Within the framework of this project the research will be conducted whose aim is to develop the method for automatic phonetic transcription of speech (in English), which will be based on combining the information coming from the analysis of the audio and the visual signal. A synergy effect is expected thanks to combining data streams from various modalities which will facilitate the achievement of greater effectiveness. In particular, the applicants plan to conduct a research on the relation between the allophonic variation in speech, i.e. the differences in the articulatory setting of speech organs for the same phoneme produced in different phonetic environments (adjacent sounds and/or prosodic position) and the objective signal parameters (both audio and video). An in-depth analysis of speech signal audio and video parameters with the use of the electromagnetic articulograph will be carried out for Polish speakers learning English, which will provide additional data pertaining to the movements of speech organs, thus enabling an extremely detailed analyses of particular articulatory settings.

Phonetic transcription consists in transforming a text into a sequence of corresponding phonetic symbols which reflect the actual phonetic realization of particular words. The system which is generally used worldwide is that of International Phonetic Association (IPA). In essence, the system includes the symbols for all phonemes which occur in natural languages and is supplemented with a set of diacritic marks which represent all possible variants of a given phoneme. Until now, detailed allophonic transcription has been done auditorily by expert phoneticians. Thanks to the method developed in this project the process will be substantially facilitated, if not fully automatized. The assumption is to create a sufficiently accurate method which will cater for minute allophonic and accentual variations. The assumption is that by using the analysis of video signals together with the acoustic signal, speech transcription can be performed more accurately and robustly than by using the acoustic modality alone. This application is supported by the results of the previous research.

The results of the project may be used both by linguists and people learning English as a second language. The advanced research on allophonic variation in the context of audio and visual signal parameters will also contribute to the advancement of the state of the art in the field of audiovisual speech recognition and consequently in human-computer interaction.

The development of automatic allophonic transcription method has a number of potential applications and contributes to the progress in many areas of knowledge, which includes:

- **phonetic and phonological studies**: the transcriptions available in dictionaries and data bases rely on subjective and impressionistic judgments of phonetic specialists. An automatized method will constitute a tool for fast processing and annotation of large speech corpora in an objective and repeating manner, thus facilitating phonetic and phonological research and possibly leading to a number discoveries in the field. Most of the existing solutions require the text input information. There also exist systems which are based on the acoustic signal only but they do not operate on the allophonic level and are not supported visually.

- **language learning methodology**: according to the British Council, around 750 million people, including many Polish speakers, worldwide speak English as a foreign language. One of the most important aspects of foreign language learning is pronunciation. Automatic transcription at the allophonic level will greatly facilitate the effectiveness of learning English pronunciation, including distant learning solutions.

- **automatic accent recognition**: while the information concerning the phones is sufficient for the recognition of speech contents, the correct recognition of an accent requires the information concerning the allophonic variation. The existing research on automatic accent recognition is based on a blind, statistical approach to the problem. The method developed in this project opens the possibility of comparing the features of an utterance with the familiar patterns and facilitates the recognition of speaker's accent, which may be applied in human-computer interaction and for instance forensic analyses. Another application is adjusting the language input to the user's geographical or ethnical background.

- **audiovisual speech recognition**: the technology of audiovisual speech recognitions needs a breakthrough in order to equal in popularity with acoustic speech recognition. The state of the art in this area is only able to increase the effectiveness of speech recognition in noise conditions and not to increase the accuracy of analysis. Such breakthrough may result from identifying the parameters of visual signal directly correlating with phonetic aspects, which becomes possible thanks to the in-depth research conducted within this project.

The project is divided into 5 phases:

1. **Preparatory phase** consists in the construction of a database of allophones and visemes, which will constitute the research material for the following phase of the project. The acoustic and visual (visemes) representations of all English allophones in consonantal and vocalic context will be recorded. For this purpose advanced HD/high-speed (100fps) cameras and thermovision cameras will be used. Additionally, electromagnetic articulographic (EMA) signals will be recorded which provide information about the exact articulatory settings. Prior to the recordings the equipment will be prepared and the subjects will be recruited. The group of subjects will include both native speakers of English and Polish speakers learning English as a foreign language (at different levels of advancement in English).

2. In the **experimental phase** a detailed analysis of the recorded allophones and visemes will be conducted. The analysis will be both phonetic and numerical. The phonetic experiments will be conducted by phonetics experts who will analyze the recorded allophones and visemes and the articulographic data with the view to finding ways of extracting relevant parameters in the audio and video signals. In the numerical experiments we will calculate the parameters of audio and visual signal which correctly reflect the allophonic variation. The result of the experimental phase will be the identification of parameters and patterns which reflect the differences between particular allophones. Decision algorithms will also be created which will be able to discriminate particular allophones in the parameters space.

3. In the **training phase** the method will be developed for automatic transcription of speech. On the basis of the training data we will design the signal processing algorithms, the classifiers and appropriate models which will enable the transformation of acoustic wave information into a sequence of transcription symbols.

4. In the **testing phase** the accuracy of the method will be verified. Appropriate accuracy measures will be calculated at the level of symbols (allophone) and words. The accuracy of the method will be analyzed in relation to the speaker's native accent and the level of advancement in English (for Polish subjects).

5. In the **conclusion phase** a summary of the research results will be formulated. An audiovisual illustration of all English allophones in the form of pictures associated with the sound will be prepared. The databases of allophones and visemes will be published and made available to research community in the form of a multimedia database.

The project results will have a global impact. The progress in the state of the art is expected in information systems, English phonetics and phonology. There also exist other fields in which the method may potentially be used. The results of the project will be published in high-rank journals of global circulation and presented at international conferences.