ON DEVELOPING NEW TEXT AND AUDIO CORPORA AND SPEECH RECOGNITION TOOLS FOR THE TURKISH LANGUAGE

Özgül Salor[*], *Bryan Pellom*[**], *Tolga Çiloğlu*[*], *Kadri Hacıoğlu*[**], *Mübeccel Demirekler*[*]

[*]Department of Electrical and Electronics Engineering
Middle East Technical University, Ankara, Turkey
{salor, ciltolga, demirek}@metu.edu.tr

[**]Center for Spoken Language Research
University of Colorado, Boulder, USA
{pellom, hacioglu}@cslr.colorado.edu

ABSTRACT

This paper describes recent work towards development of new corpora and tools for Turkish speech research. This effort represents an on-going collaboration between the Center for Spoken Language Research (CSLR) at the University of Colorado and the Department of Electrical Engineering at the Middle East Technical University (METU). A new text corpus developed from Turkish newspapers' text is described. In addition, a 193-speaker audio corpus and pronunciation lexicon for the Turkish language is developed. We then describe our initial work towards porting Sonic, the CSLR speech recognition system, to the Turkish language. Results are shown for phonetic alignment and phoneme recognition accuracy using the newly constructed corpus and speech tools. It is shown that 91.2% of the automatically labeled phoneme boundaries are placed within 20 msec of hand-labeled locations for the Turkish audio corpus. Finally, a phoneme recognition error rate of 29.3% is demonstrated.

1. INTRODUCTION

Developing text and audio corpora is one of the biggest issues in speech technology. For languages like American English, phonetically rich and large corpora like TIMIT [1], Wall Street Journal and Switchboard are used. One of the biggest problems for Turkish speech processing researchers has been the lack of phonetically rich databases. There have been recent attempts to develop audio and text corpora for Turkish speech research. Interactive Systems Laboratories has collected a multilingual audio database, the GlobalPhone, which include speech from Turkish newspapers, read by 100 speakers, 20 minutes of each [2]. Moreover, METU's Informatics Institute is working on a corpus project in collaboration with Sabancı University in Turkey to develop a morphologically and syntactically annotated text corpora for Turkish [3].

In this paper, we present our work on developing a new corpus and speech recognition tools for Turkish speech research. Section 2 describes the development of a phonetically balanced text corpus and the audio corpus. Section 3 describes the speech recognizer and phone aligner developed for American English at CSLR. Issues related to system porting to Turkish are discussed. The paper concludes in Section 4 by presenting baseline phonetic alignment accuracy and phoneme recognition results for the new corpus.

2. LEXICON AND CORPORA

2.1. Development of the Lexicon

Modern standard Turkish is a phoneme-based language like Finnish or Japanese, which means phonemes are represented by letters in the written language. It may also be true to say that there is nearly one-to-one mapping between written text and its pronunciation. However, some vowels and consonants have variants depending on the place they are produced in the vocal tract [4]. For example, the letter *a* in the word *laf* [laf] is predorsal, while in *almak* [ɑɫmɑk], *a*'s are postdorsal. Therefore, 29 letters in the Turkish alphabet are represented by 45 phonetic symbols in [4].

METU has developed a new letter-to-phone conversion rule set that is based on the phonetic symbol set described in [4]. These rules have been developed by observing the phonetic transcriptions of the letters in the dictionary and determining the conditions in which they appear. These conditions are phonetic context and position in the word. Since the IPA symbols in [4] are difficult to use for our applications, symbols in the Speech Assessment Method Phonetic Alphabet (SAMPA) dictionary [5] have been matched to the symbols in [4] and they have been used. SAMPA has not been previously applied to the Turkish language, therefore some additions to the existing SAMPA symbols were necessary. However, SAMPA symbols have poor readability since they include characters such as numbers and punctuation symbols. This has led us to develop a new simplified alphabet called METUbet. The choice of symbol formatting in METUbet is similar to that used within ARPAbet for American English.

A mapping of the SAMPA characters to the METUbet characters is shown in Table 1. METUbet has 39 phonetic representations compared to the 45 phonetic SAMPA representations. The reason is that the *open-short* and *closed-long* forms of the letters *u, ü, o, ö,* and *i* are represented by the same phonetic symbol in METUbet. The *closed-long* forms of those letters appear when they are preceding *ğ*, soft g, which causes only the lengthening of those letters [4]. This does not need to be considered for the phonetic alignment and phoneme recognition using Hidden Markov Models.

IPA	SAMPA	METUbet	Example
α	A	AA	*a*nı
a	a	A	l*a*f
e	e	E	*e*lma
ɛ	E	EE	der*e*
i	i	IY	*i*ğde
ɪ	I	IY	s*i*m*i*t
ï	1	I	*ı*s*ı*
ɔ	O	O	s*o*ru
o	o	O	*o*ğlak
U	U	U	k*u*lak
u	u	U	*u*ğ*u*r
œ	2	OE	*ö*rtü
ø	5	OE	*ö*ğren
Y	Y	UE	*ü*mit
y	y	UE	d*ü*ğme
b	b	B	*b*al
d	d	D	*d*e*d*e
g	G	GG	kar*g*a
ɟ	g	G	*g*enç
h	h	H	*h*asta
ʒ	Z	J	mü*j*de
k	k	KK	a*k*ıl
c	c	K	kedi
l	L	L	*l*ey*l*ek
ɫ	l	LL	ku*l*
m	m	M	da*m*
n	n	NN	a*n*ı
ŋ	N	N	sü*n*gü
p	p	P	i*p*
r	r	R	*r*af
ɾ	R	RR	ı*r*mak
ɣ	4	RH	bi*r*
s	s	S	*s*e*s*
ʃ	S	SH	a*ş*ı
t	t	T	ü*t*ü
v	v	VV	*v*ar
ʋ	w	V	ta*v*uk
j	j	Y	*y*at
:ɪ	yy	Y	hu*y*
z	z	Z	a*z*ık
ʐ	zz	ZH	yo*z*
ʤ	DZ	C	*c*am
ʧ	TS	CH	se*ç*im
f	f	F	*f*asıl
:	:	GH	düğme
Sil	Sil	SIL	"silence"

Table 1: Mappings of IPA, SAMPA and METUbet

2.2. Text Corpus

Text data has been collected from web pages of Turkish newspapers. This corpus has been normalized to expand numbers and abbreviations. Words coming from other languages like proper nouns have been deleted. The normalized text includes 2,529,850 words. This text has been converted to METUbet symbols. Using this text corpus, we provide some insight into the frequency of the occurrence of triphones within the Turkish language. The number of triphones in terms of METUbet characters in this corpus was found to be 29,266. The most frequent 10 triphones in this corpus is shown in Table 2.

METUbet Triphone	Example Words	Occurrence Rate (%)
EE RR IY	evl***eri***ni	2.74
LL A RR	at***lar***ını	2.67
L EE RR	ev***ler***i	2.61
B IY RH	***bir***	2.54
N D AA	yanı***nda***	2.20
IY NN IY	ev***ini***	1.96
RR IY NN	evle***rin***de	1.70
A RR I	atl***arı***	1.69
IY L EE	***ile***	1.67
I NN I	at***ını***n	1.63

Table 2: Most frequent triphones in Turkish

2.3. Audio Corpus

An audio corpus has been developed for Turkish speech recognition research. To this end, a phonetically balanced set of sentences was constructed for Turkish. The sentences were developed by translating the first 2000 sentences of the American English TIMIT database into Turkish. Then they were converted to METUbet representations and the triphone frequencies of these sentences were compared to those of the corpora from the newspaper websites. The number of triphones that occur at least once in these sentences was 9,492. The most frequent triphones from both corpora were found to be highly correlated. 462 sentences have been added in order to ensure coverage for the most frequent 5000 triphones for Turkish. The augmented list includes 11,033 triphones. The resulting 2462 sentences have been used to develop the audio database.

The audio database is being collected by METU. For each speaker, a total of 40 sentences are randomly selected from the 2462-sentence database and recorded. To-date, 193 speakers (89 female and 104 male speakers) have been recorded. The age range is from 19 to 50 years with an average of 23.9 years. Our goal is to collect 500 speakers total. The speech is being collected in office quality with a Sennheiser microphone ME 102. The data is being digitally recorded with a Sound Blaster sound card on a PC at a 16 kHz sampling rate (16 bit, PCM format). Each recording session is accompanied by a text file that lists the 40 randomly selected sentences. In addition, the recording date, the age of the speaker and the geographic region in Turkey where the speaker has grown up is recorded. Subjects are collected from mainly the students, faculty and staff in METU.

The final audio corpus consists of audio files and associated text transcriptions. Audio files are checked for misreadings and repetitions. In cases of misreadings, either the corresponding text file is corrected or the sentence is deleted completely.

3. TURKISH SPEECH RESEARCH TOOLS

Development of new speech technologies for the Turkish language will require accurately labeled and annotated corpora. In this work, we consider porting the CSLR speech recognition toolkit (Sonic) [6] to the Turkish language. The resulting port has aided in the development of a new corpus that has been phonetically labeled at the word, phoneme, and HMM-state level. In the next section, we describe the CSLR recognizer and discuss methods and issues encountered during the port to Turkish. Finally, we evaluate the performance of the system by measuring the phoneme recognition error rate and phonetic alignment accuracy on a test set derived from the newly collected Turkish speech corpus.

3.1. Speech Recognizer

For this work we use *Sonic* [6], the University of Colorado large vocabulary speech recognition system. Sonic is a continuous density hidden Markov model (CDHMM) based recognizer. The acoustic models are decision-tree state-clustered HMMs with associated gamma probability density functions to model state-durations. The recognizer toolkit can be used for large vocabulary continuous speech recognition, keyword and phrase spotting, as well as phoneme recognition. Sonic incorporates speaker adaptation and normalization methods such as Maximum Likelihood Linear Regression (MLLR) [7], Parallel Model Combination (PMC), Jacobian Adaptation, and Vocal Tract Length Normalization (VTLN) [8]. Advanced language-modeling strategies such as concept language models [9] are also incorporated into the toolkit.

The recognizer implements a two-pass search strategy. The first pass consists of a time-synchronous, beam-pruned Viterbi token-passing search. Cross-word acoustic models and trigram language models are applied in the first pass of search. During the second pass, the resulting word-lattice is converted into a word-graph. Longer span language models can be used to rescore the word graph using an A* algorithm or to compute word-posterior probabilities to provide word-level confidence scores [10].

Sonic has been benchmarked on several standard continuous speech recognition tasks for American English and has been shown to have competitive recognition accuracy to other recognition systems evaluated on similar data. Performance metrics are shown in Table 3.

3.2. Phonetic Aligner

The phonetic aligner within the Sonic toolkit generates word, phoneme, and HMM-state level alignments of audio corpora. The aligner is typically used for acoustic training (by providing state-level alignments of the acoustic feature for decision-tree state clustering). Phonetic aligners can also be used for other applications such as generating phoneme positions for lip-synchronization, general speech analysis, or to provide initial phoneme locations for development of text-to-speech synthesizers. The alignment algorithm matches sequences of HMM states to extracted features by using the Viterbi algorithm. The aligner also automatically determines locations of silence or speaker pause and can determine the pronunciation of words from sets of alternate pronunciations.

Task	Vocabulary Size	Word Error Rate	Real-Time Factor
TI-Digits	11	0.4%	0.1
Communicator	3k	15.8%	1.6
WSJ	5k	5.9%	1.5
Switchboard	40k	32.9%	9.1

Table 3: Word error rate for the CSLR recognizer on several tasks: TI-Digits, DARPA Communicator telephone based travel planning domain, Nov'92 Wall Street Journal (WSJ) 5k test set and Hub5 Switchboard task. Real-time factors are for first-pass decoding on an 800 MHz Intel Pentium III.

3.3. Issues in porting from English to Turkish

Sonic uses the Sphinx-II phoneme symbol set [11]. Initialization of the recognizer's acoustic models to Turkish was performed by mapping Sphinx-II symbols to the acoustically nearest equivalents in METUbet. The mapping is shown in Table 4. We found that there was no acceptable mapping for the Turkish phoneme GH, soft g, that is used to denote lengthening of the previous vowel sound. Therefore we have not used it for the recognizer and the aligner applications. The aligner outputs soft g in word-level, but not in phone level alignments, but instead outputs the previous vowel in lengthened form.

Sphinx-II	METUbet	Sphinx-II	METUbet
AA	AA, A	M	M
AX	OE	N	NN, N
B	B	O	O
CH	CH	P	P
D	D	R	RR, R, RH
EH	EE, E	S	S
F	F	SH	SH
G	GG, G	SIL	SIL
HH	H	T	T
IX	I	UH	U, UE
IY	IY	V	VV, V
JH	C	Y	Y
K	KK, K	Z	Z, ZH
L	LL, L	ZH	J

Table 4: Sphinx-II to METUbet phonetic symbol mapping

The Turkish corpus was used to improve the accuracy of phonetic alignment system originally developed for English. Letter-to-Phoneme (Letter-to-METUbet) rules for Turkish have been used to develop a dictionary of the words in the 2462-sentence corpus. A set of decision tree questions was developed for Turkish and used for acoustic model training. As a beginning step, the 36 questions have been determined based on place and manner of articulation [4].

The first 100 speakers of the audio corpus were used to train Turkish acoustic models of the aligner. Using the initial mapping shown in Table 4, the corpus was aligned at the HMM-state level and models were then retrained using decision tree state clustering. The resulting aligner is capable of providing word-level and phoneme-level boundaries for Turkish. The phonemes are represented by METUbet symbols at the output of the aligner.

For phoneme recognition experiments, the 2.5 million-word text corpus has been converted to METUbet symbols using text-to-phoneme rules that we have developed. This corpus was used to develop a back-off trigram phoneme language-model.

4. EXPERIMENTS

Experiments were conducted by randomly selecting 20 speakers (10 male and 10 female) from a held-out test set. A total of 40 sentences from each of the speaker were aligned using the Turkish phonetic aligner. The alignments were corrected by hand and compared to the alignments produced by the automatic method. Results of comparing boundary misalignments from human-corrected and automatically labeled segments are shown in Table 5. Here we see that 53.7% of the misalignments are within 5 msec of the hand-labeled locations. Errors in pronunciation prediction introduced by letter-to-sound rules for Turkish are not considered in this work. These results are comparable to those obtained using TIMIT [12].

Phoneme recognition using decision tree state clustered HMMs was also performed on the test-set using a back-off phoneme trigram language model trained from the newspaper text corpus. Results of phoneme recognition experiments both with and without iterative unsupervised MLLR adaptation are shown in Table 6. Phone error rates are calculated with respect to the hand-corrected reference transcriptions. Here we see that the overall phone error rate was found to be 29.3%. To the best of our knowledge, the only phone error rate of Turkish that has been reported is 44.1% [2] with 29 phonemes without phoneme language modeling.

Misalignment Tolerance	Percent of Automatically Placed Phoneme Boundaries
≤ 5 msec	53.7%
≤ 10 msec	67.6%
≤ 20 msec	91.2%
≤ 40 msec	98.1%
≤ 60 msec	99.3%

Table 5: Percent of automatically placed phoneme boundaries within a fixed distance from the hand-labeled reference.

5. CONCLUSIONS

This paper has presented the work towards new corpora and tools for Turkish speech research. Moreover, a new phonetically balanced audio corpus of 193 speakers has been presented along with a text-corpus collected from newspapers in Turkish. Based on our work for speech recognition and phonetic alignment, we propose a phonetic symbol set, METUbet and demonstrate its use within a Turkish port of the CSLR's speech recognition toolkit. The resulting Turkish phonetic aligner yields misalignments of less than 5 msec in 53.7% of our test set phoneme transitions. This aligner has been used to provide word-level and phoneme-level alignments for the new Turkish audio corpus. A phoneme recognition system trained from the Turkish corpus was found to yield a phone error rate of 29.3% after MLLR speaker adaptation.

Gender	Non-adapted	Adapted (MLLR)
Male	30.7%	29.1%
Female	31.5%	29.6%
Overall	31.1%	29.3%

Table 6: Phone error rates for 20 Turkish speakers. Results are shown for a baseline system and for the same system with iterative unsupervised MLLR adaptation.

6. ACKNOWLEDGEMENTS

The authors thank Aylin Susar and Demet Uluşen for developing the text-to-phoneme rules for Turkish, and Hüseyin Mutlu, Erdem Koç, Ahmet Mutlu and Deniz Kurt for collecting the newspaper text corpus. Special thanks to Çağla Önür and Yücel Özbek for collecting the audio corpus. Özgül Salor's work at CSLR is supported by the Turkish Government through a combined doctoral program scholarship of TÜBİTAK, the Scientific and Technical Research Council of Turkey.

7. REFERENCES

[1] TIMIT Acoustic-Phonetic Continuous Speech Corpus. http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[2] Schultz, T. and Waibel, A., "Language-independent and Language-adaptive Acoustic Modeling for Speech Recognition", Speech Comm., Vol.35, pp.31-51, 2001.

[3] Developing a Morphologically and Syntactically Annotated Treebank Corpus for Turkish, http://www.ii.metu.edu.tr/~corpus/treebank/

[4] Ergenç, İ., Konuşma Dili ve Türkçe'nin Söyleniş Sözlüğü, Simurg Yayınevi, Ankara, Turkey, 1995.

[5] SAMPA, Computer Readable Phonetic Alphabet, http://www.phon.ucl.uk/sampa/home.htm

[6] Pellom, B. "Sonic: The University of Colorado Continuous Speech Recognizer", Technical Report TR-CSLR-2001-01, CSLR, University of Colorado, March 2001.

[7] Legetter, C. J. and Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech & Language, Vol. 9, pp. 171-185, 1995.

[8] Uebel, L.F. and Woodland, P.C., "An investigation into Vocal Tract Length Normalization", Proc. Eurospeech-99, Budapest, Hungary, 1999.

[9] Hacioglu, K. and Ward, W., "A Word Graph Interface for a Flexible Concept Based Speech Understanding Framework", Proc. Eurospeech-2001, Denmark, 2001.

[10] Hacioglu, K. and Ward, W., "A Concept Graph based Confidence Measure", Proc. ICASSP, 2002.

[11] Ravishankar, M. K., "Efficient Algorithms for Speech Recognition". Ph.D. Dissertation, Carnegie Mellon University, 1996.

[12] Pellom, B., Hansen, J.H.L., "Automatic Segmentation of Speech Recorded in Unknown Noisy Channel Characteristics," Speech Comm., Vol. 25, Nos. 1-3, pp. 97-116, Aug. 1998.