Vergina: A Modern Greek Speech Database for Speech Synthesis

Alexandros Lazaridis[1], Theodoros Kostoulas[2], Todor Ganchev[3], Iosif Mporas[1], Nikos Fakotakis[1]

Wire Communications Laboratory, Department of Electrical and Computer Engineering,
University of Patras, 26500 Rion-Patras, Greece
[1]{alaza, imporas, fakotaki}@upatras.gr, [2]tkost@wcl.ee.upatras.gr, [3]tganchev@ieee.org

Abstract

The present paper outlines the Vergina speech database, which was developed in support of research and development of corpus-based unit selection and statistical parametric speech synthesis systems for Modern Greek language. In the following, we describe the design, development and implementation of the recording campaign, as well as the annotation of the database. Specifically, a text corpus of approximately 5 million words, collected from newspaper articles, periodicals, and paragraphs of literature, was processed in order to select the utterances-sentences needed for producing the speech database and to achieve a reasonable phonetic coverage. The broad coverage and contents of the selected utterances-sentences of the database – text corpus collected from different domains and writing styles – makes this database appropriate for various application domains. The database, recorded in audio studio, consists of approximately 3,000 phonetically balanced Modern Greek utterances corresponding to approximately four hours of speech. Annotation of the Vergina speech database was performed using task-specific tools, which are based on a hidden Markov model (HMM) segmentation method, and then manual inspection and corrections were performed.

1. Introduction

In Text-to-Speech synthesis (TTS) there are two major issues concerning the quality of the synthetic speech, namely the intelligibility and the naturalness (Dutoit, 1997; Klatt, 1987). The former refers to the capability of a synthesized word or phrase to be comprehended by the average listener. The latter represents how close to the human natural speech, the synthetic speech is perceived.

Over the last years the most widely used approach for high quality speech synthesis is the corpus-based unit selection technique (Campbell and Black, 1996; Hunt and Black, 1996; Black and Taylor, 1997a; Möbius, 2000). This approach is mainly based on runtime selection of the appropriate units of speech from the database and the concatenation of them with no or almost no speech processing of the selected speech units apart from the part where the concatenation takes place (Hunt and Black, 1996). In parallel with corpus-based unit selection speech synthesis, the statistical parametric speech synthesis techniques have been developed with the hidden Markov model (HMM)-based approach to be the most commonly used one (Yoshimura et al., 1999; Ling et al., 2006; Black, 2006; Zen et al., 2007). In contrast to unit selection speech synthesis, where actual instances of speech taken from a database are concatenated together to synthesize speech, in statistical parametric speech synthesis, the synthetic speech is produced by the proper manipulation of the parameters of a model offering the advantage of controlling the procedure and adapting the approach to different voices-speakers, languages or applications (Black et al., 2007).

In order to produce high quality synthetic speech, TTS methods rely on databases of high quality recordings, i.e. databases of clean and controlled speech are needed. The recordings have to be noise free (studio quality) and free of artifacts introduced by the speaker, such as breathe sounds, sounds of the lips, etc. Furthermore, the contents of the speech database must be phonetically rich and balanced with controlled prosody, and with utterances targeting at the domain for which the TTS is designed for.

Furthermore, the availability of large speech databases is a prerequisite for the unit selection (Hunt and Black, 1996; Iwahashi et al., 1993) and the HMM-based speech synthesis (Yoshimura et al., 1999; Tokuda et al., 2000) approaches. Since Modern Greek is not a widely-spoken language, to this end only limited efforts have been invested in development of corpus-based speech synthesis as well as the development of speech synthesis resources and tools (Fotinea and Tambouratzis, 2005; Tsiakoulis et al., 2008; Zervas et al., 2008).

The work presented here is an incremental step within a long-term effort towards the creation of large speech databases of Modern Greek language, extending the existence of Modern Greek speech synthesis databases (Fotinea and Tambouratzis, 2005; Tsiakoulis et al., 2008; Zervas et al., 2008). In the following we outline a recently created Modern Greek speech database, referred to as the Vergina speech database, which is intended for research on and development of corpus-based text-to-speech systems for Modern Greek language. Specifically, the present work reports details on the design of the Vergina database, the development and implementation (recordings and corrections applied) as well as the database segmentation and annotation efforts.

The remainder of this paper is organized as follows: in Section 2 we outline the design of the Vergina database. In Section 3 we offer a description of the speech recordings and in Section 4 we outline the annotation procedure. This work concludes with Section 5.

2. Design of the Vergina Database

The most crucial requirement in the design phase of a speech database for speech synthesis is the adequate *phonetic coverage* of the selected text corpus. In corpus-based speech synthesis, the quality of the output is highly correlated with the coverage of the database. In detail, it is necessary to include most of the contextual segmental variants in the database along with as more phonetic transitions as possible (diphones), and thus compensate for the co-articulation phenomenon in speech (Kominek and Black, 2003). A text corpus fulfilling this particular condition is characterized as *phonetically rich* (Black and Taylor, 1997b). In the designing phase of the Vergina speech database, this requirement was attained by utilizing an automatic selection of text data from a large corpus.

Furthermore, the design of the database was guided by the needs of building a Greek corpus-based unit-selection voice operating with phone sized units as well as by the needs of a HMM-based voice. Even though perfect quality open-domain synthesis is not yet possible (Kominek and Black, 2003), an attempt was made not to restrict the database to a specific narrow domain. This was achieved by designing the contents in such a way, so that a number of dissimilar domains are covered in the recordings. To implement this intention, we included in the database texts collected from different domains and sources such as newspapers, periodicals, and literature. For that purpose the prompt sentences were designed through the following steps: (i) selecting a source text corpus to represent the target domains, (ii) analyzing the source text corpus to obtain the unit statistics and finally (iii) selecting appropriate prompt sentences from the source text.

In the first step, a large amount of textual material, approximately 5 million words, was collected from articles in newspapers (approximately 2.2 million words) and periodicals (approximately 1.4 million words) as well as from excerpts from the literature (approximately 1.4 million words). The entire text corpus consists of approximately 280 thousand utterances. In the second step, a subset of utterances was produced, by using a Festvox (Black and Lenzo, 2000) script and a Modern Greek diphone TTS based on the Festival Speech Synthesis framework (Black and Taylor, 1997b). Specifically, this script applies a filter on the entire text corpus, in order to select a subset of sentences of length between 5 and 15 words (Kominek and Black, 2003), which are easily read. This procedure resulted in a subset of approximately 95 thousand utterances (sentences, paragraphs) of an appropriate length, which are easily pronounceable. Finally in the third step, this subset was further processed using the *dataset-select* Festvox procedure (Black and Lenzo, 2000), which is based on a greedy search algorithm and leads to the final subset of sentences. The criterion for selection is the sentences to have the best diphone coverage – with the maximum number of diphones and the maximum occurrences of these diphones. An advantage of the Greek language is that the stress is clearly defined in the text (by the stress symbol) over every stressed vowel (i.e. ά/α, έ/ε, ί/ι etc). Thus stressed vowels are represented with unique phonetic symbols. Accounting for this norm in the Modern Greek language and for the fact that stressed syllables play a very important role in the language, we used distinctive representations for the vowels of the stressed and unstressed syllables (i.e. *A*/*a*, *E*/*e*, *I*/*i* etc).

The above mentioned steps resulted in a set of approximately 3,000 sentences. This set corresponds to approximately 23,500 words – 8,000 unique words – and to approximately 60,000 and 127,000 syllables and phones respectively.

Figure 1 and Tables 1 and 2 show structural information of the Vergina speech database. In particular, in Figure 1 the number of words per sentence is presented. In Table 1 the twenty most frequent words of the database are presented along with the number of their occurrences and the pronunciation of the words. In Table 2 the twenty

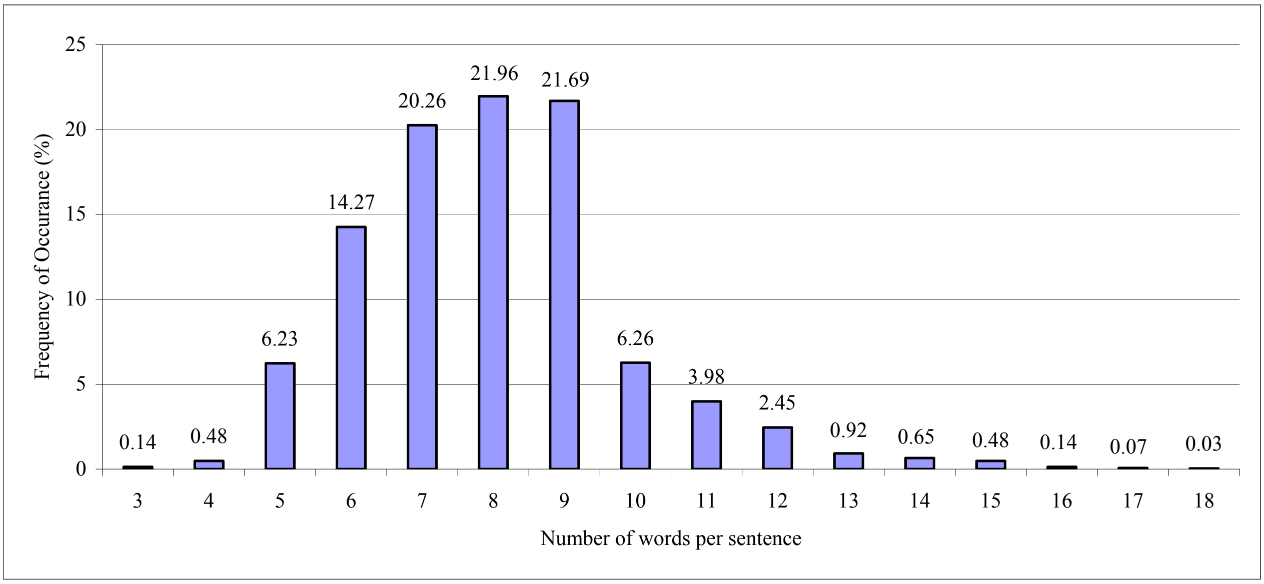

Figure 1: Structural information of the WCL-1 text corpus

Word	Pron.	Num. of Occur.	Word	Pron.	Num. of Occur.
και	Ke	586	*τα*	ta	279
το	to	566	*για*	Ya	258
να	na	545	*οι*	i	249
η	i	502	*θα*	Qa	245
του	tu	470	*με*	me	227
είναι	Ine	428	*από*	apO	225
ο	o	369	*σε*	se	180
την	tin	315	*στο*	sto	179
της	tis	311	*που*	pu	171
δεν	Den	300	*των*	ton	167

Table 1: The twenty most frequently occurred words in the database along with their pronunciation.

Diphone	Num. of Occur.	Diphone	Num. of Occur.
t-i	2575	*a-s*	1232
i-s	2118	*t-a*	1132
s-t	1962	*a-p*	1061
s-i	1658	*o-n*	1035
t-o	1578	*n-e*	1030
n-a	1542	*e-t*	1013
a-t	1354	*r-i*	960
i-n	1348	*t-e*	958
a-n	1288	*i-a*	940
o-s	1269	*m-e*	932

Table 2: The twenty most frequently occurred diphones in the database.

most frequent diphones of the database are presented along with the number of their occurrences.

The phone-set used in the database is a modification of the SAMPA (Wells, 1997) phonetic alphabet for Greek. The phone-set consisted of 39 phones plus the silent (pau) was adopted. These forty phones define eight classes as follows:

- Vowels
  - Stressed Vowels: /A/, /E/, /I/, /O/, /U/,
  - Unstressed Vowels: /a/, /e/, /i/, /o/, /u/,
- Consonants
  - Affricates: /c/, /j/,
  - Fricatives: /D/, /f/, /Q/, /s/, /v/, /x/, /X/, /y/, /Y/, /z/,
  - Liquids: /l/, /L/, /r/,
  - Nasals: /m/, /n/, /N/, /h/,
  - Plosives: /b/, /d/, /g/, /G/, /k/, /K/, /ks/, /p/, /t/, /w/,
  - Silence: /pau/.

The percentage of the diphone coverage for Vergina database is nearly 75%. This percentage is derived based on the consideration that, in theory, the maximum number[1] of the diphones is 1599=40x40-1. However, the real percentage is even higher since the realizable diphones in Greek language are less than 1599. In Table 3 each phone of the Vergina speech database phone-set along with the respective symbol of the IPA alphabet is presented.

3. Recordings

The Vergina database has been recorded in the audio studio located in the premises of the Artificial Intelligence Group at the University of Patras. The studio walls (floating screed) are 12 cm thick filled with glass-wool material. Heavy curtains and carpets are installed on the inside area as absorbent material.

The female voice talent, a native Greek speaker, being recorded was sitting in front of a personal computer with her mouth 10 to 20 cm away from the microphone[2]. A pop filter was installed between the speaker and the microphone to reduce the force of airflow to the microphone. Furthermore, a high fidelity audio capture card[3] was used, and the audio was sampled with sampling rate of 44.1 kHz and a resolution of 16 bit per sample.

The talent used a customized graphical user interface (GUI) to process the utterances. This GUI consists of one box showing the utterance to be captured and four buttons; two buttons for starting and stop recording, a "play" button allowing the talent validating the recording and a "next" button for saving the current recording/proceeding to the next utterance to be recorded.

Due to the large amount of recordings, the database collection campaign had duration of two weeks. Consequently, in order to reduce the unevenness which could result due to the multiple recording sessions, the speaker was instructed to speak in a neutral voice with minimal inflection. In total, the database was recorded in fifteen sessions, each one with length of approximately two hours. The Vergina speech database consists of approximately 3,000 sentences corresponding to approximately four hours of high quality speech.

After the end of the recording campaign, all recordings were checked in order to identify and re-record the misspellings and other mistakes in the database. The

Vergina Phoneset	*IPA Alphabet*	*Vergina Phoneset*	*IPA Alphabet*	*Vergina Phoneset*	*IPA Alphabet*
A	a	*i*	i	*Q*	θ
a	a	*j*	dʒ	*r*	r
b	b	*K*	c	*s*	s
c	tʃ	*k*	k	*t*	t
D	ð	*ks*	ks	*U*	u
d	d	*L*	ʎ	*u*	u
E	e	*l*	l	*v*	v
e	e	*m*	m	*w*	ps
f	f	*N*	ɲ	*X*	ç
G	ɟ	*n*	n	*x*	x
g	g	*O*	o	*Y*	ʝ
h	ŋ	*o*	o	*y*	ɣ
I	i	*p*	p	*z*	z

Table 3: Vergina Speech Database phone set along with the respective symbols of the IPA alphabet.

[1] This product is reduced by one since /pau-pau/ does not count as a valid diphone.

[2] AKG C3000B

mistaken recordings, which were approximately the 10% of the whole database, were recorded again in purposely-planed additional recording session.

4. Annotations

Annotations were semi-automatically created utilizing task-specific tools; based on a hidden Markov model (HMM) segmentation method (Mporas et al., 2008). In detail, we used the word-level prompts of each speech recording file to compute the phonetic-level annotation as well as the positions of the phonetic transitions.

With the use of a pronunciation dictionary we converted the word level transcriptions to the corresponding phone sequences. The pronunciation dictionary consisted of the phonetic representation of all the pronounced words of the prompts, using a set of forty phones of the Greek language. As described in Section 2, the phone-set used for the annotations of Vergina speech database is a modification of the SAMPA (Wells, 1997) phonetic alphabet for Greek. The phonetic representation of each word of the dictionary was made manually by expert phoneticians.

After producing the phonetic transcription for each speech waveform of the database we estimated the phonetic boundary positions by time-aligning the phone sequences with HMM phone models. In order to accurately estimate the phonetic transition positions we used the hybrid-HMM method of Mporas et al. (2008). In this method, for each phone an HMM model is initially constructed by embedded training (Young et al., 2006) of the corresponding HMMs. The resulting initial set of HMM models is time-aligned against the phonetic sequences in order to produce a first estimation of the phonetic boundaries. These boundaries are in turn used to train isolated-unit models (Young et al., 2006), which in turn are time-aligned to produce a refined, i.e. more accurate, estimation of the phonetic boundary positions. This procedure is iteratively repeated until the average refinement of the boundary positions reaches a predefined threshold.

Here, we utilized context-independent HMM phone models. The parameterization of the speech recordings was performed using the Mel frequency cepstral coefficients (Davis and Mermelstein, 1980), within a 20 millisecond sliding Hamming window, with step 5 milliseconds.

Except for the word-level and phone-level segmentation we annotated the database in syllable-level. Furthermore after the HMM-based segmentation, effort for manual inspection and correction of the automatic annotations, concerning the full size of the database, took place for improving the automatic annotation on the phone-level and on the syllable-level and thus improving the quality of the derivative speech voice (synthetic speech). The manual inspection and correction of the automatic annotations of the speech database was made using the PRAAT software (Boersma and Weenink, 2008). The most important criterion for the hand-correction of the boundaries of each phone, and subsequently of each syllable and word was the listening perception of the speech signal, along with the visual observation of it and its spectrum.

In Table 4 the mean durations, the standard deviations and the number of occurrences of all the phones of Vergina speech database after the manual correction of the automatic segmentation are presented.

5. Conclusion

In this work, we outlined the Vergina speech database, which was recently developed at the Wire Communications Laboratory of the University of Patras. The design, development and annotation of the database

Phone	Mean Duration	Standard Deviation	Number of Occurrences	Phone	Mean Duration	Standard Deviation	Number of Occurrences
A	113.35	26.86	2911	*m*	76.39	15.39	4041
a	64.27	21.90	11080	*N*	97.31	25.61	158
b	89.95	24.75	290	*n*	64.21	18.26	7367
c	123.82	29.01	165	*O*	100.40	29.73	3485
D	75.47	20.04	2194	*o*	60.20	21.20	7960
d	89.69	21.20	888	*p*	84.62	23.63	4648
E	99.65	26.69	2914	*Q*	94.57	23.65	1537
e	57.87	18.05	8221	*r*	44.42	12.98	5676
f	92.63	26.82	1715	*s*	84.58	30.10	9120
G	114.72	27.33	64	*t*	76.61	20.77	8955
g	102.15	28.29	323	*U*	95.12	31.45	648
h	59.43	22.58	156	*u*	49.38	21.90	2237
I	91.44	31.80	5474	*v*	82.16	21.31	1090
i	56.14	24.42	12860	*w*	152.16	25.71	257
j	123.99	36.98	129	*X*	118.02	29.82	680
K	103.73	29.33	2002	*x*	101.15	22.38	1014
k	86.47	22.43	2698	*Y*	98.65	29.07	958
ks	153.26	26.34	618	*y*	70.45	17.93	1042
L	95.38	30.37	160	*z*	82.73	20.71	893
l	77.08	16.70	3184	*pau*	270.09	188.43	16673

Table 4: Mean duration, standard deviation and number of occurrences of the phones of Vergina speech database.

were described. In summary, the Vergina database, recorded in audio studio, consists of approximately 3,000 phonetically balanced utterances in Modern Greek language. The Vergina speech database was annotated using HMM-based speech segmentation tools, and then manual corrections were introduced to improve the annotation. This database was created in support of speech synthesis research, for the needs of development of corpus-based unit selection and HMM-based speech synthesis systems for Modern Greek language. The broad coverage and contents of the recordings in the database (text corpus collected from different domains and writing styles such as newspapers, periodicals, and literature) makes this database appropriate for various application domains. Eventually, the speech corpus will be made available for research purposes (AIG, 2010).

6. References

AIG (2010). Artificial Intelligence Group, Wire Communications Laboratory, Electrical and Computer Engineering Department, University of Patras, Greece. http://www.wcl.ece.upatras.gr/ai/Vergina

Black, A.W. (2006). CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. In *Proc. of Interspeech'06*. Pittsburgh, Pennsylvania, USA, pp. 1762--1765.

Black, A.W., Lenzo, K. (2000). Building voices in the Festival speech synthesis system. http://festvox.org/bsv/.

Black, A.W., Taylor, P. (1997a). Automatically clustering similar units for unit selection in speech synthesis. In *Proceedings of Eurospeech'97*. Rhodes, Greece, pp. 601--604.

Black, A.W., Taylor, P. (1997b). The Festival Speech Synthesis System: system documentation. Technical Report HCRC/TR-83. Human Communications Research Centre, University of Edinburgh, Scotland, UK.

Black, A.W., Zen, H., Tokuda, K. (2007). Statistical parametric speech synthesis. In *Proc. of ICASSP'07*. Hawaii, pp. 1229--1232.

Boersma, P., Weenink, D. (2008). Praat: doing phonetics by computer (Version 5.0.32).

Campbell, N., Black, A.W. (1996). Prosody and the selection of source units for concatenative synthesis. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (eds). Progress in speech synthesis, pp. 279--282. Springer Verlag.

Davis, S.B., Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4), pp. 357--366.

Dutoit, T. (1997). An Introduction to Text-To-Speech Synthesis. Kluwer Academic Publishers, Dordrecht.

Fotinea, S.L., Tambouratzis, G. (2005). A Methodology for Creating a Segment Inventory for Greek Time Domain Speech Synthesis. *International Journal of Speech Technology,* 8, pp. 161--172.

Hunt, A., Black, A.W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database In *Proc. of ICASSP'96*. Atlanta, Georgia, pp. 373--376.

Iwahashi, N., Kaiki, N., Sagisaka, Y. (1993). Speech segment selection for concatenative synthesis based on spectral distortion minimization, *IEICE Transaction Fundamentals*, E76-A, pp. 1942--1948.

Klatt, D.H.. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*. 82(3), pp. 737--793.

Kominek, J., Black, A.W. (2003). CMU ARCTIC databases for speech synthesis, CMU-LTI-03-177. Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Ling, Z.H., Wang, R.H. (2006). HMM-based unit selection using frame sized speech segments. In *Proc. of Interspeech'06*. Pittsburgh, Pennsylvania, USA, pp. 2034--2037.

Möbius, B. (2000). Corpus-based speech synthesis: Methods and Challenges. Technical Report, University of Stuttgart, AIMS 6 (4).

Mporas, I., Ganchev, T., Fakotakis, N. (2008). A hybrid architecture for automatic segmentation of speech waveforms, In *Proc. of ICASSP'08*. Las Vegas, Nevada, USA, pp. 4457--4460.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. of ICASSP'00*. Istanbul, Turkey, pp. 1315--1318.

Tsiakoulis, P., Chalamandaris, A., Karabetsos, S., Raptis, S. (2008). A statistical method for database reduction for embedded unit selection speech synthesis. In *Proc. of ICASSP'08*. Las Vegas, Nevada, USA, pp. 4601--4604.

Wells, J.C. (1997). SAMPA computer readable phonetic alphabet. In D., Gibbon, R., Moore, and R., Winski (eds.). Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. of Eurospeech*'99. Budapest, Hungary, pp. 2347--2350.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. (2006). The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department.

Zen, H., Toda, T., Nakamura, M., Tokuda, T. (2007). Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. Syst*. E90-D (1), pp. 325--333.

Zervas, P., Fakotakis, N., Kokkinakis, G. (2008). Development and evaluation of a prosodic database for Greek speech synthesis and research. *International Journal of Quantitative Linguistics*, 15(2), pp. 154--184.