ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392-1215 2006. Nr. 6(70)
ELEKTRONIKA IR ELEKTROTECHNIKA

SIGNAL TECHNOLOGY
T 121
SIGNALŲ TECHNOLOGIJA

Syllable-Phoneme based Continuous Speech Recognition

S. Laurinčiukaitė, A. Lipeika
Institute of Mathematics and Informatics,
A. Goštauto str. 12, LT-01108 Vilnius, Lithuania; e-mails: sigita.lau@mch.mii.lt; lipeika@ktl.mii.lt

Introduction

State-of-the-art speech recognition systems are based on the use of the sub-word units, i.e. words in such systems are modelled as sequences of previously defined phonemes, syllables or from acoustic signal derived units. We can assign all listed units to one of the two linguistically or acoustically defined groups of units. Majority of the systems prefer linguistically defined units and, especially, phoneme-based recognition. This preference is reasoned by simplicity in building such systems for phoneme set, i.e. sub-word unit set is known a priori. Such choice justify itself in achieved word error rate (WER), which varies according to differences in vocabulary size, speech type, speaker variability and complexity of speech recognition systems between 2-31.4% WER [1]. Recently more attention was paid to alternative units to phoneme unit with intention to use information, coded in acoustic of speech and not properly sized. Co-articulation, prosody phenomenon is more likely to be reflected in units of longer duration.

Modelling of speech recognition for Lithuanian remains in a stage of attempts to follow different approaches trying to establish mainstream in speech recognition. Speaking about speech recognition unit, word [2] and phoneme [3, 4] units were mainly picked up as basic recognizable units in speech recognition systems. Mentioned speech recognition systems yield 0,17-20% WER. The least error rates were achieved with word based speech recognition system with vocabulary of 12 words. Increase in vocabulary size, number of speakers and speech type required to change speech recognition approach and increased WER.

Comparison of phoneme, syllable and word units in recognition of isolated words showed that better performance was inherent to word and syllable units [5]. We predicted that similar results could be obtained for continuous speech recognition systems.

Baseline system

In recent speech recognition modelling we investigated syllable-phoneme based continuous speech recognition. This article presents directions of further investigation in the same field and results of experiments, which are based on earlier gained results.

Speech corpus. Speech corpus that was used in our experiments is LRN0 (Lithuanian Radio News, Version 0). Corpus contains records of 23 speakers with correct and clear pronunciation. The records of 10 speakers make 89% of the speech corpus.

Speech corpus was divided into training (10 hours, 6564 sentences), development (2 minutes, 50 sentences) and evaluation (19 minutes, 360 sentences) data sets. Speech corpus is accompanied by words-to-phonemes transcription lexicon. It contains ~18 000 entries. Phonetic transcriptions and stress marks were created manually referring to [6, 7]. Semi-automatic lexicon transformations were carried out in the process of word syllabification.

System description. System is based on Hidden Markov model (HMM) methods and has been built using HTK toolkit [8]. Syllabification of words in lexicon was implemented according algorithm description, given in [9]. It produced a finite set of syllables and phonemes of 2 959 items. Finite syllable-phoneme set H_1 of 290 phonemes and syllables was chosen after sequence of experiments. It was formed according to syllable frequencies in lexicon. Not basic syllables were exchanged into syllable and phoneme combinations after investigation of two different decomposition schemes.

We explored number of states and mixtures in a model from model parameters. A standard left-to-right model topology with no skips was used. The number of states, which tends to model sub-word unit duration, in each syllable model was set according to phoneme count. Traditionally duration of phoneme is expressed in 3 states. Following this, counting phoneme number in syllable and multiplying by 3 could express duration of syllable. Different schemes for increase of mixtures in states of models were investigated. As they didn't show improvement, standard scheme of increase of mixtures was set.

Models were trained using 13th-order feature vectors of Mel Frequency Cepstral Coefficients (MFCC) and their delta and delta-delta values (feature vectors were 39-dimensional), extracted from raw speech waveforms. Training and testing stages were followed as in [10]. Training process involved augmentation of number of mixtures in each model and each state to 4, after each iteration performing model training. It was agreed, that testing would be carried out after each augmentation of mixtures and training step on derived model set. For comparison score we consider recognition results achieved with acoustic models with 4 mixtures per state. Development and evaluation sets were used for testing. Development set was used for testing after each training procedure. Evaluation set was used after the best syllable-phoneme set was obtained.

All possibilities of formation of phoneme and triphone models were tested, which might have impact on recognition accuracy. Distinct experiment was performed forming models on the same speech corpora.

The performance of recognition system was measured by word level accuracy, defined as:

$$WA = \frac{N - S - I - D}{N} \times 100\%, \qquad (1)$$

where N is number of words in test or development set in total, S – number of word substitution errors, I – number of word insertion errors and D is the number of word deletion errors.

Task definition

Research issues, we specified in earlier experiments, building syllable-phoneme based speech recognition system, involved syllable set formation approach and a role of syllables in the lexicon. We paid little attention to model accuracy of distinct syllables, restricting ourselves just to improvement of accuracy of the phoneme models. Different lexicons, syllable-phoneme sets, phoneme model sets were tested. Consequently we got speech recognition system in which syllable-phoneme unit set was formed according to lexicon, word transcriptions in lexicon were made of sequences of syllables and phonemes. The best syllable-phoneme unit set and according it formed set of acoustic models were evaluated on two data sets: development and evaluation, which differ in duration: 2 min and 19 min. Recognition results in WER were 29,85% and 42,94% respectively.

As it was mentioned, we didn't investigate accuracy of syllable models, syllable sets were formed according repetition counts of each syllable in the lexicon. These two points we set for further investigation.

Accuracy of any models, used in speech recognition, is of great importance. Main task in training of acoustic models is to extract all possible information from acoustic speech signal and to encode it in models in such a way that models could reflect acoustic characteristics of defined speech signal segment as good as possible. Commonly, about accuracy of models one decides from performance of all models in speech recognition. Accuracy of distinct acoustic models isn't calculated and estimation criteria aren't known. It leaves a gap in more deep understanding of speech recognition errors.

We set a goal to test one criterion for such model evaluation and according to accuracy of distinct models to modify our syllable-phoneme unit set. We predicted that such solution could guide to another investigation point – syllable-phoneme unit set formation according to structure, pattern of syllable, i.e. according to quality of syllable that could be supplementary to frequency criteria.

Theoretical framework

We evaluated our syllable-phoneme acoustic models that we got in previous experiments. Firstly, we applied following approach to get material that we can use for calculations. Acoustic model set was used in recognition of the same training material that was used in building of acoustic models without lexicon of words with transcriptions. Instead of this lexicon we used simple list of syllable and phoneme units. Such recognition pattern ensures that syllable-phoneme sequence is compiled just according acoustic characteristics of speech signal without a priori knowledge reflected in lexicon. Ability of acoustic models to find in acoustic signal segment, which represents syllable or phoneme that acoustic model tries to imitate, is observed. In such style we get a set of sentences, composed from syllables and phonemes with their boundaries in speech signal.

Reference patterns of the same sentences are formed in process of recognition in the following way. For reference patterns we used acoustic models, lexicon and syllable-phoneme level transcriptions of sentences. Viterbi alignment fixes boundaries of units, given in transcription of that sentence, in speech signal. These reference patterns can't be considered as perfect. To the moment modern technologies aren't able to perform this task better then human.

Accomplishing comparison of test and reference patterns, two characteristics for each acoustic model were calculated: count of that model in all reference patterns P and count of the same model in all test patterns AT.

Second step was to choose distinct model evaluation criteria, based on above mentioned two characteristics. The most natural way is to calculate simple unit accuracy value URA:

$$URA = \frac{AT}{P} \times 100\%. \qquad (2)$$

This criterion was chosen for distinct model evaluation.

Experimental setup

The best syllable-phoneme set and according it trained acoustic model set, named H_1, was a result of first experiment. This material was used in our further investigation. Syllable-phoneme set H_1 consists of 290 units: 63 phonemes and diphthongs and 227 syllables. Syllable set is composed of syllables with different number of phonemes: two-phonemes (169), three-phonemes (58).

Evaluation of above laid out approach allowed to calculate recognition accuracy of distinct units. The results revealed that syllables with more then two-phonemes have higher accuracy in comparison to phonemes or two-phoneme syllables. Investigation of phoneme models was performed in earlier experiment. Accuracy of treephoneme models was high enough. This motivated us to investigate two-phoneme syllables. In Table 1 few examples of syllables with recognition accuracy are shown. All syllables were ranged according alphabetical order of the first component of syllable – consonant and recognition accuracy.

Table 1. Examples of two-phoneme syllables with recognition accuracy *URA*, given in %[1]. SN – syllable name

SN	%	SN	%	SN	%	SN	%
bu:	71.90	ju:	47.35	lio	70.16	žu	90.14
bu	68.02	jo	46.87	lu	64.69	ža	87.54
bi	65.48	je	35.56	liu:	64.52	ži	82.93
ba	64.82	ji	27.82	la	59.28	ži:	81.60
bo	55.75	ju	21.61	liu	57.09		

Investigating recognition accuracy of distinct syllables, we looked closer at every syllable structure seeking to find out exceptional features in pattern or structure of syllable that influence high possibility of syllable to be recognized. It is the questioning of the point that every researcher follows – is there any alternative way of choosing syllables for syllable-phoneme set to quantative criteria of syllable repetition count in training data or lexicon.

Approach 1. Observation of syllables recognition accuracy showed: 1) syllables that starts with *š*, *ž*, *z* and *č* have high recognition accuracy, but 2) combination of any sound with *i* (*di*, *gi*, *ki*, *li*, *mi* et cetera), has very low recognition accuracy, 3) all syllables that start with consonant *j* have recognition accuracy of less then 50%, 4) some syllable pairs as *le* and *lia*, *re* and *ria*, *se* and *sia* model very similar acoustic. These remarks prompts few recommendations for syllable-phoneme set re-formation and we compiled new syllable-phoneme unit set H_1P, where:

- six units (*di, gi, ki, pi, ti, bi*) were represented by one model;
- each of three syllable pairs (*sia-se*, *lia-le*, *ria-re*) were represented by one unit;
- seven units that start with consonant *j* were removed;
- eighth units that start with vowel were removed;
- instead of 24 removed units we added new ones with three-phonemes in a syllable (to maintain the same number 290 of units and models).

According to the new syllable-phoneme set H_1P correspondent acoustic model set was formed and tested. Also another syllable-phoneme set H_2P was derived from H_1P not adding 24 new syllables.

Approach 2. The same list of recognition accuracy of syllables, a part of which is shown in 1 Table, was transformed ranging syllables according alphabetical order of second component of syllable – vowel and its recognition accuracy. Then separate groups according vowel was investigated paying attention to the first component – consonant. The aim was to find units of similar acoustic and to exchange them by one unit that has higher recognition accuracy. Similarity of consonants was stated if their pronunciation and articulator movements were similar. To find these relations we used [11]. Decline in number of units in the set was compensated with addition of new units. One more syllable-phoneme set H_3P was formed.

Approach 3. Formation of one more syllable-phoneme set has no links to laid out approaches. With this one we tried to get affirmation that syllable-phoneme based recognition is better then selection of any symbol combination according to their repetition counts. Combinations of two- and three symbols were extracted revising all words in lexicon separately, i.e. symbol combinations reflect just inter-word relations of symbols. From two separate lists of two- and three symbol combinations 227 items were selected according to repetition count of two- and three-symbol syllables observed in set H_1. This set was ranged according to symbol combination repetition counts, named as PS_1, and according to number of symbols in combination and symbol combination repetition counts, PS_2. After construction of lexicon using these symbol sets, difference in number of units of PS_1 and PS_2 sets appeared: 225 and 290 respectively.

All above mentioned syllable-phoneme sets are summarized in Table 2.

Table 2. Syllable-phoneme sets, derived from H_1

Syllable-phoneme set name	Size	Description
H_1	290	Base set
H_1P	290	24 unit difference from H_1
H_2P	277	23 units removed from H_1
H_3P	290	Some syllable models are paired
PS_1	225	Two- and three symbol combinations
PS_2	290	Two- and three symbol combinations. Difference from PS_1in size of set.

Results and discussion

According to syllable-phoneme unit sets, summarized in Table 2 acoustic model sets were formed and evaluated in recognition of development and evaluation sets. Results are shown in Table 3, given in word accuracy (WA) and listed for each acoustic model set with 1, 2, 3, 4 mixtures per state for development set and for acoustic model set with 4 mixtures per state for evaluation set.

Table 3. Recognition results for syllable-phoneme sets H_1P, H_2P, H_3P and PS_1. WA – word level accuracy

Syllable-phoneme set name	WA (development set)				WA (evaluation set)
	1	2	3	4	4
H_1	52.24	62.94	64.93	70.15	58.06
H_1P	46.52	60.20	65.42	69.15	55.71
H_2P	46.77	60.95	64.43	69.15	56.12
H_3P	45.27	62.19	64.18	66.42	53.62
PS_1	46.02	54.98	58.21	62.19	52.33
PS_2	48.26	57.21	58.71	63.18	53.39

[1] The sign ":" in the Table 1 mark long vowel. In the case of *ju:* one unit represent two *jų* and *jū*; in the case of *ži:* – *žų* and *žū*; *bu:* - *bų* and *bū*.

As syllable-phoneme set H_1P gave the closest recognition score to set H_1, we calculated recognition accuracy of distinct acoustic models of set H_1P and observed recognition accuracy change of the same models in sets H_1 and H_1P. We observed trend that recognition accuracy of all acoustic models tend to decline. At the same time recognition accuracy of new three-phoneme syllables are quite high.

Conclusions

We proposed new criterion for evaluation of distinct acoustic models. According to it, new discriminative features in syllable structure and pattern were investigated. Three approaches of investigation were taken, in which few syllable-phoneme unit and model sets were formed according to: 1) *URA*, 2) *URA* and first component of syllable – consonant. By the third approach we examined recognition using any symbol combination instead of syllables. These syllable-phoneme model sets were tested and compared to the best model set from previous investigation in respect to recognition accuracy *WA*. Experiment results affirmed that standard syllable formation pattern is more effective and simple.

References

1. **Holmes J., Holmes W**. Speech Synthesis and Recognition. Second Edition. New York: Taylor & Francis, 2001.
2. **Lipeika A., Lipeikienė J., Telksnys L.** Development of isolated word speech recognition system // *Informatika.* – 2002. – 13(1). – P. 37–46.
3. **Raškinis G., Raškinienė D.** Building medium-vocabulary isolated-word Lithuanina HMM speech recognition system // *Informatika.* – 2003. – 14(1). – P. 75–84.
4. **Filipovič M., Lipeika A.** Development of HMM/Neural Network-Based Medium-Vocabulary Isolated-Word Lithuanian Speech Recognition System // *Informatika.* – 2004. – 15(4). – P. 465–474.
5. **Laurinčiukaitė S.** On different kinds of speech units based isolated words recognition of Lithuanian language // Proceedings of the First Baltic Conference on Human Language Technologies: The Baltic Perspective. – Riga, 2004. – P. 139–143.
6. **Vaitkevičiūtė V.** Fundamentals of Pronunciations for Lithuanian. Lexicon. – Vilnius: Pradai, 2001 (in Lithuanian)
7. **Keinys S.** Lexicon of Modern Lithuanian. – 4th edition. – Vilnius, 2000 (in Lithuanian).
8. **Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Valtchev V., Woodland Ph.** The HTK Book ant Toolkit. – Cambridge University Engineering Department Speech Group, 2003.
9. **Kasparaitis P.** Syllabification and hyphenation of word (in Lithuanian), 2005 // Lecture notes. Reachable by Internet: http://www.mif.vu.lt/~pijus/CL/cl.htm.
10. **Šilingas D., Laurinčiukaitė S., Telksnys L.** Towards Acoustic Modeling of Lithuanian Speech // Proceedings of SPECOM'2004. – St Petersburg, 2004. – P. 326–333.
11. **Pakerys A.** Lietuvių bendrinės kalbos fonetika. – Vilnius: Enciklopedija, 2003.

Submitted 2006 02 27

S. Laurinčiukaitė, A. Lipeika. Syllable-Phoneme based Continuous Speech Recognition // Electronics and Electrical Engineering. – Kaunas: Technologija, 2006. – No. 6(70). – P. 91-94.

Syllable-phoneme based continuous speech recognition research is presented in this article. This investigation is continuation of previous experiments. Achieved results are set as a point for further investigations.

Evaluation of distinct acoustic models is considered and an evaluation criterion is proposed. According to proposed criterion recommendations were made for syllable-phoneme unit and model set formation. Few such sets were formed and tested, evaluating their performance in total speech recognition accuracy and recognition accuracy of distinct models. At the same time we tested syllable set formation pattern according to quality of syllable that could be additional to selection according to its repetition counts. Achieved results suggest using frequency criteria. Bibl. 11 (in English; summaries in English, Russian and Lithuanian).

С. Лауринчюкаите, А. Липейка. Распознавание непрерывной речи, основанной на фонемах-слогах // Электроника и электротехника. – Каунас: Технология, 2006. – № 6(70). – С. 91–94.

Рассматривается распознавание непрерывной речи, основанной на фонемах-слогах. Настоящее исследование является продолжением раннее проведенных экспериментов, результаты которых являются опорной точкой для дальнейших исследований.

Выдвигается вопрос оценки отдельных акустических моделей и предлагается критерий их оценки. Согласно предложенному критерию и с его помощью полученными рекомендациями формируется несколько множеств, состоящих из фонем-слогов. Эти множества формируются по оценке их эффективности в распознавании речи как согласно критерию точности распознавания, так и по новому критерию оценки отдельных моделей. Таким способом проверяется количественный метод создания множества слогов, когда слога, рядом с подбором по частоте, оцениваются по их структуре.

Полученные результаты дают преимущество способу создания множества слогов, основанному на стандартном количественном критерии. Библ. 11 (на английском языке; рефераты на английском, русском и литовском яз.).

S. Laurinčiukaitė, A. Lipeika. Skiemenimis ir fonemomis grįstas ištisinės šnekos atpažinimas // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2006. – Nr. 6(70). – P. 91–94.

Pristatomi skiemenimis ir fonemomis grįsto ištisinės šnekos atpažinimo tyrimai. Šie tyrimai yra tęsinys ankstesnių eksperimentų, kurių rezultatai laikomi atskaitos tašku atliekant tolesnius tyrimus. Iškeliamas atskirų akustinių modelių vertinimo klausimas ir siūlomas vertinimo kriterijus. Pagal pasiūlytą kriterijų ir jį taikant gautas rekomendacijas formuojamos kelios skiemenų ir fonemų modelių aibės, vertinant jų efektyvumą tiek pagal bendrąjį šnekos atpažinimo tikslumo, tiek pagal naująjį atskirų modelių vertinimo kriterijų. Taip tikrinamas kiekybinis skiemenų aibės sudarymo būdas, kai skiemenys, be dažniais grindžiamo parinkimo, vertinami jų struktūros požiūriu. Gauti rezultatai teikia pirmenybę standartiniam kiekybiniu kriterijumi grįstam skiemenų aibės sudarymo būdui. Bibl. 11 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).