DEVELOPMENT OF AN AUTOMATIC IDENTIFICATION SYSTEM OF SPOKEN LANGUAGES: PHASE I

Deidre Cimarusti
Russell B. Ives

Speech Communications Research Laboratory
Los Angeles, California

ABSTRACT

The feasibility of a new approach to automatic language identification is examined in this pilot study. The procedure involves the application of pattern analysis techniques to features extracted from the speech signal. The database of the extracted features for five speakers from each of eight languges was divided into a learning subset and an evaluation subset. A potential function was then generated for all features in the learning subset. The complexity of the decision function was systematically increased until all members within the learning subset could be separated into the properly identified languages. Although the constraints on this pilot study necessarily precluded feature ordering and selection, the application of the decision function to the evaluation subset resulted in an overall 84% classification accuracy.

INTRODUCTION

This paper presents Phase I, a feasibility study to investigate a procedure to automatically identify spoken languages on the basis of pattern analysis techniques applied to features extracted from the speech signal. The approach of employing pattern analysis techniques was selected because it is reasonable to assume that humans are capable of language discrimination on the basis of some language-unique features of which they are apparently minimally aware. Not only does a speaker have the ability to identify the language(s) in which he has competency, but also, when presented with an utterance from a foreign language, he need not have fluency in that language to determine whether the utterance is from a language within the set of languages in which he has competency, or whether it is from a language which lies outside this set. Once he has determined that the utterance is not from his own linguistic repetoire, it is conceivable that he may be able to "guess at" the language to which it does belong on the basis of his stored linguistic knowledge, i.e., his knowledge of languages not within his competency but within his limited familiarity.

Previous studies in language identification research presented evidence that linguistic units most likely discriminate languages. Certain basic linguistic units have a higher frequency of occurrence in the speech utterances from one language than from any others under consideration. Therefore, it was hypothesized that languages can be discriminated on the basis of the distributions of these linguistic units (1,2,3,4,5). Such discriminating linguistic units can be described in terms of their acoustic characteristics in the speech signal. Because of this, it is reasonable to assume that the acoustic description of the speech signal will retain the discriminating nature of the language-unique properties. The currently investigated procedure differs from those employed in previous language identification research, in that it is not assumed that the discriminating features need necessarily be linguistic units.

CONDITIONS FOR SYSTEM DEVELOPMENT

The ultimate goal of applied automatic language identification research is to achieve a system which will be able to discriminate among languages with a high degree of efficiency. Toward this goal, consideration must be given to the eight conditions shown in Table 1.

****

Table 1. System Conditions

1. Content-independence
2. Context-independence
3. Form-independence
4. Language-independence
5. Speaker-independence
6. Style-independence
7. Degraded speech signal
8. Total automation

****

The terms in Table 1 are generally accepted as follows:

1. Content-independence implies that the message being transmitted may contain information about an infinite number of topics.
2. Context-independence indicates that the speech signal surrounding the extracted portion may be from among an almost infinite number of possible contexts.
3. Form-independence refers to the form of the input signal presented to the language identification system, e.g., live communication, recorded speech, or transmissions via a telephone system.
4. Language-independence implies that the speech sample to be presented to the system may belong to any language from among the world's more than 3,000 languages.

CH 1746-7/82/0000 - 1661 $ 00.75 © 1982 IEEE

5. Speaker-independence implies that the speaker can be known or unknown, of any sex, age, emotional state, or condition of health.

6. Style-independence implies that the speech to be processed by the language identification system may be casual conversation, formal speaking, read speech, or any other speaking style.

7. It is a necessary requirement that a speech signal can be somewhat degraded without having the results significantly degraded.

8. Total automation signifies a system in which no human interaction is necessary.

The following experiment met the above conditions in varying degrees, as will be discussed below.

THE EXPERIMENT

The three primary subsets of the experiment were the compilation of the database, the selection of the features to be extracted, and the generation of the classification decision methodologies to be employed.

The Database

The database consisted of the three minutes of read speech for each speaker of each language. The speech was extracted from audio recordings of five speakers for each of the following eight languages: American English, Czech, Farsi, German, Korean, Mandarin, Russian, and Vietnamese. The techniques decided upon for extracting the selected features from the speech signal required that the speech be digitized before analysis could be performed; thus the audio recordings were passed through a 5 KHz lowpass anti-aliasing filter and digitized at a sampling frequency of 10 KHz.

As previously mentioned, the conditions for system development were met in varying degrees. Content-independence was achieved. The content of the speech utterance was not monitored in order that the language discrimination accuracy would not be dependent on any key words, nor would the subject of the message be in any way a factor in the identification process. Context-independence was also incorporated in its totality so that the signal to be processed was not limited by boundary conditions.

The form of signal input was restricted to speech recorded by means of an obvious microphone so that the form of the input signal was not a factor in the language discrimination accuracy. The ultimate goal is to achieve form-independence such that the language identification system may respond to a variety of input signals.

Language-independence is obviously one of the ultimate goals in the development of a language identification system. This pilot study necessarily had constraints as to the number of languages to be investigated so that the study would remain feasible and manageable under today's technology. The consideration as to which languages should be included in an investigation is of importance. Ideally, a language identification system should be capable of discriminating different, as well as closely related, languages. In Phase I, the number of languages was restricted to eight. These eight languages (American English, Czech, Farsi, German, Korean, Mandarin, Russian, and Vietnamese) represent a fairly diverse group. Two pairs bear a familial relationship: American English and German (Germanic), and Czech and Russian (Slavic), but the remaining possess no familial relationships (6).

The condition of speaker-independence covers many variables. The goal of attaining speaker-independence is vital since the application of a language identification system to the real world will necessitate this characteristic. In Phase I, only adult males assumed to be of normal emotional states and conditions of health, at the time their speech was recorded, were included. The number of participants was necessarily restricted to five speakers per language because of the constraints placed on this initial study.

The ultimate language identification system should be capable of accepting any style of speech. In Phase I, the style was restricted to read speech.

The condition of signal degradation was not addressed at all with Phase I because all input signals were of reasonably good quality. It was determined that this condition could best be investigated in subsequent phases.

Feature Extraction

One hundred features were selected to be extracted from the speech signals of each of the forty speakers. Based upon the hypothesis that there is some set of language-unique features, not necessarily known by the speaker or listener, almost all of the features available to the experimenters utilizing the Interactive Laboratory System (ILS) Program at SCRL were extracted. These were selected knowing that many are redundant, that not all would be useful, and that there might be other features which could prove to be more useful. These selected features will later be investigated as to their relevance in language discrimination.

The one hundred features were: 15 area functions, 15 autocorrelation coefficients, 5 bandwidths, 15 cepstral coefficients, 15 filter coefficients, 5 formant frequencies, 15 log area ratios, and 15 reflection coefficients. The analysis conditions included a 30 msec frame length and a context shift of 30 msec as well.

Pattern Recognition Methodology

The classification process for Phase I consisted of first randomly dividing the available database into the learning and evaluation subsets. Each subset contained an equal number of feature vectors from each language.

The feature logic was developed on the learning subset and the the appropriate decision func-

tions were generated. The learning subset was separated until 100% classification accuracy was achieved. This was possible because of the recursive iteration capabilities of the potential functions utilized by the pattern analysis program which was employed in this investigation. The results of the logic development were tested on the evaluation subset and the unknown samples were assigned to the language classes with an overall accuracy of 84%. The results for each language are shown in Table 2.

The actual form of the decision function was that of a Type II Exponential Function. The Type II suggests that the decision function may be expressed as a polynomial with an infinite number of terms. In practice, the number of terms can never exceed the number of unique point-pair distances within the learning set. A significant amount of computation time was saved by precomputing all valid distances within the training set and then adjusting the multiplicative coefficient of each term in the decision function when a modification was required. Further, the reciprocal of the largest distance within the training set was used as a scaling coefficient for the exponent of each term in the polynomial.

****
Table 2. The Results of Phase I.

Language	Correct Classification
American English	76.8%
Czech	84.3%
Farsi	85.2%
German	91.6%
Korean	93.4%
Mandarin	83.9%
Russian	78.1%
Vietnamese	81.7%

****

DISCUSSION

The pioneering work in language identification which was based on linguistic units was scientifically solid. However, to attain greater than 80% accuracy, a rejection criterion was introduced. We were, therefore, inclined to take an entirely different approach. This approach, which was not based on linguistic units, provided an overall accuracy better than any achieved in the previous studies where no rejection criteria was included. One does have to take into account, however, that there are many variables that can make percentages appear to be better. The importance of these percentages must be cautiously considered. However, the results are encouraging enough to be interpreted as indicating that the application of pattern analysis techniques on a certain set of features proved to be a viable approach to the development of an automatic language identification system.

Total automation is a characteristic that would be valuable in the ultimate language identification system. Not only would humans be freed to work elsewhere, but this would serve to eliminate the errors caused by such problems as fatigue and/or inattention. This condition was almost completely achieved in Phase I. The only human decision making was the manual selection of the speech signal to be used as input.

Because Phase I was a pilot study, it was desired that the speech input be free of non-speech sounds such as coughs, sneezes, and extremely long silences, and speech was selected from the middle of the recording in order to allow the speaker's voice to settle down from the initial tenseness associated with knowing one's speech is being recorded. The process could easily be totally automated given that a program could be written to automatically and arbitrarily select three minutes of speech from midway through the recording. For the rest of the decision making, the process was totally automatic.

Although rather good results were obtained in discriminating between the eight target languages in Phase I, it is reasonable to assume that not all of the one hundred features per sample contributed to the classification process. In fact, it is possible that some features may have actually degraded the potential results. This leads us to our next obvious step of feature ordering and selection. The goal of this task would be to minimize the number of features required for classification through the selective elimination of those not contributing to the classification process.

Subsequent phases will investigate feature ordering and selection as well as the expansion of the system conditions for which the ultimate goals have not yet been met.

REFERENCES

(1) R.G. Leonard and G.R. Doddington. "Automatic Language Identification," Final Report, RADC-TR-74-200, August, 1974, 785397/1GI.

(2) R.G. Leonard and G.R. Doddington. "Automatic Classification of Languages," Final Report, RADC-TR-75-264, October, 1975, B0087081.

(3) R.G. Leonard and G.R. Doddington. "Automatic Language Disrimination," Final Report, RADC-TR-78-5, January, 1978, A050841.

(4) R.G. Leonard. "Language Recognition Test and Evaluation," Final Report, RADC-TR-80-83, March, 1980.

(5) A.S. House and E.P. Neuberg. "Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations," Journal of the Acoustical Society of America, 62,3,9,1977.

(6) R.R.K. Hartmann and F.C. Stork. Dictionary of Language and Linguistics. New York: John Wiley and Sons, 1976.