Text Selection for Speech Recognition Training Procedures under Sub-Word Units Variety

Nina Vasylyeva

Speech Science and Technology Department
Int. Research/Training Center
for Information Technologies and Systems, Kyiv, Ukraine
ninel@uasoiro.org.ua

Abstract

In this paper we investigate approaches to select a set of sentences for speech samples to train acoustic models for Ukrainian both TTS and ASR systems. An algorithm that is not widely known is introduced and another one is applied. Several sub-word units are analysed: phoneme, phoneme-triphone and open syllable. Some experimental results are given and discussed.

1. Introduction

One of crucial problems that have to be solved when recognition or a speech synthesis system is developed is the availability of a proper speech corpus for the system training and testing. Particularly, speaker must pronounce some text for system to form a speaker voice file (passport) that describes all phonetic-acoustic variety, pronouncing peculiarities for a person [1].

The coverage of all phonemic segments might be attained for the account of artificially generated words like in [2]. But more contributory way, particularly for a speaker, is to select automatically some sentences or words from a database of phonetically transcribed natural text.

The methods which are used to select sentences from the phonetically transcribed database can be divided into 2 groups. One of them consists of methods that enable to select sentences containing all phonetic events with approximately uniform frequency distributions. Such sentences are usually called phonetically rich sentences [3]. The other group includes methods that can be used to select “naturally” balanced sentences, i.e. sentences containing phonetic events according to their frequency of occurrence in natural speech. Such sentences are called phonetically balanced sentences.

Here we also rise a problem of structurally-hierarchical knowledge representation of speech signals. What level of detalization must be taken for basis, which unit is appropriate for processing and subsequent use at the phoneme level of speech patterns hierarchy? These are still open questions.

The basic unit overview is presented in Section 2. Next, in Section 3, is described the text selection algorithms. Some experimental results and their analysis is presented in Section 4.

2. Overview of Basic Units

We will consider such elementary units of division as: (1) phoneme, (2) phoneme-triphone, (3) opened syllable.

A phoneme is the least sense-separating phonetic unit that is physically realized in speech and articulation.

The basic set of Ukrainian phonemes counts 58 units: a, A, o, O, u, U, i, I, yi, Yi, e, E, b, b', v, v', h, h', g, g', d, d', zh, zh', z, z', j, k, k', l, l', m, m', n, n', p, p', r, r', s, s', t, t', f, f', kh, kh', ts, ts', ch, ch', sh, sh', dz, dz', dzh, dzh', # (sign ' means softening or palatalization of a consonant) [4]. Every phoneme is characterized by place and method of its producing, articulation motions and duration.

To take into account the phenomenon of coarticulation we consider a phoneme in context of adjacent phonemes, so called phonemes-triphones. The set of phoneme-triphones for every language is unique and makes a phoneme-triphone alphabet.

Phoneme-triphone transcription in a phoneme-triphone alphabet is formed on the basis of phonetic text by universal rule of phoneme-triphones joining: phoneme of right context of previous phoneme-triphone passes to the terminal one of the following phoneme-triphone, and terminal phoneme of previous phoneme-triphone passes to the left context of the following phoneme-triphone.

Since the basis alphabet of Ukrainian phonemes includes 58 phonemes, the theoretical amount of phoneme-triphones ($58^3 = 195112$) makes prospects of recoding a training sample containing all these units unreal: it is about 10 hours if one assume that each item occurs only once. Actually, analyzing a Ukrainian orthoepical dictionary [5] only 27 thousand phoneme-triphones were found. The analysis of the continuous speech textual corpus revealed grown up to 58 thousand of phoneme-triphones owing to the cross-word effect.

Third unit that will be considered in this article is a generalized opened syllable. The division of words on syllables is more natural, than division on phoneme-triphones. The opened syllable is always ends on a vowel letter with some exceptions as denoted in the language-dependent rules published in [6] for Ukrainian. Count of open syllables in the orthoepical dictionary is 9777 and a considered continuous speech textual corpus contains 41212 open syllables.

3. Methods for Text Selection

On the input we have an array of text *T*. This array turns out from the files of textual corps, which have the standard extensions *.txt, *.htm, *.html.

Task: to form some optimum great number of the sentences $K^* \subset T$, which contains the set of all acoustic-phoneme units - $K1$, occured in *T*. The least of orthographic characters in the final selection is the criterion of optimum.

To find a solution of the set problem, we introduce a method call it covering algorithm [10] and apply the greedy algorithm [7], [8], [9].

The coverage algorithm realizes an idea of full search.

On the first step for each unit occurred in the input text *T* we form a sequence of sentence numbers containing the phonetic unit. This is illustrated in Table 1. Here a number of sentence containing the phonetic unit is indicated with an asterisk.

Table 1. Artificial example of phonetic units „1” to „n” distribution in 12 sentences for coverage algorithm illustration

Sentences / Units	1	2	3	4	5	6	7	8	9	10	11	12
“1”	*		*	*	*		*		*			
“2”	*	*				*		*	*		*	
“3”		*			*	*			*		*	
“4”	*	*					*	*	*	*	*	*
…												
“n”	*		*	*			*			*		*

We get:

$$\text{“1”} \prec (1 \cup 3 \cup 4 \cup 5 \cup 7 \cup 9)$$

$$\text{“2”} \prec (1 \cup 2 \cup 6 \cup 8 \cup 9 \cup 11)$$

$$\text{“3”} \prec (2 \cup 5 \cup 6 \cup 9 \cup 11)$$

$$\text{“4”} \prec (1 \cup 2 \cup 7 \cup 8 \cup 9 \cup 11 \cup 12)$$

…

$$\text{“}n\text{”} \prec (1 \cup 3 \cup 4 \cup 7 \cup 10 \cup 12)$$

On the second step it is needed to get a direct production of all arrays which were got on the first step:

$$\begin{aligned}(\text{“1”} \bullet \text{“2”}) \prec (&(1\bullet 1)\cup(1\bullet 3)\cup(1\bullet 4)\cup(1\bullet 5)\cup(1\bullet 7)\cup (1\bullet 9)\cup \\ &\cup (2\bullet 1) \cup (2\bullet 3) \cup (2\bullet 4) \cup (2\bullet 5) \cup (2\bullet 7) \cup (2\bullet 9) \cup \\ &\cup (6\bullet 1) \cup (6\bullet 3) \cup (6\bullet 4) \cup (6\bullet 5) \cup (6\bullet 7) \cup (6\bullet 9) \cup \\ &\cup (8\bullet 1) \cup (8\bullet 3) \cup (8\bullet 4) \cup (8\bullet 5) \cup (8\bullet 7) \cup (8\bullet 9) \cup \\ &\cup (9\bullet 1) \cup (9\bullet 3) \cup (9\bullet 4) \cup (9\bullet 5) \cup (9\bullet 7) \cup (9\bullet 9) \cup \\ &\cup (11\bullet 1)\cup(11\bullet 3) \cup (11\bullet 4) \cup (11\bullet 5) \cup (11\bullet 7) \cup \\ &\cup (11\bullet 9))\end{aligned}$$

Finally, on *n*-th step we evaluate ((((“1” • “2”) • “3”) • “4”)... • “*n*”) which consists of minimal subsets of sentences containing units the initial set *T* includes. We take the shortest subset.

It is possible to shorten calculations using conjunctive and disjunctive rules of absorption.

Applying these rules for our illustration we take such a result:

$$\begin{aligned}(\text{“1”} \bullet \text{“2”}) \prec (&1\cup (2\bullet 3) \cup (2\bullet 4) \cup (2\bullet 5) \cup (2\bullet 7) \cup \\ &\cup (6\bullet 3) \cup (6\bullet 4) \cup (6\bullet 5) \cup (6\bullet 7) \cup (8\bullet 3) \cup \\ &\cup (8\bullet 4) \cup (8\bullet 5) \cup (8\bullet 7) \cup 9 \cup (11\bullet 3) \cup \\ &\cup (11\bullet 4) \cup (11\bullet 5) \cup (11\bullet 7))\end{aligned}$$

The greedy algorithm is based on the assumption that the selection of the optimal sentence in each step will lead to a list of sentences that is close to the global optimal selection.

4. Text Selection Experimental Research

In experiments we operated with the mentioned speech units. Grapheme-to-phoneme conversion was performed automatically on basis of the orthoepical dictionary [5] and reading rules for Ukrainian texts with pointed stresses.

4.1. Coverage and greedy methods comparison

On a relatively small corpus of total 321 sentences (12257 phonemes) we tested the coverage algorithm and compared its results with the results of the greedy algorithm. Speech unit here was a phoneme. Both algorithms selected 9 sentences containing all units but total length of sentences the coverage algorithm produced was shorter: 220 versus 248.

As we can learn from Figure 1 the coverage algorithm reaches its satiation faster and so is more effective theoretically giving the optimal decision in global understanding. Unfortunately, it takes too much memory resources.

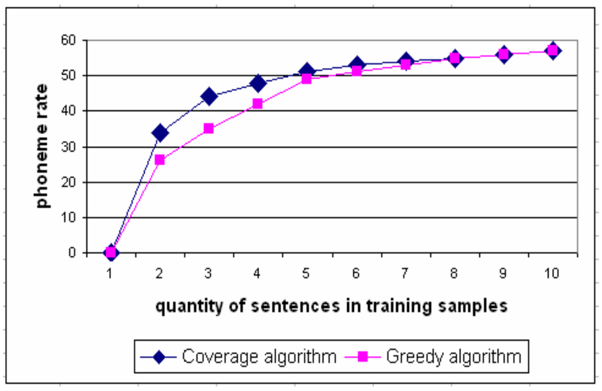

Figure. 1. Dynamics of new phoneme appearance with processed sentences growing for greedy algorithm and coverage algorithm

For coverage algorithm the main memory allocation during the work with diphone and phoneme-triphones rises substantially under condition of a set of 321 sentences containing 1239 diphones and 6078 triphones.

Such memory huge expenses makes the coverage algorithm not trackable on contemporary computers.

4.2. Phoneme-triphone and open syllable phonemic units

The text selection for phonemes-triphones and open syllables was carried out using the greedy algorithm. The following text sources were exercised in series of experimental text selections:

1) orthoepical electronic dictionary (isolated words) – 1874743 items [5];

2) a list of most used Ukrainian isolated words – 137639 items;

3) text corpus – a set of electronic documents, with the total amount of sentences exceeding 300000.

These results are given in Figures 2, 3 and Table 2.

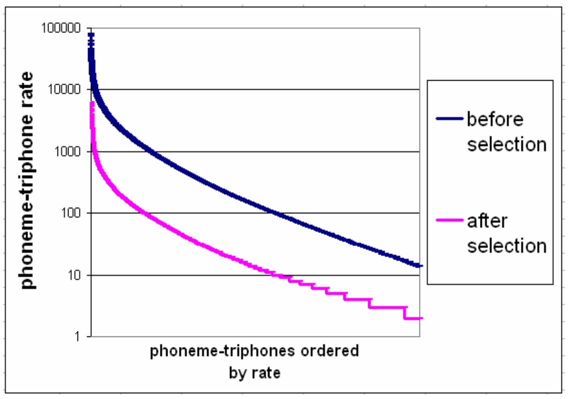

Figure 2. Distribution of phoneme-triphones in initial and output text corpora.

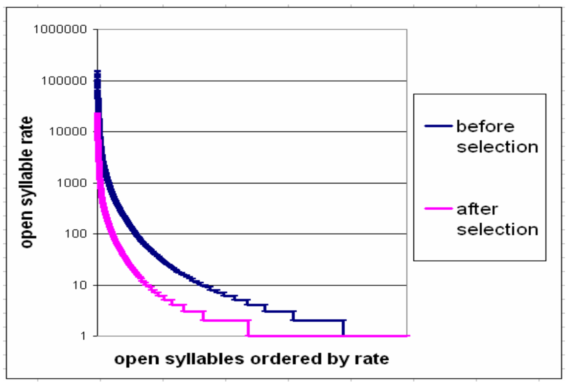

Figure 3. Distribution of open syllables in initial and output text corpora.

Table 2. Results of text selection for both phoneme-triphones and open syllables

Unit type / Corpus style	Amount of units / output sentences	
	Phoneme-triphone	Open syllable
Text corpus, 313057 sentences	57829 / 14615	41212 / 32096
A dictionary, 1874742 word forms	27300 / 11482	9777 / 9594
Rate dictionary, 137639 words	18307 / 7325	4961 / 4728

5. Syllable-based Speech Recognition Experimental Research

For training procedure was used a text formed in accordance with [12]. The text contained separate phoneme balanced words under conditions of the alphabet containing 55 basic Ukrainian phonemes including a phoneme-pause. Then a phoneme-based speaker voice file (passport) was formed. A speaker pronounced the phonetically rich training sample of above 2113 words containing 20353 phoneme realizations in each of three microphones having unlike acoustic features. Acoustic models accordingly to were trained and refined for each basic phoneme [11], particularly taking into account its both acoustic variability and occurrence. Each phoneme model had three states and 1 to 6 Gaussian mixtures in accordance to the phoneme occurrence and variability.

To perform a syllable recognition we built a free-syllable order grammar based on first 3200 syllables among sorted by occurrence ones. In this series of experiments we considered an open syllables united with a set of terminal consonant phones. This was taken into account in the grammar.

The control sample contained 2000 separate words. These words were taken from the top of the Ukrainian rate dictionary. The next word taken had to contain at least a one new triphone. Then the conventional HTK-based automatic phoneme recognition was carried out [11].

As the result:

- 54,7% of words were correctly recognized;
- 42,2% of words had some defects (doubling of vowels, extralinguistic words (at the beginning and at the end of words, and sometimes in the middle of complicated words), interchange of letters in a word);
- 3,1% of words were quite wrong recognized.

6. Conclusion

The paper deals with the problem of text selection for training procedures considering different phonemic units and types of initial text. Selected text is less than initial one in 10 and more times. Open syllables are more applicable for corpora of isolated words. Open syllables approximation capability significantly degrades relatively to phoneme-triphone unit by the reason of cross-words effect.

The syllable-based recognition results are promising.

As far the coverage algorithm appeared not trackable we plan to simplify it and then to compare with greedy algorithm on phoneme-triphone and open syllable units.

We considered phonemic variety but intonation variety is also important and must be investigated in future work.

7. References

[1]. Taras K. Vintsiuk, M.M. Sazhok. Speaker Voice Passport for a Spoken Dialogue System. Proceedings of the 3rd International Workshop "Speech and Computer" - SPECOM'98, St.-Petersburg, 1998.

[2]. Sazhok M.M. Automated means for data and knowledge base forming for Ulrainian text-to-speech conversion. – PhD thesis. Kyiv - 2004.

[3]. Gibbon D., Moore R., Winski R. (eds.): Design consideration and text selection for BREF, a large French read-speech corpus. In: Proc. Of the ICSLP (1990), 1097-2000

[4]. Taras K. Vintsiuk, Tetiana V. Liudovyk. Phonetic Knowledge Base for Ukrainian. Proceedings of the 3rd International Workshop “Speech and Computer” – SPECOM'98, St.-Petersburg, 1998.

[5]. Shyrokov V., Monako V. Organization of national lexicographic framework resources. – Movoznavstvo №5, 2001.

[6]. L. Shevchenko, V. Rizun, Yu. Lysenko. Current Ukrainian. Kyiv, Lybid, 1993.

[7]. Goncharov E., Kochetov Yu. Behavior of probabilistic greedy algorithm for stage location problem. Sampling analysis and operations research, vol. 6 (1999), №1, 12-32.

[8]. Korupolu M., Plaxton C., Rajaraman R. Analisys of a local search heuristic for facility location problem. – Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, 1998, 1-10.

[9]. Paraschenko M. Lagrange heuristics for location problem with power limitations. Works of XI Baikal international school-seminar “Optimization methods and its applications”. Irkutsk, 1998, 175-178.

[10]. Taras K. Vintsiuk. Recognition of hand-written character using methods dynamic programming. Cybernetics and Computer Engineering, vol. 3: Pattern Recognition. Kyiv, Naukova dumka, 1969, pp. 52–77.

[11]. Young S.J. et al., HTK Book, version 3.1, Cambridge University, 2002.

[12]. Nina Vasylyeva, Mykola M. Sazhok, Text Selection for Training Procedures under Phoneme Units Variety, *Proceedings of the 10nd International Workshop "Speech and Computer", SPECOM’2005*, Patras, Greece, 2005, pp. 629-631.