ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2005, Vol. 34, No. 3

ANALYSIS OF VOCAL PHONEMES AND FRICATIVE CONSONANT DISCRIMINATION BASED ON PHONETIC ACOUSTICS FEATURES

Kęstutis Driaunys, Vytautas Rudžionis, Pranas Žvinys

Vilnius University Kaunas Faculty of Humanities
Naugardo g.2, LT-3000, Kaunas, Lithuania

Abstract. Direct recognition of phonemes in speaker independent recognition systems still cannot guarantee good enough recognition results. Here we want to investigate assumption that recognition could be improved via group features of phonemic system. We propose to try to recognize group of phoneme first (voiced/unvoiced, vowel/ consonant, etc.) then try to recognize phoneme itself. In this experiment a system for discrimination of fricative consonants and sonants was developed using phonetic – acoustic features.

1. Indroduction

Speech technologies is a rapidly developing field and more and more often they are implemented in commercial applications. Depending from the purpose and complexity of speech recognition system we could distinguish speaker – dependent, speaker – independent and adaptable speech recognition systems. Speaker dependent recognition systems typically are designed to fulfill the needs of a single user. Such systems are simpler, cheaper, they could provide higher recognition accuracy for selected speaker but they lack of flexibility and the range of their applications is narrower than of the systems of other types. Speaker independent recognition systems are designed to serve some group of users that could be grouped under some common characteristics (e.g. native Lithuanian speakers) [11]. Such systems are more complex, more expensive and still have significantly lower recognition accuracy than previously mentioned systems. Adaptable to speaker systems are developed in such a way that they could be adjusted to speaking style and parameters of new user. This could allow achieving higher recognition accuracy. Such systems are quite popular and there are several commercial products of this type on the market (IBM ViaVoice, Dragon Naturally Speaking, etc.).

Insufficient accuracy of speaker independent speech recognition systems is mainly caused by large variations of voice characteristics of different speakers. Here we need to emphasize that good enough recognition accuracy cannot be achieved even by the best speech recognition systems based on continuous density hidden Markov model. The states of HMM (Hidden Markov model) or elementary chains implicitly models speech signal as a linear sequence of phonemes with additional acoustic events – silence, noise, pause, etc. [12]. In these systems phonemes don‘t represents exact structure of acoustic – phonetic properties except traditional HMM topology of three left – to –right directed emitting states. Researchers seek to find alternative methods of speech recognition since current state-of-the–art systems still cannot fulfill requirements for many applications, as we mentioned above.

One of the alternative approaches could be exploitation of phoneme templates using various statistical classification methods. Our experience showed that [5, 6] direct phoneme classification can‘t provide good enough recognition results. As were observed in some of our later investigations [7] it could be meaningful prior to direct phoneme classification perform phoneme classification into some of phoneme groups or some superclasses using group features of phonemes.

In this paper we present a method for recognition of speaker independent phoneme groups (sonants and fricative consonants) using phonetic – acoustical features. This approach is based on an assumption that the set of phonetic features has enough information to fix and to exploit structural properties of speech signal (that are common only to some group of phonemes) that in HMM approach aren‘t exploited sufficiently.

2. Problem formulation

Each speaker differs from others with individual voice tract characteristics. The speech signal generated by speaker could differ due to various properties of vocal tract, throat length and diameter, properties of vocal cords, age, sex, dialect, health condition, education, speaking style, emotional state, etc. So the acoustical realization of the same word or utterance pronounced by different speakers could differ very much. Even not taking into account inter-speaker differences the same speaker can't pronounce the same word or phrase identically several times [9]. So phonemic speech recognition should confront with big variation of the same phoneme and this causes degradation in phoneme recognition accuracy.

Currently, the majority of speech recognition systems are based on template or pattern recognition principles and methods. The main idea of these methods is that at first we prepare templates of those phonemic units that we want to recognize and later they are compared with tested feature vectors to find the closest match during recognition stage. Phoneme recognition is a problem which aims to find the class of phoneme to whom belongs part of speech signal. The simplest algorithm for template based phoneme classification – to compare features describing part of speech signal with template parameters of each phoneme and after that to prescribe to the class of phoneme that is closest under some selected criteria.

Such recognition requires relatively long time since the observed feature vector must be compared with all template parameters that system has in own repository. Another drawback of this approach is in the fact that the weights of errors speech – inside of similar phonemes and outside such group are different. Contrarily errors inside the group could be often corrected in the subsequent recognition stages. Misrecognitions between phonemes from different groups have stronger impact on the final recognition accuracy since often to correct such errors in a proper way is significantly more complex task. These drawbacks could be partly lessened by hierarchical phoneme recognition structure. Here recognition process is divided into two steps:

1. It is identified dependence of analyzed speech signal to the one of the main groups of phonemes (vowel, semivowel, consonant, etc.).
2. Then recognition inside this group of phonemes is carried on to make a final decision.

Theory of phonetics interprets Lithuanian phonemes as tree type phonetic hierarchy where nodes of the tree represents phonemes and are grouped into the some groups (vowels, consonants, etc.). Simplified and adopted to the set of phonemes from LTDIGITS speech corpora example of such tree is presented in Figure 1. Usually these phonemes trees are called dendrograms. More about the hierarchies of phonemes one could find in appropriate literature on phonetic theory [8, 14].

Based on these assumptions and motivations we presented here attempts to construct a discrimination system for sonant phonemes (sonants are all vowels and semivowels l, m, n, r ,v) and fricative consonants.

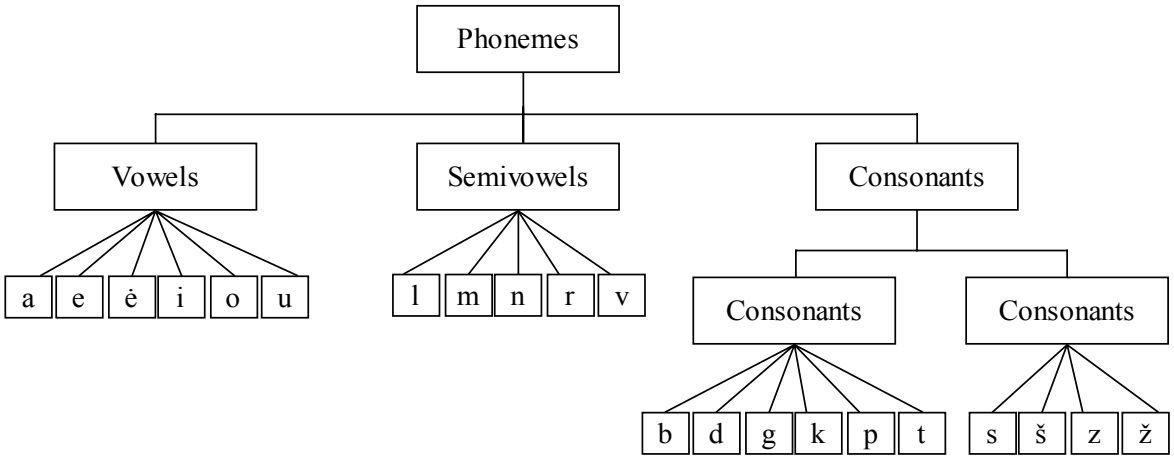

Figure 1. Hierarchical structure of phonemes from LTDIGITS speech corpora

3. Similar investigations

Currently more and more attention in the literature is devoted to the modeling of speech recognition systems based on the acoustical and phonetic knowledge. Briefly, acoustical phonetic speech recognition method could be characterized as follows: spectrum coefficients or various acoustic events describing parameters are used to divide speech signal into appropriate parts representing phonemes – basic element of human speech. Then individual segments or linguistically important events are used as allophones that consist of vectors of phonetic parameters.

T. Koizumi and others [12] used structural phoneme recognition. Experiments were carried out using the corpora of Japanese speech that contained 52140 Japanese words pronounced in the silent environment by a single speaker. Feature vectors were obtained by

filtering short-term speech signal spectrum with 16 filters evenly spaced in the Bark scale. Classifier has been realized using *Multilayered Neural Network*s or RNN (*Recurrent. Neural Networks*). During experiments phonemes were brought into 6 groups – voiced and unvoiced plosive consonants, voiced and unvoiced fricative consonants, nasal consonants and vowels. They achieved an increase in recognition accuracy of 3.2 percent when initial classification into the phoneme groups was implemented (from 84.9 percent till 88.1 percent).

Ahmed M. Abdelatty Ali and others [1,2] implemented a structural consonant recognition system. In this system classification is based on the logical rules obtained from the analysis of such phonetic – acoustic properties as spectrum, magnitude, place of articulation, voiciness/unvoiciness and duration. Experiments were performed using TIMIT speech corpora, 60 speakers with 7 different phrases recorded in different American English dialects. Phonemes were brought into such groups as plosive and fricative consonants, affricates. Further they were brought into voiced and unvoiced and even further into labials, palatals, alveolars, etc. They achieved about 92 percent recognition accuracy of consonants.

Liu [13] presented a system for detection of acoustic events in continuous speech. The system was able to recognize three different events – sleekness, plosiveness and voiciness. Starts and ends of vowels are denoted by sleek regions, plosiveness is denoted near the place of plosion of plosive consonants and for voiced regions there are denoted starts and ends of voiced consonants. He got error rates of 5 percent, 14 percent and 57 percent when comparing with the results of manual expert segmentation.

A. Juneja and C. Espy-Wilson [10] performed experiments with the recordings from TIMIT corpora. In these experiments they compared performance of HMM based approach and hierarchical classification methods. Speech signal was classified into 5 classes: silence, vowels, sonorants, fricative and plosive consonants. When performing experiments with HMM they used feature vectors containing 12 MFCC coefficients and their first and second derivatives (delta and delta-delta features). For recognition they used context-independent, three states, left-to-right HMM with diagonal covariance matrix and 8 Gaussian mixture components for each HMM state. This model recognized phonemes with 64.9 percent accuracy. For hierarchical recognition they used methods belonging to the group of SVM (*support vector machines*) methods. Feature vectors contained various phonetic – acoustic parameters such as energy of signal, format trajectories and energies in the various frequency bands. This method allowed achieving 68.1 percent recognition accuracy.

Similar investigations were carried out in Lithuania also. Recently G. Daunys [4] described features that could enable to classify phonemes according the place of co articulation. Possibilities to develop a methodology for Lithuanian speech segmentation and phoneme recognition were discussed also.

As could be seen, phoneme recognition based on the phonetic – acoustic knowledge is sufficiently applicable and perspective method that could allow achieving higher general speech recognition accuracy level.

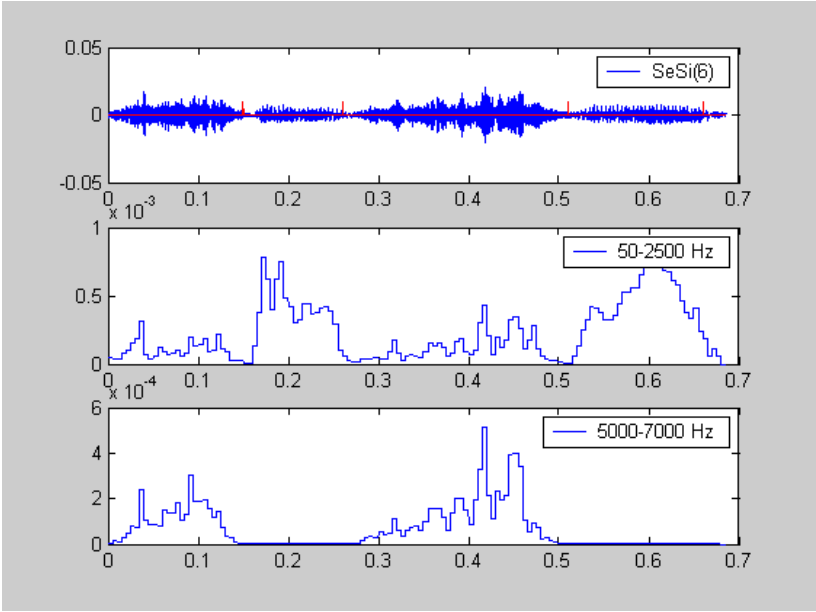

Figure 2. Energies of word "šeši" in low and high frequencies

4. Experiments

The main objective of our investigation is to try to construct a model and to find suitable features which could better discriminate sonant phonemes and fricative consonants and to create logical rules which could ensure best results of classification into these two groups.

A speech signal is called sonant if its energy in low frequencies is significantly bigger than energy in higher frequencies. It is a specific property of fricative consonants that in higher frequencies is concentrated more energy than in higher frequencies of sonants. Taking into account these facts we realized a system consisting of two bandpass filters with such passbands: low frequencies filter 50 –2500 Hz and high frequency filter 5000-7000 Hz. The passband of low frequency filter was selected from the results of research in experimental phonetics area [3]. There was observed that the first two formants of Lithuanian phonemes tend to concentrate in this zone. The passband of high frequencies filter we defined ourselves analyzing results of impact of filter passband on the accuracy of fricative consonants recognition and selecting the most efficient.

The filters were realized using Butterworth second order filter prototype:

$$K(w) = \frac{1}{\sqrt{1 + \left(\frac{w}{w_0}\right)^{2n}}};$$

were w_0 – cutoff frequency, n – filter order.

The filtered signal was divided into 10 ms length frames with 5 ms overlapping. For each frame the energy was calculated using the following formula:

$$E = \log \sum_{m=1}^{M} s(m)^2 ;$$

were $s(m)$ – m-th sample of frame.

An example of performed analysis is presented in Figure 2, where we see the oscilogram of the Lithuanian word „šeši" (six), below there are energies of low frequencies and high frequencies (rows 2 and 3). As could be seen in the figure the energy of sonants in higher frequencies is lower than the energy of fricatives.

Trying to reduce variations caused by different speakers such as distance from the microphone, loudness, etc. we used the ratio of energies of high frequencies and low frequencies the so called ratio of fricativity Frik(i). It was obtained using formula:

$$Frik(i) = \frac{Ead(i)}{Ezd(i)};$$

where $Ead(i)$ – high frequencies energy of the i-th frame, $Ezd(i)$ – low frequencies energy of the i-th frame.

At the next step we tried to find experimentally threshold of fricativity ratio and to compare with this level fricativity of each analyzed frame. If this level is above or is equal to the threshold level then the current frame was assumed to be part of fricative sound, vice versa – to sonorant sound.

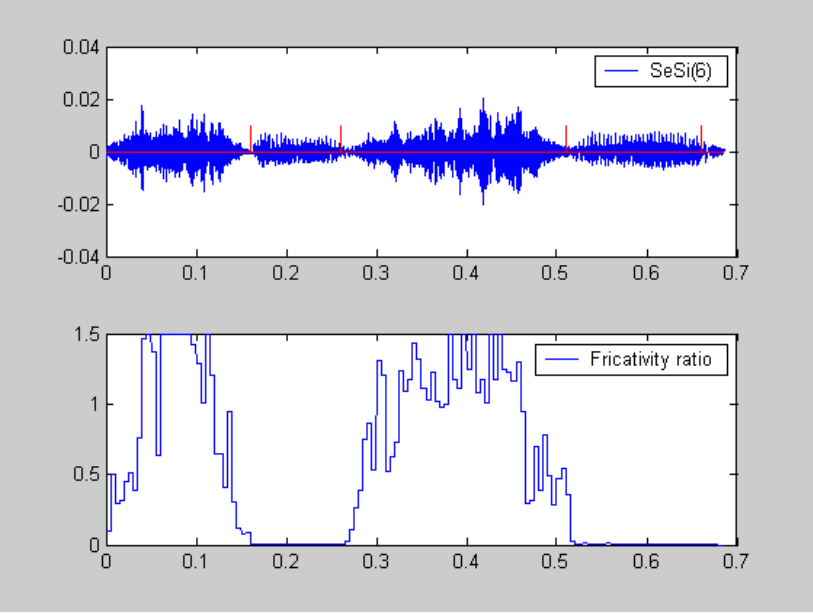

Figure 3. Fricativity ratio of the word "šeši"

We used LTDIGITS corpora in our experiments. Corpora contain recordings of 220 female and 130 male speakers. Each speaker pronounced 10 phrases. First six phrases contain sequences of Lithuanian numbers in random order. Next phrase contains Lithuanian control words such as start, stop, pause,

etc. Two phrases were specially designed to investigate peculiarities of nasals recognition in different contexts. Recordings were done in silence using 16000 Hz sampling rate. All speech data were labeled by experts both on word and phonemic level.

We used phrases fro LTDIGITS corpora pronounced by 100 speakers (50 male, 50 female). There were used about 19000 phoneme realizations in this experiment.

5. Results

The experiment provided overall 98.75 percent phoneme group's recognition accuracy. More details can be found in Table 1.

Table 1. Recognition accuracy o phoneme groups (in percent)

	Vowels	Semivowels	Fricative consonants
Number of phonemes	11482	4369	3000
Recognized as sonant	98.74	98.31	0.57
Recognized as fricative consonant	1.26	1.69	99.43

Analysis of vowels recognition errors showed that about 70 percent of errors occurred when recognizing vowel *i* at the end of the word. Some speakers don't pronounce this vowel at all.

Analyzing recognition errors of semivowels we observed that all errors occurred in the words *trys* (three) and *keturi* (four) with the phoneme *r*. Since acoustic features of Lithuanian speech have nor been explored sufficiently this effect still needs further explanation.

Analyzing recognition errors of fricative consonants we observed that the system frequently misrecognizes voiced fricative phoneme *z*. This phoneme exists at the end of the word *pauzė* in the LtDigits corpora speech recordings. If speaker stresses this word properly (stressed phoneme a) then phoneme *z* becomes very short and marginal and its fricativity ratio does not go beyond threshold of fricativity.

6. Conclusions and further work

Improved phonetic discrimination is one of the perspective methods to achieve overall improvement of recognition accuracy of automatic speech recognition system. Various strategies were proposed for phonetic speech recognition. They include automatic segmentation of speech signal, different feature types for recognition, various statistical methods for classification, etc. Our previous experience also proved that this is a promising approach.

One of the possibilities to improve phoneme discrimination is to assign phonetic unit under investigation to one of the phoneme groups. At the next step we will recognize this phoneme exactly inside this group using additional features or additional classification methods.

This paper presents a system of band-pass filters, which aims to discriminate Lithuanian speaker – independent groups of sonant phonemes and fricative consonants. Performed experiments allowed to achieve overall 98.75 percent recognition accuracy of these phoneme's groups.

In the future we plan to look for the features that could enable us to discriminate subsets of voiced and unvoiced phonemes and to augment experiments including plosive consonants as well.

References

[1] **A.M. Abdelatty Ali, J. Van Der Spiegel, P. Mueller.** An Acoustic-Phonetic Feature-based System for Automatic Phoneme Recognition in Continuous Speech. *IEEE ISCAS*, *May* 1999, *Proc. Vol*. III, 118-121.

[2] **A.M. Abdelatty Ali, J. Van Der Spiegel, P. Mueller.** Acoustic-Phonetic Features for the Automatic Classification of Stop Consonants. *IEEE Transactions on Speech and Audio Processing*, *Vol*. 9, 2001, 833-741.

[3] **D. Balšaitytė**. Vocalism of contemporary Baltic languages. *Problems and methods of experimental phonetics research. Sankt-Peterburg*, 2002, 17-23, (*in Russian*).

[4] **G. Daunys, D. Balabonas.** Classification of Sound using Decision Tree. *In Proceedings of conference Information technology, Kaunas. Technologija*, 2005, 277-282, (*in Lithuanian*).

[5] **K. Driaunys, V. Rudžionis, P. Žvinys.** The classification of Lithuanian language phonemes through the application of Fisher linear discrimination function and mel frequency cepstral coefficients. *Information Sciences*, ISSN 1392-0561. *Vilnius, VU leidykla*, 2004, T 31, 213-218, (*in Lithuanian*).

[6] **K. Driaunys, V. Rudžionis, P. Žvinys.** Mel frequency cepstral coefficients analysis of Lithuanian phonemes. *Human language technologies: the Baltic perspective: the* 1*st Baltic conference, Riga, April,* 2004, 162-165.

[7] **K. Driaunys, V. Rudžionis, P. Žvinys.** Lithuanian Speech Recognition by Improved Phoneme Discrimination. *In Proceedings of the second Baltic conference on Human Language Technologies. Tallinn*, 2005, 173-178.

[8] **A. Girdenis.** Theoretical foundations of phonology. *Vilnius: Petro ofsetas*, 1995, (*in Lithuanian*).

[9] **X. Huang, A. Acero, H. Hon.** Spoken Language Processing. *A Guide to Theory, Algorithm and System Development*. ISBN 0-13-022616-5. *Prentice Hall*, 2001.

[10] **V. Juneja, C. Epsy-Wilson**. Speech segmentation using Probabilistic Phonetic Feature Hierarchy and Support Vector Machines. *In Proceedings of International Joint Conference on Neural Networks, Portland, Oregan*, 2003, 675-679.

[11] I. **Kandratavičius.** Recognition of Lithuanian words in real-time computer telephony system. *Doctoral thesis*, 2001, (*in Lithuanian*).

[12] **T. Koizumi, M. Mori, S. Taniguchi, M. Maruya**. Recurrent Neural Networks for Phoneme Recognition. *ICSLP* 96, *Proceedings, Fourth International Conference, Vol*. 1, 3-6 *October* 1996, 326 –329.

[13] **S.A. Liu.** Landmark detection for distinctive feature based speech recognition. *J. Acoustic Soc. Am*., 100(5), 1996, 3417-3430.

[14] **A. Pakerys.** Phonetics of appellative Lithuanian language. *Vilnius: Mokslas*, 1986, (*in Lithuanian*).