TURKISH LVCSR:
TOWARDS BETTER SPEECH RECOGNITION FOR AGGLUTINATIVE LANGUAGES

Kenan Çarkı, Petra Geutner, and Tanja Schultz

Interactive Systems Laboratories
University of Karlsruhe (Germany)
tanja@ira.uka.de

ABSTRACT

The Turkish language belongs to the Turkic family. All members of this family are close to one another in terms of linguistic structure. Typological similarities are vowel harmony, verb-final word order and agglutinative morphology. This latter property causes a very fast vocabulary growth resulting in a large number of out-of-vocabulary words. In this paper we describe our first experiments in a speaker independent LVCSR engine for Modern Standard Turkish. First results on our Turkish speech recognition system are presented. The currently best system shows very promising results achieving 16.9% word error rate. To overcome the OOV-problem we propose a morphem-based and the Hypothesis Driven Lexical Adaptation approach. The final Turkish system is integrated into the multilingual recognition engine of the GlobalPhone project.

1. INTRODUCTION

For languages like English many large vocabulary continuous speech recognition engines have been evaluated on several different tasks. Recently the interest increased on LVCSR systems in Asian languages like Chinese, Japanese and Korean. Furthermore, projects like SQALE [1] focus on transferring the evaluation paradigms and training methods to languages spoken in Europe like French, German, Spanish etc. However, so far there have been no attemps for Turkish Large Vocabulary Continuous Speech Recognition (LVCSR). This has several reasons: First, there is a lack of speech databases and text corpora for the Turkish language, and knowledge sources like pronunciation dictionaries are not available yet. Second, Turkish is very different from Indo-European languages because its morphology is agglutinative and suffixing. This means that the inflection, the derivation and other relationships between words in a sentence are done by constantly concatenating suffixes to the word stem. Therefore, the vocabulary growth rate is very high resulting in a large number of out-of-vocabulary words (OOV). As a consequence poor recognition results are achieved when using Turkish words as dictionary units. The following example illustrates the morphological structure of the Turkish language:

Osman-lı-laş-tır-ama-yabil-ecek-ler-imiz-den-miş-siniz
(English: behaving as if you were of those whom we might consider not converting into Ottoman, see [2]). It can be seen easily, that agglutination results in a word length that is exceptional for the Turkish language.

In this paper we present our experiments on Modern Standard Turkish, which is the most widespread language of the Turkic family spoken by about 65 million speakers. For all experiments we use the Turkish part of our GlobalPhone database, which is briefly introduced in the first section of this paper. The second section describes important properties of the Turkish language and resulting problems for LVCSR. The paper concludes by presenting several recognition experiments and results.

2. GLOBALPHONE DATABASE

Turkish LVCSR presented here is evaluated in the framework of the GlobalPhone project. The database of this project currently consists of 15 languages. In each language about 100 native speakers were asked to read 20 minutes of political and economic articles from a national newspaper. The speech was recorded at 16kHz sampling rate and 16 bit resolution in office quality, using a close-talking microphone. The corpus is fully transcribed including spontaneous effects like false starts and hesitations. For further details of the GlobalPhone project refer to [3].

	Total	Training	Test	Evaluation
Speakers	100	78	11	10
Utterances	6872	5418	240	124
Words	112K	87K	4K	2.5K
Vocabulary	16K	12.6K	2K	1.3K
Length	17h	13h	40min	20min

Table 1: Turkish speech database

2.1. Speech database

For the training and evaluation of the Turkish LVCSR systems we used the Turkish part of this database which consists of 72 female and 28 male speakers, most of them speaking Standard Turkish, only a few speaking slight Anatolien dialect. The range in age is 13 to 53 years, on average 27.5 years. Table 1 gives the number of utterances, spoken words, vocabulary, and total hours of spech data used for training and testing. The average length of an utterance is 8.9sec with about 16.5 spoken words per utterance.

2.2. Text Corpus

For the LVCSR experiments trigram language models have been built using a Kneser/Ney backoff scheme for unseen n-grams. Whereas the collected speech data is sufficient for acoustic modeling, the corpus of 112K spoken words is still too small for accurate trigram estimation. Thus, we collected additional text data from the internet as shown in table 2. The final language model was built by interpolating the read articles with the internet corpus of 15.63 words with an interpolation factor of 9:1. The trigram perplexity of this language model is 280, and the OOV-rate on the test set based on a 30K dictionary is 14.2%.

Source	Words	
	in-domain	out-of-domain
Zaman	1.02 Mio	1.20 Mio
Milliyet	3.92 Mio	3.81 Mio
Hürriyet	0.35 Mio	0.58 Mio
Superhaber	1.39 Mio	3.09 Mio
Xn Online	0.11 Mio	0.16 Mio
Total	6.79 Mio	8.84 Mio
	15.63 Mio	

Table 2: Turkish text corpus

3. PROPERTIES OF TURKISH

3.1. Phonemic inventory

The Turkish vowel inventory is small and very symmetric. Eight phonemic vowels are grouped into foursomes with respect to the features of height, backness and rounding. No diphthonges can be found in the Turkish language. All vowels are shown in table 3, written with symbols according to the IPA conventions. The vowels of the native vocabulary in Turkish language are phonemically short. However, lengthening occurs as a result of borrowed words for example. For further details refer to [4].

The consonant inventory of Turkish shown in table 4 is very compact. The Turkish language does not allow consonant clusters. Native speakers tend to break up those clusters occurring in proper names and borrowed words by inserting vowels like in the Japanese language. The English

	+Round		-Round	
	-Back	+Back	-Back	+Back
+High	y	u	i	ɨ
-High	ø	o	e	ɑ

Table 3: Turkish vowel inventory [IPA]

word *club* for example is written in Turkish as *klüb* and is pronounced like */k u l y b/*. The vowel insertion is applied according to the vowel harmony rules. Recently vowel insertion occur in written text as well.

	Bil	Lad	Alv Pos	Pal	Vel	Glo
Plosive	p b		t d		k g	
Nasal	m		n			
Trill			r			
Fricative		f v	s z ʃ ʒ			h
Approximant			l	j		
Affricate			ʧ ʤ			

Table 4: Turkish consonant inventory [IPA]

In our speech recognition engine we introduced models for the 8 vowels and 20 consonants showed in figure 3 and 4 plus one model for silence, and one for spontaneous effects respectively. The occurrences of long vowels in our Turkish speech database are too rare to be modeled by separate phonemes.

3.2. Pronunciation Dictionary

When reforming the writing system in 1928 the Latin script was adopted for Modern Standard Turkish adding two diacritics for the affricates ç /ʧ/, ş /ʤ/ and the sign ğ. Pronunciation alternations are written out, differing in this aspect from for example the German orthography. Therefore, the orthographic conventions correspond roughly with those of a broad phonetic transcription. As a consequence we were able to build most pronunciations by applying a grapheme-to-phoneme tool using a small number of simple context-free rules. Only the ğ must be dealt with context sensitive rules, since it causes lengthening of the preceeding vowel. Numbers, except for date and ordinal numbers, are pronounced regularly from left to right. The automatic generated dictionary was finally proof-read and postedited by native experts adding pronunciation variants for proper names, date expressions and acronyms.

3.3. Word length distribution rates

Figure 1 shows the number of phonemes per word in the training corpus and the dictionary. The distribution over length is symmetric, showing that the most frequent phoneme length in the corpus is 5. Typical words from the corpus of this length are *genel, büjük, jüsde, and sezim.* Particularly the distribution of phoneme length in the lexicon illustrates the agglutinative morphology of Turkish resulting in very long words.

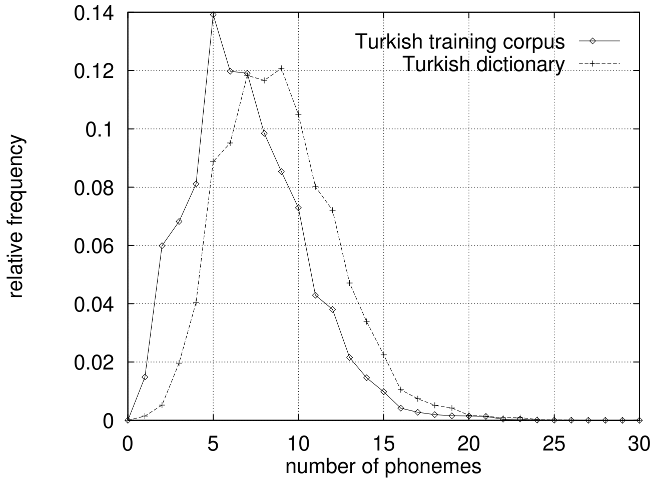

Figure 1: Phonemes per word in corpus and dictionary

3.4. Vocabulary growth rates

Long words might make it easier to acoustically distinguish them from each other, but results in high out-of-vocabulary rates, as can be easily seen from figure 2. It shows the self- and cross coverage for the Turkish language based on 15.63 Million words and a 6K test set of 3K different words. The cross coverage curve shows that even given a 500K vocabulary the OOV-rate is still above 5%. Assuming a dictionary size of 64K words, which is commonly used in speech recognition, we can expect to deal with an OOV-rate of around 15%.

4. RECOGNITION ENGINE

The Turkish LVCSR system was trained and evaluated using the Janus Speech Recognition Toolkit (JRTk) in two passes: First a preliminary recognition engine was built using Turkish words as dictionary units, without attacking the problem of high OOV-rates by adding all words of the test set to the dictionary. In the second step a morphem-based and the HDLA approach are applied to the resulting system to overcome the problems of agglutinative languages.

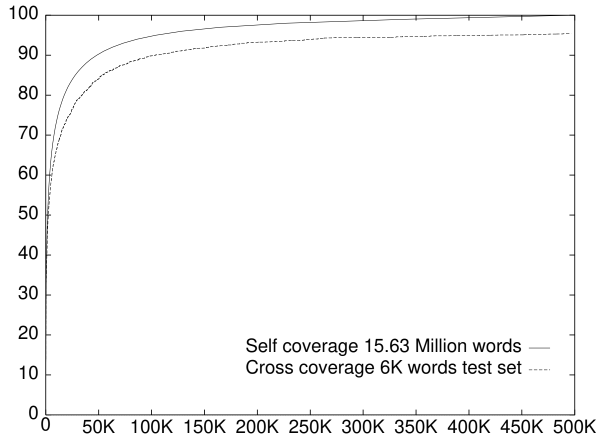

Figure 2: Vocabulary growth rates

We initialize the Turkish acoustic models applying our fast crosslingual bootstrap technique [5]. A four-lingual recognizer was used, which consists of German, Japanese, Spanish, and English models from a spontaneously spoken scheduling task. The phoneme mapping was done heuristically by picking the best matching IPA-counterpart. Initial word labels are written, Gaussian distributions are calculated using k-means clustering and trained along the previously written labels. The resulting models are used for writing improved labels, and the training procedure is repeated. For the first context dependent system the polyphonic tree of all occurring quintphones has been clustered down to 1500 models using linguistic motivated questions about the phonetic context. This results in a fully continuous 3-state HMM system, where each state is modeled by one codebook containing a mixture of 16 Gaussian distributions. The feature space of the Gaussians is based on 16 cepstra, power, and their first and second derivatives calculated from the 16kHz sampled input speech. After mean subtraction the number of features is reduced to 24 and 32 coefficients respectively by computing a linear discriminant analysis. Speaker normalisation is applied through a piecewise linear frequency transformation which compensates for different vocal tract length (VTLN). During training, optimal warping parameters for

Stage of Development	WE
Bootstrapped from multilingual seed models	61.1%
Iterative Training, Labelboosting, 15,7 Mio LM, Context Dependent system	25.0%
Data and Dictionary Corrections	20.0%
More Acoustic Models (1500 → 3000)	19.4%
More Dimensions (24 → 32)	18.7%
Speaker Normalisation through VTLN	18.0%
Currently best system	16.9%

Table 5: Word error rates [WE] for Turkish LVCSR

each speaker are estimated and iteratively optimized. Table 5 summarizes the progress of the Turkish LVCSR system in terms of word error rates. The currently best system achieves 16.9% word error rate.

5. AGGLUTINATIVE LANGUAGES

5.1. Morphem-based approach

Turkish morphology is agglutinative and suffixing with only a few exceptions to the one-to-one relationship between morphem and function. As a consequence morphems as recognition base units would provide a solution for speech recognition. Currently available morphological analyzer for the Turkish language suffer from morphological ambiguities. About 10% of input words need follow-up treatment by human experts [6]. In order to handle a 15.63 Million words corpus fully automatically we break up words into syllables applying Latex distinction rules. Afterwards these syllables were merged into larger units by defining word-positioned syllable classes. The problem of limited scope of a syllable-based language model is approached by class based 4-gram LMs. Table 6 summarizes the recognition results of three different systems resulting from merging the word position based classes. It can be seen that using smaller units for recognition decreases the OOV-rate dramatically. However, the lower OOV-rate does not reflect a better recognition performance, since the smaller units suffer from a higher acoustic confusability. Experiments for Korean LVCSR indicate that the locking of acoustic similarities by merging units is able to overcome this problem [7].

Systeme	Vocabulary	Splits	OOV	WE
Baseline	30000	1	15,3	34,1
Z1	14881	1,63	6,0	39,0
Z2	21868	1,24	7,6	35,7
Z3	17033	1,45	6,8	37,0

Table 6: Syllable-based LVCSR

5.2. Hypothesis Driven Lexical Adaptation

Using morphem-like small base units is one approach to counteract the fast vocabulary growth and resulting high OOV-rates of highly inflected languages. A second possibility is to allow a virtually unlimited recognition dictionary by adapting the vocabulary of the speech recognizer to the utterances to be recognized. To this end the Hypothesis Driven Lexical Adaptation (HDLA) algorithm [8] is applied to our Turkish system. Within this framework a first recognition run is performed on a baseline dictionary. The resulting word lattices and utterance-specific vocabulary lists are then used to look up similar words in a large fallback dictionary, consisting of all words included in the largest available Turkish text corpus. Different selection criteria for the adaptation procedure of the dictionary for the second recognition run can be used, ranging from morphological knowledge to distance measures either based on grapheme or phoneme level [9]. For our Turkish system a phonetic distance based adaptation has been performed. The baseline OOV-rate of 14.9% for a 30K vocabulary could be improved to 10.9% unknown words. This 27% improvement in the OOV-rate is comparable to results achieved in Serbo-Croatian and German. As for both of those languages significant word error rate reductions could be achieved, the OOV-rate improvement in the Turkish language is expected to also reflect in an improved word error rate for the ongoing recognition experiments.

6. CONCLUSION

We have presented first results for a Turkish large vocabulary speech recognition system. Our final system yields a word error rate of 16.9%. It is integrated into the multilingual recognition engine in the GlobalPhone framework. The special agglutinative morphology of the Turkish language has been exploited to conduct morphology-based recognition experiments as well as vocabulary adaptation in the HDLA framework. The latter was able to decrease the OOV-rate by 27%. This reduction is expected to reflect in an improved word error rate for the ongoing recognition experiments.

7. REFERENCES

[1] S.J. Young et al.: *Multilingual large vocabulary speech recognition: the European SQALE project*, Computer Speech and Language, vol 11, pp. 73-89, 1997.

[2] K. Oflazar, E. Göçmen, C. Bozşahin: *An Outline of Turkish Morphology*, Report on Turkish Natural Language Processing Initiative Project, 1994.

[3] T. Schultz and A. Waibel: *Language independent and language adaptive LVCSR*, Proc. ICSLP, pp. 1819-1822, Sydney 1998.

[4] J. Kornfilt: *Turkish and the Turkic Languages*, B. Comrie (Editor): The Worlds Major Languages, pp. 619-645, 1990.

[5] T. Schultz and A. Waibel: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets*, Proc. Eurospeech, pp. 371-374, Rhodes 1997.

[6] K. Oflazar and G. Tür: *Combining Hand-crafted Rules and Unsupervised Learning in Constrained-based Morphological Disambiguation*, Proc. ACL, Philadelphia, 1996.

[7] D. Kiecza, T. Schultz, and A. Waibel: *Data-driven Determination of Appropriate Dictionary Units for Korean LVCSR*, Proc. ICSP, Seoul 1999.

[8] P. Geutner: *Adaptive Vocabularies in Large Conversational Speech Recognition* PhD Thesis, University of Karlsruhe, Germany, 1999.

[9] P. Geutner, M. Finke, and A. Waibel: *Selection Criteria for Hypothesis Driven Lexical Adaptation*, Proc. ICASSP '99, Phoenix 1999.