – translation of text from source language to destination language,
- **Automatic Speech Synthesis** – artificial production of human speech based on text-to-speech (TTS) conversion of normal language text into speech.

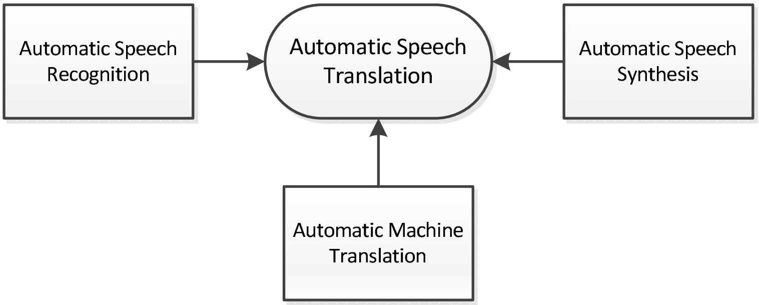
Automatic Speech Recognition
Automatic Speech Translation
Automatic Speech Synthesis
Automatic Machine Translation

Fig. 1. Tasks of typical automatic speech-to-speech translation system

Figure 2 presents block diagram of typical automatic speech-to-speech translation from speech-to-speech in language A to speech in language B.

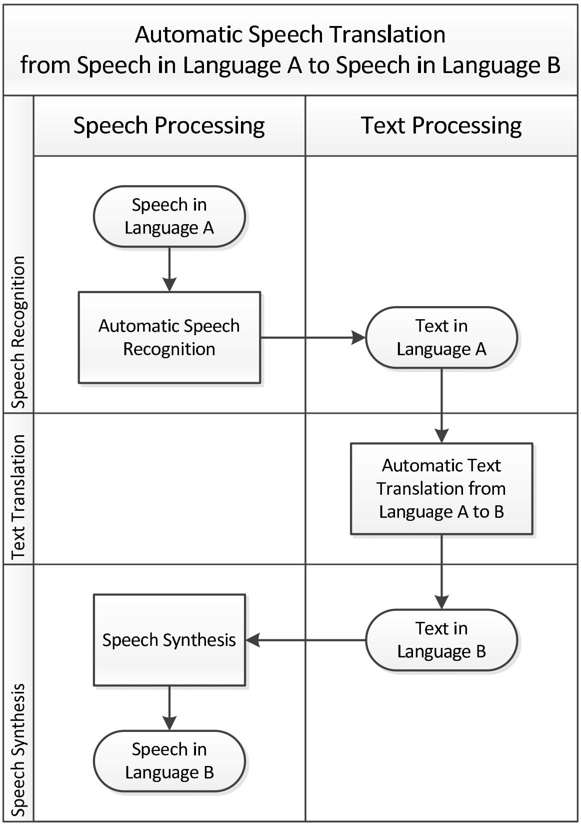

Fig. 2. Block diagram of typical automatic speech-to-speech translation from speech in language A to speech in language B

It is very possible to improve the process of automatic speech translation by using efficient algorithms for automatic segmentation of speech signals coming from different speakers in different languages. Improving is possible by adding automatic speech segmentation process based on speaker recognition and language recognition algorithms. Tasks of improved automatic speech-to-speech translation is presented in Fig. 3. The tasks are as follows:

- **Automatic Speech Segmentation and Speaker Recognition** – the identification of the person who is speaking by characteristics of their voices (voice biometrics), also called voice recognition. The general area of speaker recognition encompasses two more fundamental tasks: speaker identification and speaker verification. Speaker identification is the task of determining who is talking from a set of known voices or speakers.
- **Automatic Language Recognition** – process of determining which natural language given content is in speech.
- **Automatic Speech Recognition** – translation of spoken words into text.

- **Automatic Machine Translation** – translation of text from source language to destination language.
- **Automatic Speech Synthesis** – artificial production of human speech based on text-to-speech (TTS) conversion normal language text into speech.

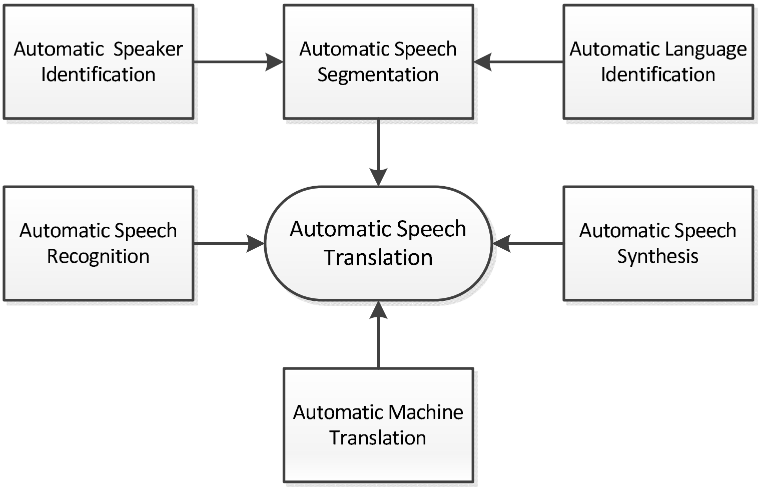
Automatic Speaker Identification
Automatic Speech Segmentation
Automatic Language Identification
Automatic Speech Recognition
Automatic Speech Translation
Automatic Speech Synthesis
Automatic Machine Translation

Fig. 3. Tasks of improved automatic speech translation system

Figure 4 presents block diagram of improved automatic speech-to-speech translation from speech in language A to speech in language B.

3 Automatic Speech Segmentation and Speaker Recognition

The purpose of the automatic segmentation is to separate the speaker from the audio signal containing speech fragments and to allocate them to the speakers. It is also desirable before the speech translation process to remove untranslatable fragments as music and other sounds, ambient noise and noise. This necessity stems from the fact that most of the speech recognition systems only work well when the recognized speech is free of noise [18]. In addition to segmentation in many applications it is also necessary to identify the speakers, which is separated from the signal assignment parts of speech to individuals whose identity is unknown. This occurs in applications including teleconferencing, where in addition to the transcription of speech it is important to the identity of the speaker.

Automatic speaker recognition is field of knowledge akin to speech recognition. An important element in distinguishing these two issues is the fact that speech recognition is important to extract the contents of the analyzed linguistic expression, while recognizing the speaker characteristics of the speech signal

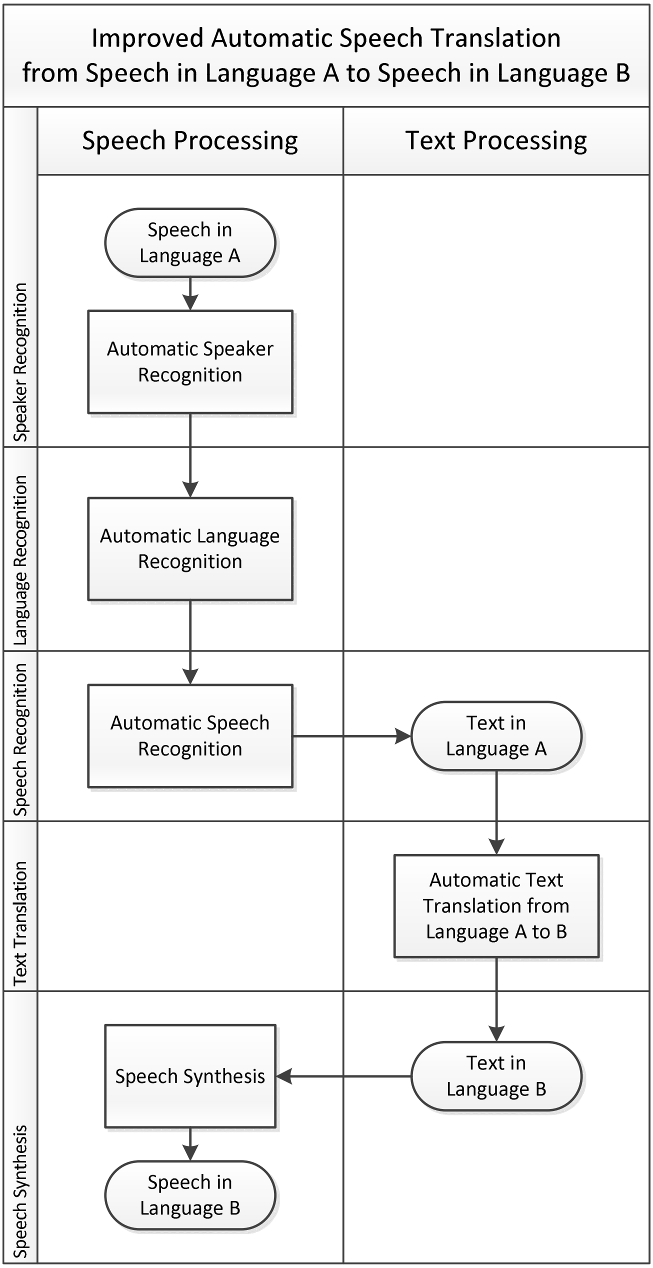
Improved Automatic Speech Translation from Speech in Language A to Speech in Language B
Speech Processing
Text Processing
Speaker Recognition
Language Recognition
Speech Recognition
Text Translation
Speech Synthesis
Speech in Language A
Automatic Speaker Recognition
Automatic Language Recognition
Automatic Speech Recognition
Text in Language A
Automatic Text Translation from Language A to B
Text in Language B
Speech Synthesis
Speech in Language B

Fig. 4. Block diagram of improved automatic speech translation from speech in language A to speech in language B

output specific to the person who will recognize it in the future. Automatic speaker recognition includes automatic speaker verification, automatic speaker identification and speaker authentication. During the verification process, a user must initially declare their identity by entering their personal identification number and is then obligated to provide one or more statements. The result of the verification process is confirmation or rejection of the identity declared by the user. Such a decision is based on comparison of similarity between the model voice (already registered in the system) and the recognized utterance of a fixed threshold.

In the process of speaker identification, identity is not predeclared and speaker, whose voice is subject to examination, may have been previously registered in the system (has its own model of voice), or is someone completely unknown to the recognition system. During the identification of a set of closed (called closed set identification), it is assumed that access to the system is granted to those whose voice models were developed. The recognition system makes a choice type 1 from N, where N is the number of registered users. When this assumption is not true, there is identification in the open set (called open-set identification). It may happen that similarities of the unknown speaker's speech to the characteristics of one of the models of speakers registered in the system is large enough that you can decide to identify a person or to regard it as not belonging to any of the speakers in the system. In the second of these situations the system may decide to reject the speaker or his registration.

The last of the speaker recognition procedures is speaker authentication, which consists of determining whether the statement is one of the speakers already registered in the system or not. Speaker recognition systems are also divided as text dependent or text independent. The relationship of the text means that when trying to identify, the system requires that a person diagnosed uttered a word or words that were in sequence learning, which is used to create a model of the speaker. The system without requirements on pre-entered dictionary is called "independent of the text". Due to the same vocabulary learning and test sequences the effectiveness of recognition process from the text-dependent systems is greater. These systems typically rely on a fixed password assigned to each user. In the case of an incorrect recognition, the system often requires keyword repeated several times. Automatic speech segmentation used in automatic speech translation systems must be based on text independent speaker recognition algorithms.

4 Automatic Language Recognition

One of the important components of an automatic speech translation module is automatic language recognition. The task of spoken language recognition is to determine the language of an utterance. Since multilingual applications become increasingly popular due to globalization, spoken language recognition has become an essential technology in areas such as multilingual conversational systems, spoken language translation and multilingual speech recognition. Most of the systems rely on two types of features: the acoustic features and the phonotactic features. The acoustic features reflect low-level spectral characteristics, while the phonotactic features represent the phonological constraints that govern a spoken language. Both features have been shown to be effective in spoken language recognition. The most important factors affecting the level of language recognition system errors are: the duration of the test expression and a limited amount of learning material, the problem of the diversity of speakers, the low quality of the speech signal and the large variation in the time to vote. The impact of these factors in an ideal language recognition system should be kept to a minimum and the recognition system should work with the smallest possible error rate.

Automatic language recognition process can be divided into: the language identification and language verification. Language identification is to determine speakers language on an open set of languages. Language verification to determine whether the speaker actually speaks in a language that speaker declares. The most important characteristics of the language are: a set of phonemes, vocabulary and grammar rules specifying the connection between the words and sentence structure. In addition, each language has a significant acoustic patterns such pronounciation and melody of words and sentences (prosody). The fact that the vocabulary and grammar are almost too extensive source of information, however, leads to the fact that most of the developed systems use features such as phonotactics and simpler properties as acoustic and prosody.

Process of automatic language recognition consists of several stages. Statement in an unknown language is divided into short, on the order of 20 ms, segments called frames. Each frame is subjected to parameters extraction process. The most frequently used parameters are Linear Prediction Cepstral Coefficients (LPCC) or Mel Frequency Cepstral Coefficients (MFCC). The sequence of multidimensional parameters is used later in the process of calculating the similarity between the model and the language of this sequence. Maximum similarity is the criterion for recognizing the decision-making system. Constructing models of language during learning phase is very important and determines the effectiveness of recognition. Two families of models are used in language recognition. The first group of models is based on vector quantization. Each language is represented by a collection of multidimensional vectors defined using clustering techniques (k-means algorithm) based on training speech sentence. This collection creates a codebook. During recognition for each of the test vectors a distance to the nearest neighbor from codebook is calculated. Total distance, normalized to the length of speech is the basis for the decision of the recognition system. The second approach to the modeling language is through utilization of parametric models using statistical properties of the voice. In this case, the language is represented by sum of Gaussian Mixture Models (GMM) [19].

It seems that the obtained results could be significantly improved if some higher level information was used, e.g. pronunciation, vocabulary or accent. However, this requires a broader approach to language modeling and will be the aim of further research.

5 Research Methodology

The specific objectives of the research project can be defined as follows:

1. Development of the structure, construction and functional modules of automatic speech translation, using automatic segmentation of the speech signal derived from a variety of speakers who speak different languages.
2. Development of efficient algorithms for the identification / verification of the speakers allowing for segmentation of speech coming from different speakers.
3. Development of efficient algorithms for automatic language identification of speech fragments.

Objectives will be achieved by the following research tasks:

- Record multimedia content in the form of recordings and speech samples from different speakers in different languages. No studio condition recordings are required.
- Development of algorithms for the identification / verification of the speakers allowing for segmentation of speech coming from different speakers.
- Implementation of developed speaker identification / verification algorithms allow segmentation of speech coming from the various speakers in the MATLAB. Evaluation of algorithms in action. The MATLAB is also used for speech feature extraction.
- Developing algorithms for automatic language identification of speech fragments.
- Implementation of algorithms for automatic language identification of parts of speech in the MATLAB. Evaluation of algorithms in action. The MATLAB is also used for speech feature extraction.
- Development of design and simulation environment that allows to assess the effectiveness of the developed algorithms.
- Experimental research. Evaluation of results and their statistical analysis. Preparing articles for publication in scientific conferences and journals.

The research methodology consists of three steps. The first stage is the formulation made in a thesis of the theoretical development of the algorithm. The second stage is the experiment, as the practical implementation of the proposed algorithms in the selected environment. The third step is thesis verification by evaluation of the algorithm effectiveness, leading to the confirmation or rejection of the thesis formulated in the first step. It is expected that the developed algorithms will improve the automatic speech translation. Their implementation in the MATLAB computing environment and analyze the effectiveness of their actions will evaluate the usefulness in automatic speech translation. Research project will also require collection of source audio files as multimedia recordings of speech from different speakers in different languages. The culmination of the project research will be a series experiments testing the effectiveness evaluation of the developed solutions and their potential use in automatic speech translation.

6 Summary

An expected result of the research project is the development of efficient algorithms which allow to improve the automatic speech translation. The use of automatic speech translation systems can be versatile. The research project is very innovative, because the problem of automatic speech translation has not been effectively resolved. Each solution to improve the performance of automatic speech translation seems to be very important. Number of research projects carried out in the field of automatic speech translation shows that this research area is very actively developing and research projects in this area are supported by the governments of many countries around the world. Many of these projects have an international character. The combination of these efforts in many countries has a chance to create an interesting prospect for future research projects aimed at solving the fundamental problems of global communication multilingual societies in a globalized world.

References

1. Dziwoki, G.: An analysis of the unsupervised phase correction method in quadrature amplitude modulation systems. Przeglad Elektrotechniczny 88(7a), 245–249 (2012)
2. Izydorczyk, J., Izydorczyk, M.: Limits to microprocessor scaling. Computer 43(8), 20–26 (2010)
3. Sułek, W.: Pipeline processing in low-density parity-check codes hardware decoder. Bulletin of the Polish Academy of Sciences Technical Sciences 59(2), 149–155 (2011)
4. Zawadzki, P.: Security of ping-pong protocol based on pairs of completely entangled qudits. Quantum Information Processing 11(6), 1419–1430 (2012)
5. Kucharczyk, M.: Blind signatures in electronic voting systems. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2010. CCIS, vol. 79, pp. 349–358. Springer, Heidelberg (2010)
6. Dustor, A.: Speaker verification based on fuzzy classifier. In: Cyran, K.A., Kozielski, S., Peters, J.F., Stańczyk, U., Wakulicz-Deja, A. (eds.) Man-Machine Interactions. AISC, vol. 59, pp. 389–397. Springer, Heidelberg (2009)
7. Kłosowski, P.: Speech processing application based on phonetics and phonology of the polish language. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2010. CCIS, vol. 79, pp. 236–244. Springer, Heidelberg (2010)
8. Kłosowski, P., Pułka, A.: Polish Semantic Speech Recognition Expert System Supporting Electronic Design System. In: Prooccedings of The International Conference on Human Systems Interactions, HSI 2008, Kraków, Poland. IEEE Eurographics Technical Report Series, pp. 479–484 (2008)
9. Stuker, S., Herrmann, T., Kolss, M., Niehues, J., Wolfel, M.: Research Opportunities In Automatic Speech-To-Speech Translation. IEEE Potentials 31(3), 26–33 (2012)
10. Koehn, P.: Statistical Machine Translation. Cambridge Univ. Press, Cambridge (2009)
11. Waibel, A., Fügen, C.: Spoken language translation-enabling crosslingual human-human communication. IEEE Signal Processing Mag. 25(3), 70–79 (2008)

12. Huang, X., Acero, A., Hon, H.W.: Spoken Language Processing. Prentice-Hall, Englewood Cliffs (2001)
13. Gordon Jr., R.G.: Ethnologue, Languages of the World, 15th edn. SIL International, Dallas (2005)
14. Janson, T.: Speak-A Short History of Languages. Oxford Univ. Press, London (2002)
15. Hutchins, J.: International Association for Machine Translation compendium of translation software (2010), `http://www.hutchinsweb.me.uk/Compendium.htm`
16. A new framework strategy for multilingualism, Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee, and the Committee of the Regions. Commission of the European Communities (November 2005)
17. Steinbiss, V.: Human language technologies for Europe. Work Comissioned by ITC-irst, Trento, Italy to Accipio Consulting, Aachen, Germany (April 2006)
18. Rabiner, L.R., Juang, B.H.: Fundamentals of speech recognition. Prentice-Hall (1993)
19. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing 3(1), 72–82 (1995)