Solving the Problem of the Accents for Speech Recognition Systems

Irakli Kardava[1], Jemal Antidze[1,2], and Nana Gulua[1,2]
[1]Sokhumi State University, Politkovskaia 9, Tbilisi, 0186, Georgia
[2]I. Vekua Scientific Institute of Applied Mathematics, I. Javakhishvili Tbilisi State University, Tbilisi, 0186, Georgia
Email: {sokhummi, jeantidze}@yahoo.com, ngulua77@mail.ru

***Abstract*—Since the speech recognition system has been created, it has developed significantly, but it still has a lot of problems. As you know, any specific natural language may owns about tens accents. Despite the identical word phonemic composition, if it is pronounced in different accents, as a result, we will have sound waves, which are different from each other. Differences in pronunciation, in accent and intonation of speech in general, create one of the most common problems of speech recognition. If there are a lot of accents in language we should create the acoustic model for each separately. When the word is pronounced differently, then the software can become confused and misunderstand (perception) also correctly what is pronounced. The same can also occur, if the human speaks slowly or vice versa quickly, then the program expects. There are any partial decisions (solutions) but they don't solve all problems. We have developed an approach, which is used to solve above mentioned problems and create more effective, improved speech recognition system of Georgian language and of languages, which are similar to Georgian language. In addition, by the realization of this method, it is available to solve the artificial intelligence issues, such as arrange sound dialogue between computer and human, independent from any accents of any languages.**

***Index Terms*—accents, acoustic model, phonemic composition, speech recognizer, waves**

I. Introduction

Most speaker independent (SI) speech recognition systems comprise a set of acoustic models (for example hidden Markov models, HMMs) whose parameters are estimated by using speech data from a large set of speakers. There are two principal differences which exist between speakers: acoustic differences which are related to the size and shape of the vocal tract, and pronunciation differences which are generally referred to as accent and are often geographically based. In practice, it is difficult to get full coverage of all the regional accents (in England alone there are at least ten broad regional accents). SI speech recognition systems do not perform as well as speaker dependent (SD) systems, largely because of the need to model speaker variations within a single model [1].

Because the words are composed of letters, we have carried out an observation of an individual sound of vowels, consonants [2] and studied each of these structures [3]. As a result, we have divided their sounds into I and II phases, which clearly showed that the second phases of consonants are equal to each other i.e. they are identical, and the difference between them is only the first phase (including the relevant Allophones) [4]. The changes in accents mainly are caused due to the elongation of the second phase in time, while the first phase during speaking remains unchanged. Therefore, it is enough to observe the sound of the first phase of speech recognition [5], because the extension of time of the second phase cannot change the phonemic composition of word or sentence (For composition of words we use the software, developed by us, it represents a set of programs, by which it is possible with unchanged part of a word and morphological categories to get appropriate grammatically right word-form or word-forms if such exist. Also with this software it is possible by unchanged part of a word get all possible grammatically right word-forms. This used approach is based on description of natural language morphology by using formal grammar and characterizing symbols of grammar with feature structures. For description natural language morphology we use special type of context free grammar, which describes all correct natural language word-forms. With given unchanged part and its features existed in database, also with given morphological categories (in the case of first problem) we compose morpheme classes and their representatives, which must be in related word-forms. Using of the software is effective for languages, which have developed morphology like Georgian. [6]). Similarly, it is achievable to resolve the problem [7] of the tempo of pronounced words or sentences [8] during speech recognition [9]. (The first phase can be considered as the first airflow passage of the speech organs).

Monophone acoustic models are built using 3-state continuous Hidden Markov Models (HMMs) without state-skipping with a mixture of 12 Gaussians per state. We extract standard MFCC (Mel Frequency Cepstral Coecients) features from 25 ms frames, with a frame shift of 10ms. Each feature vector is 39D: 13 features (12 cepstral features plus energy), 13 deltas, and 13 double-deltas. The features are normalized using cepstral mean normalization. For our phone recognizer, the acoustic models are context-independent (i.e., monophone acoustic models are trained). For our ASR (Automatic

Manuscript received March 12, 2015; revised July 8, 2015.

doi: 10.18178/ijsps.4.3.235-238

Speech Recognition) experiments, tied context-dependent cross-word triphone acoustic models are created with the same settings as monophones. The acoustic models are speaker and gender independent, Maximum Likelihood (ML)- trained from at-start. We build our framework using the HMM Toolkit (HTK).

We analyze dialects' timing features using Ramues' hythmic measures [Ramus, 2002]. Particularly, we compare percentages of vocalic intervals (%V), the standard deviation of vocalic intervals (ΔV), and the standard deviation of intervocalic intervals (ΔC) across pairs of dialects. These measures have been shown to capture the complexity of the syllabic structure of a language/dialect in addition to the existence of vowel reduction. Languages/dialects that have a high variability of consonantal intervals are likely to have more clusters of consonants, which lead to more complex syllables. The complexity of syllabic structure of a language/dialect and the existence of vowel reduction in a language/dialect are good correlates with the rhythmic structure of the language/dialect. We identify vocalic intervals using our phone recognizer. A sequence of consecutive vowels is considered as a single vocalic interval. Similarly a sequence of consecutive consonants is considered as one intervocalic interval. Again, we use Welch's t test to indicate significant differences in features between each dialect pair. [10]

Besides this, implementation of the above mentioned approach will have an important effect also for solving the problem of the non-native accents [11] for speech recognition systems. Analogically, it will be more reliable to recognize a speech with unnatural accent. For example, when it is desired to establish communication between machines by synthetic voices. This represents one of the major fields of artificial intelligence.

II. Phases of Sound

In order to observe of sounds of given letters (A, B, C...) for their further dividing into parts, it is needed to pronounce them slowly and separately from each other. This concerns to vowels and also to consonants, too.

Let mark the characteristic sound of X sound i.e. the first phase by x1, but the X sound's elongation in time by x~. See Fig. 1, where X sound and relevant phases are shown graphically, based on the example of the result of pronunciation of consonant "D". (Georgian letter, consonant "დ")

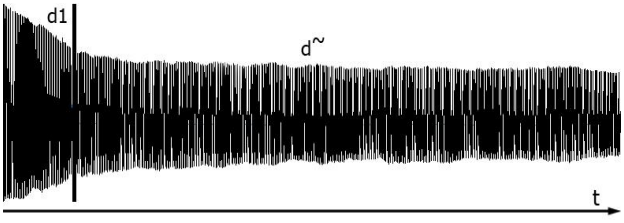

Figure 1. Sound "D", first and second phases

The left side of the vertical black line represents the first phase of sound "D", the right side represents the second phase of sound "D". The "t" is the time line. In Fig. 1 we can obviously see, that after sounding d1, d~ lasts until the end of the acoustics of the sound, as the second phase had been started. Observation showed, that each of the consonants is characterized exactly equally.

III. Deployment of Consonants Phases

As it is shown in Fig. 2, phases of the consonants with respect to the time, have a linear layout, in the X sound x1 sound is followed by the x~, so that they are not combined with each other. In addition, according to this approach, we got that (any consonant)~ = (any consonants)~. This means that in Georgian and Georgian-like languages only first phase of consonants differs from each other, while their second phases are all the same. In other words, the second phases of all consonant sounds identically.

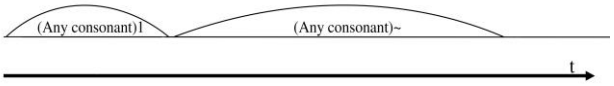

Figure 2. Deployment of phases of consonants.

IV. Deployment of Vowels Phases

Unlike consonants, the sound phases of vowels are combined with each other and their layout is parallel on time. See Fig. 3.

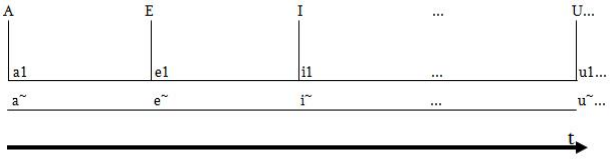

Figure 3. Deployment of phases of vowels

The relevant phases of represented vowels are marked by small letters. Vowel itself is given by big letter. The first horizontal line from the top, indicates the first phase, the second line from the top indicates the second phase. The line of "t" indicates a time. See Fig. 3.

V. Coverage of Phase during Pronunciation

When a syllable ends with vowels, for example, such as "FA" (in Georgian language "ფა"), it can be written as it is shown in Fig. 4.

$$f1\begin{matrix} a1 \\ a\sim \end{matrix}$$

Figure 4. Correct representing of syllable

Thus, the given syllable consists with the first phase of a consonant and with both phases of a vowel. This means that a~ has combined the f~ sound and we got a case when it is possible that the equation (consonants)~=(vowels)~ is correct.

For illustrating of approach developed by us, here is given the example, which clearly shows, that represented syllable sounds in different ways, when it is pronounced by various accents. See Fig. 5.

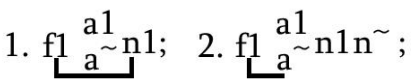

Figure 5. Phases with different combinations

At the first case, a~ - has combined f~ - and n~ -'s absence, in the second case n1 has its independent second phase, which in this concrete case sounds differently. However, at the both cases the syllable is the same: "FAN" (in Georgian language "ფან"). Exactly these different kinds of sounds create diversity of accents.

Thus, we can say that the second phase of vowel has the possibility to overlap second phase of consonants from left, right, or both sides simultaneously.

We should note that if more than one vowel is found in the word together, then uttering the word, the sound of the first vowel in the second phase does not stop and continuous exactly the same, as at the beginning and the first phase of the second vowel replaces the first phase of the first vowel, etc. and the type of the accent depends on the pace of their (first phases) replacement during the dynamics of speech. See Fig. 6.

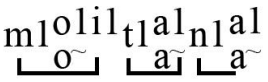

Figure 6. Word "moitana", in Georgian "მოიტანა".

The same principle applies to the proposal, when the previous word ends in a vowel and the next words begin with a vowel. This issue is one of the reasons of the pronunciation in different ways of a given word. We can say that these differences between pronunciations are called as the accent.

(Georgian letters must be pronounced exactly as they are written)

The cases which are mentioned above, presented at figures, which confirm that the second phases of these sounds or duration of them, have not any importance, for the recognition of pronounced word and sentence. Mostly it is demanded to observe at the first phase of sounds.

VI. Results of Observation

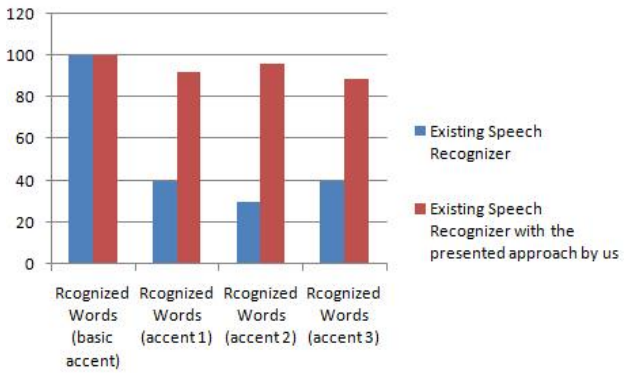

Figure 7. Comparison the presented approach to existing system.

Observation has shown that the approach gives significantly better results than existing speech recognition system to date. There was given to hundred words. They were uttered in various accents. Acoustic model was created based on only one accent, which we called "basic accent". See the Fig. 7.

VII. Conclusion

From the fact that a speech recognition system uses previously created acoustic model, the quality of recognition for different accents, which are not supported yet by the system, is unsatisfactory. In this case it means that these opportunities are not sufficiently flexible to get the desired results.

By using created problem-solving approach, it is possible to solve these accents problems of speech recognition systems. In addition, its realization gives an interesting result to solve inverse problems, such as text to speech (TTS) systems, with different accents. Through the implementation of the presented method, it is achievable to develop the universal system of speech recognition, which will be able to recognize different accents of a given language.

In general, by the realization of this approach, which is used for creation of our software, it is achievable to solve the artificial intelligence issues, such as arrange an improved sound dialogue between the computer (A machine) and human, independent from any accents of any languages. Not only for Georgian language, but for all languages, because the problem of the accents for speech recognition systems is a common problem.

Furthermore, in natural languages may exist such words which phonemic composition may be invariable, but at the same time, pronunciation in a different tone (the emphasis made in different places etc.) changes their semantics. In that way, we mean Homographs. (For example, English words: Buffet – 1. bəˈfā, 2. ˈbəfit; - to hit, punch or slap/a self-serve food bar. Second – 1. siˈkänd, 2. ˈsekənd; - 1/60th of a minute/after the first. Etc.). This means that it is possible to have different meanings to the same word, according to the change of (in) tone.

Yet, in the speech recognition systems it is impossible to define the semantics of the word according to the sound, if it is not combined with auxiliary facilities (semantic analyzer, lexical analyzer and the general use of the possibilities of formal grammars).

By using of our approach we can define the semantics of the word not after the turn into phonemes to graphemes, but before we get graphemes.

The view of the fact, that it is possible to determine the meaning of the word not just words or sentences, but the appropriate signals. This means that we can use the intonations' (emphasis) type to pronounce words to determine semantics.

According to the above circumstances, it is clear that the define the semantics of the word in this way requires less computer resources and time. Also, it is equally important, that we need a significantly smaller software code to achieve the goal, than to bring together other existing capabilities.

The possibility of implementation and speech recognition systems, we can create a voice commands, which will be ordinary words by phonemic composition and change the tone of the voice commands. It will significantly improve the process to avoid the ambiguities that arise during a dispute between voice orders and ordinary words (when they may be identical).

In other words, speech recognition system will be able just give us the text into the spoken word and the first meaning of its semantics, second intonation uttered the same word, convert into text and give us it's semantically second meaning or absorb the same word as a voice command to perform the functions (without generating any ambiguity).

It can be said that in addition to the above advantages, this comprehensive approach will be work to increase capacity and their meanings for speech recognition systems.

Our approach can be an important element also in forming the multilingual speech recognition system. For which in future it will be possible to recognize a speech for any accents of any language.

For sound surveillance is used system "Praat".

ACKNOWLEDGMENT

Thanks to Kate Tchilaia, Tamta Kvaratskhelia, Lika Daraselia and Nina Bitskinashvili for helping to prepare this article.

In addition, I want to thank all the wonderful people who, despite all the difficulties, gave an important aid during the experiments. Special thanks to professor J. Lejava, also to Mr. Nickolay Shmyrev, his suggestions were very useful for us during creating a speech recognition system for Georgian language.

REFERENCES

[1] J. J. Humphriesy, P. C. Woodlandy, and D. Pearcez, "Using accent-specific pronunciation modelling for robust speech recognition," in *Proc. Fourth International Conference on Spoken Language*, 1996, pp. 2324-2327

[2] N. Gamkrelidze, S. Kotetishvili, J. Lejava, L. Lortkipanidze, and L. Javakhidze, *Phonetic Analysis of Georgian Normative and Dialectal Speech*, Tbilisi, Georgia: Nekeri, 2006.

[3] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, New Jersey: Prentice Hall PTR, 2001.

[4] D. Melikishvili, *The Georgian Verb: A Morphosyntactic Analysis*, New York, USA: Dunwoody Press, 2008.

[5] R. Reddy, "Speech recognition by machine: A review," *Proceedings of the IEEE*, vol. 64, pp. 501-531, Apr. 1976.

[6] J. Antidze, N. Gulua, and I. Kardava, "The software for composition of some natural languages' words," *Lecture Notes on Software Engineering*, vol. 1, pp. 295-297, Aug. 2013.

[7] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *Proc. 3rd ESCA Speech Synthesis Workshop*, Jenolan Caves, Australia, 1998, pp. 77-80.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis," in *Proc. Eurospeech*, 1999, vol. 5, pp. 2347-2350.

[9] K. Shinoda and T. Watanabe, "MDL-Based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn.*, vol. 21, no. 2, pp. 79-86, 2000.

[10] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Dissertations & Theses, Columbia University, 2011.

[11] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proc. IEEE ICASSP*, 2003, pp. 540-543.

Irakli Kardava was born in Sokhumi, Georgia on November 1st, 1986. He graduated from Sokhumi State University with a master's degree in computer science, 2012. He is a PhD student of computer science at The Sokhumi State University. He has 5 scientific publications. Irakli Kardava is a member of IACSIT.

From July 15, 2015, He is chosen for Erasmus Musnus – Eminence II, at Adam Mickiewicz University (Nanobiotechnology center) in Poznan, Poland. For more information about him, please visit the web site: http://www.irakli-geo.hol.es

Jemal Antidze was born in Tbilisi, Georgia, on 10th of March 1935, graduated from The Tbilisi State University, in 1958, mathematician, PhD degree in mathematics was earned at The Tbilisi State University, 1966, fields of study – software engineering and computer linguistics. He was head of the Systemic Programming Department, director of Institute of mathematics and information technology. Now, he is a professor of The Tbilisi State University, has more than 92 scientific publications.

Prof. Antidze is a senior member of IACSIT, an expert of UNESCO in informatics, a redactor in-chief of LNSE, a member of the scientific counsil of I. Vekua Scientific Institute of Applied Mathematics of The Tbilisi State University, Tbilisi, Georgia. For more information about him, please, visit the web site: http://fpv.science.tsu.ge/.

Nana Gulua was born in Sokhumi, Georgia, on December 13th, 1968. She graduated from The Sokhumi State University, in 1994, mathematician. PhD degree in computer science was earned at The Tbilisi State University, Georgia, in 1999, fields of study – software engineering and computer linguistics. She is a full professor of The Mathematics and Computer Science Faculty at The Sokhumi State University. She has more than 31 scientific publications.