Large Vocabulary Continuous Speech Recognition in Greek: Corpus and an Automatic Dictation System

V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis,

C. Vosnidis, N. Chatzichrisafis and V. Diakoloukas

Department of Electronic and Computer Engineering
Technical University of Crete
{vas,doikon,pratsdim,ntsourak,vosnidis,nhatzi,vdiak}@telecom.tuc.gr

Abstract

In this work, we present the creation of the first Greek Speech Corpus and the implementation of a Dictation System for workflow improvement in the field of journalism. The current work was implemented under the project called Logotypografia (*Logos* = logos, speech and T*ypografia* = typography) sponsored by the General Secretariat of Research and Development of Greece. This paper presents the process of data collection (texts and recordings), waveform processing (transcriptions), creation of the acoustic and language models and the final integration to a fully functional dictation system. The evaluation of this system is also presented. The Logotypografia database, described here, is available by ELRA.

1. Introduction

The emerging technology of speech recognition offers nowadays a variety of applications, from telephone solutions to voice-controlled appliances and dictation systems. It is claimed that we speak seven times faster than we type, therefore there is an obvious necessity for the study and creation of systems that can transform speaker utterances to text. Many efforts have appeared commercially and in academic institutions in the last years. Little effort, however, has been done for the Greek language. Our work under the project "Logotypografia" was concentrated in the effort to create the first Greek dictation system and the "Logotypografia Database", which is publicly available.

The project had the following goals:

- The creation of the basic infrastructure for continuous speech recognition in Greek with the help of a journalistic organization.
- The adaptation of an efficient speech recognition system involving a large lexicon in Greek and its installation in the editing environment of the "Eleftherotypia" daily newspaper.
- Research and development in Greek large vocabulary continuous speech recognition.
- Preparing the foundations for the development of a 100% Greek speech recognition system.

Despite the challenge, all of the four goals were accomplished. It should be noticed that we integrated a large vocabulary, speaker independent Greek dictation system using the SRI Decipher speech recognition engine.

2. Speaker Characteristics

The speakers that took part in the recording sessions where staff of the "Eleftherotypia" daily newspaper, one of the most reputable in Greece. Specifically, 55 male and 70 female speakers took part, which constitutes a distribution of 44% and 56% respectively. The age distribution of the speakers is shown in Figure 1.

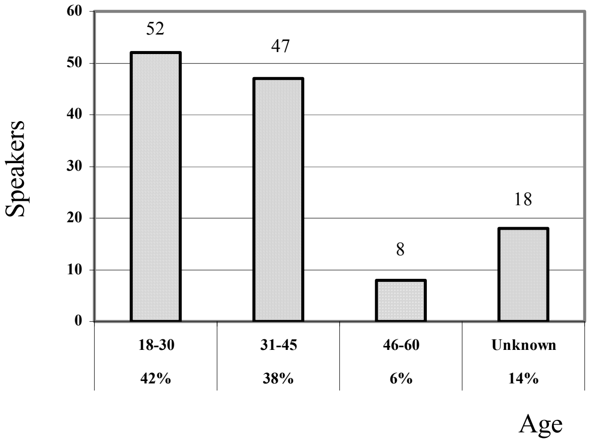

Figure 1: Speakers' Age Distribution

3. Recording Conditions

The recordings took place in three different environments, a sound proof room, a quiet environment and an office environment. 125 speakers completed 291 sessions. Each soundproof session consisted of 180 utterances, each quiet session of 150 and each office session of 150 utterances. 30 of the 180 utterances of the sound proof session were specially selected in order to have rich phonetic coverage. The total number of utterances that were collected was 46,020. Table 1 summarizes the data above.

	Sound Proof Room	Quiet Room	Office Room	
Number of Utterances	180	150	150	
Number of Sessions	79	110	102	
Total Utterances	14,220	16,500	15,300	**46,020**

Table 1: Number of utterances

From the initial number of utterances, 12884 utterances where discarded (46 sessions) due to recording and clipping problems, leaving us with 33136 utterances. With an average of 7.8 seconds per utterance, the total collected utterances were approximately 72 hours of speech. 30 additional sessions were recorded with spontaneous speech. Two different microphones were used, a desk microphone and a head-mounted close-talking microphone. 143 session where recorded with the Audio Technica ATM73a desktop microphone, having a cardioid polar pattern and 60 – 15kHZ frequency response. 102 sessions where recorded with the AKG C410 head-mounted microphone, having a cardioid polar pattern, 20 – 20kHZ frequency response and with the use of the Audio Buddy preamplifier. The SNR was calculated for each session. Figure 2 shows the SNR distribution.

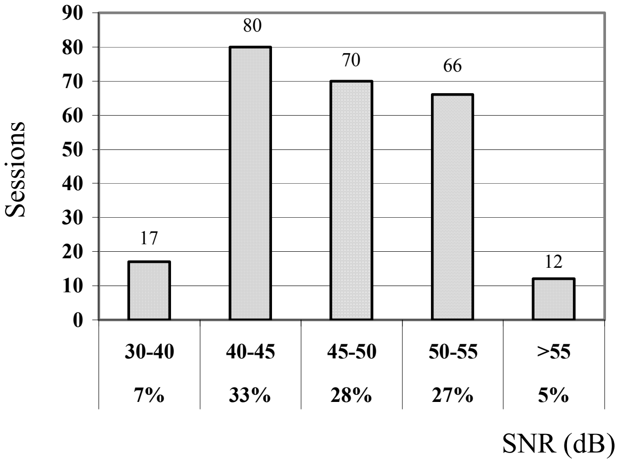

Figure 2: SNR Distribution

The format of the waveform files is NIST Sphere. Waveforms are encoded using PCM coding format, 16 kHz sampling rate and 2 bytes per sample.

4. Transcriptions

The corpus was split in two sets, one of 23,136 utterances transcribed by the speech recognition group of the Technical University of Crete and one set of 10,000 utterances transcribed by the Institute of Language and Speech Processing in Athens. There are markings for the utterance's orthography and several speech and non-speech events (e.g. mispronunciations, truncation, noise etc). During the transcription process we followed some basic rules and few of them are described in the list below:

- All transcribed sentences contain Greek words.
- Punctuation marks are not used except from the stress mark.
- Use of lower case letters.
- The numbers are expanded to words.
- Marks can be placed in the beginning of an utterance or can enclose any word or phrase.

During the process we used specific marks for the description of several sounds presented in the waveforms and marks to depict articulation problems during the reading of an utterance. Some of those marks were: "Noise" when an external noise is encountered, "Breath and "Clear Throat" with obvious usage, "Bad Reading" when the speaker is reading without consistency, "Bad Audio" when the waveform of an utterance is unacceptable, "Paf Noise" when the characteristic sound produced by a speaker to close to the microphone is encountered, "Hesitation" when the speaker hesitates to begin or continue reading, "Mispronunciation" used to show that a word of phrase was mispronounced etc.

Phonetic Category	Ph.	Example	Ph.	Example
Vowels				
	i	σπίτι	o	τόπος
	E	γενναίος	u	κουτί
	A	καλός		
Consonants				
Plosives	p	πόνος	d	ντύνω
	b	μπαίνω	k	κώνος
	t	τομή	g	γκρεμός
Fricatives	f	φεύγω	s	σκούπα
	v	βάζω	z	ζητώ
	T	θέλω	x	χάνω
	D	διαβάζω	G	γράμμα
Nasals	m	μένω	n	νάρκη
Liquids	l	λάμδα	r	ράφι
Palatals	c	καιρός	ly	ελιά
	C	χέρι, χιόνι	N	εννιά, μιά
	J	γέρος, γιατρός		
Others	-	(silence)	hes	(hesitation)
	exh	(exhale)	rej	(rejection)

Table 2: Greek Phoneme set

5. Acoustic Model

We used SRI's DECIPHER speech recognition system [2][3] to train the acoustic model. The system's front-end configured to output 12 cepstral coefficients, cepstral energy and their first and second derivatives. The cepstral features are computed with a fast Fourier transform (FFT) filterbank and subsequent cepstral-mean normalization on a sentence basis is performed. We used genonic HMM's with arbitrary degree of Gaussian sharing across different HMM states [3]. After experimentation we found the best number of genones to be 960, each with 32 Gaussian distributions. We used a set of 32 phonemes, specially designed for the Greek language by a linguist. This set is specified using the Computer Phonetic Alphabet (CPA). The CPA provides a system for easily expressing the phonemes in notation by the IPA (International Phonetic Alphabet) using a standard keyboard. Table 2 lists the phonemes used to express pronunciations with an example for each one.

It should be mentioned that we trained both gender-independent and gender-dependent models.

6. Text Preprocessing

The initial form of the articles collected from the "Eleftherotypia" daily newspaper contained text, which would degrade the performance of our language model; therefore we were forced to do text preprocessing. A list of actions, which led to the quality improvement of the text used for the creation of the language model, is mentioned below:

- The headings of the articles were removed, because they usually did not contain grammatically correct sentences.
- Articles that contained results of sport events, crossword questions etc. were also removed.
- The articles were split to sentences (one sentence per line). A sentence could be terminated by a full stop ("."), a question mark (";") or an exclamation mark ("!").
- The abbreviations in the articles were replaced with their respective expanded forms. Moreover, the person's first name, which was represented by a capital letter, was replaced by the corresponding word.
- The numbers were also expanded to words. For example, the number "112" was expanded to "one hundred twelve".
- Dates that were written with digits were replaced by word phrases. For example, the date "1/4/96" was replaced by the phrase "the first of April of ninety six".
- All upper case letters were converted to lower case.
- Finally, we corrected all the words that had erroneous stress marks and removed all the punctuation marks, except from the stress mark and the apostrophe.

In the preprocessing phase we used tools like Perl, Awk and Shell Scripts in a Solaris Environment.

7. Language Model

We used a standard back-off [5] trigram language model with Good-Turing smoothing [6] and vocabulary size of 64K words. The language model was trained with a text of 35M words, taken from the Eleftherotypia daily newspaper. The perplexity of the language model, evaluated on a different test text, was 174. The out-of-vocabulary (OOV) words were 3.55%. These rather high values for perplexity and OOV rate are typical for Greek, which is an inflectional language. Details about the language model can be found in [4].

8. Evaluation Tasks

Table 2 shows the word error rate (WER) and the recognition time (xRT) for the three acoustic models: Unisex for the gender-independent model and Male/Female for the gender dependent models.

Model	WER (%)	xRT
Unisex	21.01	6.16
Male	19.27	5.78
Female	20.85	5.85

Table 3: Word error rate and recognition time

9. Dictation System

As the final step of our work, we developed a graphical user interface (GUI) in order to complete the Greek Dictation System, named Logotypografos 1.0. This tool was created using the MFC C++ library for the graphical interface. A snapshot of the tool can be seen in the image below.

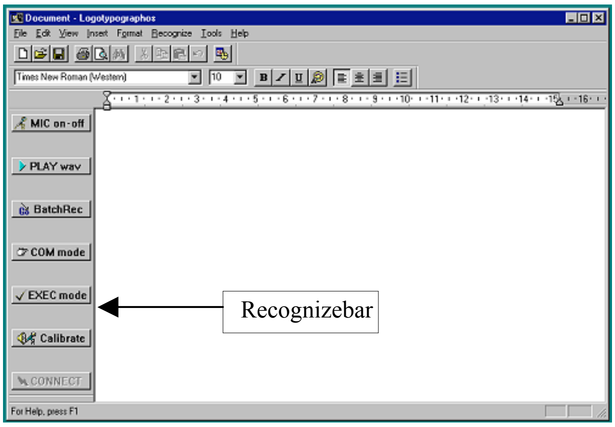

Figure 2: GUI snapshot

The tool can be used either as a voice editor or as a simple editor. The user can determine which model set (male or female) will be used during the recognition process. Beside the common toolbars (e.g. standard, formatting, etc.) the tool offers a special recognition bar. When the "*Connect"* button is pressed, the tool loads the specified recognition grammars (for male or female users). The user can initiate a new recognition task by just pressing the "*Recognize*" button. The result string is gradually created and presented to the user, who can see the partial results until the final one is formed. The last recorded and recognized utterance can be heard by pressing the "*Play*" button. When a sufficient number of utterances has been reached, the user can initiate a batch recognition process by pressing the "*GoBatchRec*" button. The user will be prompted with an estimation of the time needed in order for the process to be completed. During this second pass, the recognition engine uses a greater pruning threshold (threshold which determines the number of the active hypotheses used by the recognizer) and a larger Greek lexicon. In this way the user can speak a number of utterances and have a better offline result.

Another feature of the tool is the ability to execute a predefined set of voice commands. When the "*CommandMode*" button is pressed the user can say words like "bold", "italics", "comma" etc. with obvious results to the printed text. There is also the ability to run external applications from the tool. A predefined list of words is offered so that each of them can be associated with an application. When the "ExecutionMode" button is pressed and the user says an associated word, the specified application starts.

Finally, in order to achieve an acceptable input signal the tool offers the "Calibrate" button. The user is asked to read a sentence and the signal to noise ratio (SNR) is calculated for that specific waveform. In this way the user can calibrate any external devices (e.g. preamplifier), alter the distance between his mouth and the microphone or change the environment conditions that can affect the input signal.

The minimum requirements in order to use the dictation system are Windows NT, Pentium III 800MHz, 256MB RAM, Sound Blaster live or compatible.

10. Conclusions

In this work we presented the steps taken in order to create the first Greek Dictation system, "Logotypografos 1.0" and the speech corpus, "Logotypografia Database". Considering the complexity of the Greek language and the duration of the project (18 months), we have achieved respectable results. The system's performance will be greatly improved with methods of speaker adaptation, that we currently work and intent to integrate.

Additional information and demos can be found in the following address: http://www.speech.tuc.gr/new_page_2.htm

11. References

[1] Douglas B. Paul and Janet M. Baker. The Design for the Wall Street Journal-based CSR corpus. In Proc. Fifth DARPA Speech and Natural Language Workshop, pages 357-362. DARPA, Morgan Kaufman Publishers, Inc., 1992.

[2] H. Murveit, J. Butzberger, V. Digalakis and M. Weintraub. "Large Vocabulary Dictation Using SRI's DECIPHER Speech Recognition System: Progressive Search Techniques". *1993 IEEE ICASSP pp. II-319-II-322.*

[3] V. Digalakis, P. Monaco and H. Murveit. "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers". *IEEE Trans. on Speech and Audio Proc., pp. 281-289, July 1996.*

[4] D. Oikonomidis and V. Digalakis. "Stem-based Maximum Entropy Language Models for Inflectional Languages". *Submitted to Eurospeech 2003*.

[5] S. Katz. "Estimation of probabilities from sparse data for the language model component of a speech recognizer". *IEEE Transactions on Acoustics, Speech and Signal Processing, 1987.*

[6] I. Good. "The population frequencies of species and the estimation of population parameters". Biometrica, 1953.