EUSIPCO 2013 1569746469

ADVANCES IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION IN GREEK: MODELING AND NONLINEAR FEATURES

Isidoros Rodomagoulakis[1,3]*, Gerasimos Potamianos*[2,3]*, and Petros Maragos*[1,3]

[1]School of ECE, National Technical University of Athens, 15773 Athens, Greece
[2]Department of CCE, University of Thessaly, 38221 Volos, Greece
[3]Athena Research and Innovation Center, 15125 Maroussi, Greece

irodoma@cs.ntua.gr, gpotam@ieee.org, maragos@cs.ntua.gr

ABSTRACT

The main goal of this work is the development of an improved Large Vocabulary Continuous Speech Recognition (LVCSR) framework in Greek. Language modeling is carried out in a collection of journalistic text and in the acoustic signal processing, a nonlinear approach is implemented for deriving features of the AM-FM type. Experimentation is carried out in both clean and simulated far-field speech offering insight about the acoustic modeling under adverse conditions with reverberation and additive ambient noise. Beyond the baseline implementation, a first step is made in exploring how standard (MFCCs and PLPs) and modulation features (AM-FM) behave in a LVCSR framework when the input speech is distant, like in real life home applications.

Index Terms— Speech Processing, Acoustic modeling, Language modeling, Large Vocabulary Continuous Speech Recognition

1. INTRODUCTION

One of the main difficulties in Greek Automatic Speech Recognition (ASR) is the complex nature of the language due to multiple inflectional rules. Thus, many efforts have been made in language processing and modeling, but only a few works report extensive results in Large Vocabulary Continuous Speech Recognition (LVCSR) problems which involve many other issues regarding feature extraction, acoustic modeling and recognition methods. Among these works is the implementation of a dictation system [3] which achieved 19.27% Word Error Rate (WER) by using speaker-independent genomic Hidden Markov Models (HMMs) [2]. A WER reduction of 0.28% was obtained on the same database by using maximum entropy language models [9] that employ stem information to cope with the very large number of distinct words. Another LVCSR module has been implemented in [6] for a Greek Broadcast transcription system where the reported WER for speaker-independent recognition in mixed recording conditions was 38.42%. On Greek phoneme recognition, experiments have been conducted on the SpeechDat(II)-FDB-5000 Greek database [1], yielding 39.06% classification accuracy. Finally, some efforts have also been made for ASR in home environments. An implemented system [8] with a low-cost microphone recognized 3.2k Greek words and instructions in spontaneous speech, using a task-dependent grammar yielding 5.25% WER in recordings with 12dB SNR. The acoustic modeling was based on the SpeechDat(II)-FDB-5000 corpus. Overall, a possible limitation of these works is the employment of standard front-end methods for extracting the traditional cepstral features that are applied only for small and medium vocabulary tasks.

Our motivation for this work lies basically in the fact that Greek ASR lacks extensive experimentation in large vocabulary databases. Additionally the front-end is mostly based on the linear speech model with the traditional cepstral features without attempting to extend or to combine with other nonlinear features that overcome some assumptions of the linear speech production. In addition, AM–FM features and their variants have been proved effective for other recognition problems in other languages (Spanish, English). Although some efforts have been made to improve the acoustic and the language modeling for Greek, only a few works have completely integrated all the components in a ASR framework for LVCSR experimentation. The next sections describe all the components of the implemented recognizer and the obtained results in a large vocabulary Greek speech corpus. Language modeling with n-grams is described in Sec. 2, while in Sec. 3, the extraction of standard Mel-Frequency Cepstrum Coefficients (MFCCs), Perceptual Linear Prediction (PLP) coefficients and nonlinear AM-FM based features for speech is analysed. Section 4 involves the acoustic modeling with context independent triphones and Sec. 5 describes the conducted experiments and the obtained results for clean speech and far-field simulations. Finally, Sec. 6 concludes the paper.

2. LANGUAGE MODELING FOR GREEK ASR

This section describes the development of back-off n-gram language models for Greek in the field of journalism. The

This research was supported by the European Union under the research program DIRHA with grant FP7-ICT-2011-7-288121.

specific field is selected for two reasons: (a) Journalistic text is considered as formal representative of written language and (b) the speech corpus is recorded by journalists for a variety of news including politics, athletics, entertainment etc. The effectiveness of the implemented models is measured in terms of perplexity, out of vocabulary rate (OOV) and WER.

2.1. Language characteristics

Greek is an inflectional and morphologically rich language. Verbs are varied by the usage of eight tenses and six persons while articles, adjectives and nouns must agree with the subject in number, gender and person. All these forms of variation produce a very big vocabulary size. To compare with English, the Wall Street Journal corpus consists of 37.2M words from which only 165K are unique and thus, the corresponding sizes of Table 1 prove that Greek is much more complex. Stem-based approaches [9] have been used to cope with the large vocabulary issues but the reported WER improvement in recognition was minor. In this work, we developed n-grams to model word sequences.

2.2. N-gram models and lexical content

Bigram and tri-gram models are built in a collection of Greek journalistic text from various sources, containing 12.2M words. As Table 1 depicts, the sources of text are transcriptions of the "Logotypographia" [3] corpus in which LVCSR experiments are conducted (set L), transcriptions of a broadcast news corpus named "GridNews" [6], and text from the web site of a reputable Greek newspaper named "Eleftherotypia". The L set is split into sets L_t and L_e with $|L_t| = 30000$ and $|L_e| = 1000$ sentences. The latter corresponds to the evaluation set for both language modeling and recognition. Set L_t consists of the remaining sentences of L that do not overlap with L_e.

Language modeling is implemented with the Carnegie Mellon University language modeling toolkit [11] which supports Good-Turing discounting for better modeling of low-frequency word sequences. The presented models in Table 2 are trained either on the whole text (LEG) or on subsets where the text of set L does not participate at all (EG) or partially (L_tEG) with sentences from the training set L_t only. Another varying factor is the vocabulary size. This consists either of 4.8k words (i.e., the unique words of test set L_e), or of 37k words (i.e., the unique words of the entire set L). In this way we are able to explore how the size and the type of training text affect the perplexity of the models and also the recognition performance. Moreover, we compare models of the closed-vocabulary type against more general ones having larger vocabularies than in the recognition set.

The Perplexity (PP) and WER results of Table 2 confirm the expected fact that the LEG-37k set is overtrained because $L_e \subset L$. Comparing L_tEG-37k with L_tEG-4.8k, perplexity and WER degrade relatively almost 20% and 31% respectively. Finally, it is observed that the undertrained EG-4.8 set acts in recognition almost in the same way with L_tEG-4.8k although the perplexity is significantly increased (e.g., 598.00 vs. 178.00). All the models above have zero OOV rates. In addition, we implement a more general model set LEG-242k trained and tested on the 20% and 80% of the LEG text respectively for all the 242k unique words that appear in the text. The perplexity for tri-grams was measured at 192.0 and the OOV was 8%. The achieved performance is near the state-of-the-art for Greek language modeling as it is reported for texts of the same content and size [9].

Notice that for the LVCSR experiments presented next, only tri-grams are incorporated in decoding due to their superior perplexity. In particular, L_tEG-37k was chosen due to its larger vocabulary size. Further, and in order to allow system training and decoding, a Greek dictionary with multiple pronunciations for 411k words is used to map the "Logotypographia" vocabulary into phoneme sequences.

text resources	# words	voc. size
(L) "Logotypographia" corpus	550 k	37 k
(E) "Eleftherotypia" web-site	11 M	233 k
(G) "GridNews" corpus	670 k	47 k
total	12.2 M	242 k

Table 1. The text resources employed in Greek language modeling with their approximate text and vocabulary sizes.

model	PP bigrams	PP tri-grams	WER
LEG-37k	143.40	30.10	3.34
L_tEG-4.8k	252.18	178.90	10.81
EG-4.8k	1173.00	598.00	11.49
L_tEG-37k	321.20	212.35	14.16

Table 2. Performance of the four developed Greek bigram and tri-gram language models in terms of perplexity (PP) and WER (for tri-grams) measured on the `eval` set (L_e) of the "Logotypographia" corpus which is used for Greek LVCSR.

3. ACOUSTIC FEATURES

3.1. Cepstral & Perceptual Linear Prediction coefficients

The built-in HTK front-end [12] is used to derive 12 MFCCs and 12 PLP coefficients from $N = 20$ filterbank channels that span the whole signal bandwidth. Short-time analysis is applied every 10 ms over 32 ms long speech frames that are first Hamming filtered and pre-emphasized. The C_0 coefficient is also added for better description of cepstral energies. To capture speech dynamics, delta and acceleration coefficients are also computed in 5-sample windows using the regression formula that is implemented in HTK. Overall, 39-dimensional feature vectors are produced per frame for each feature type.

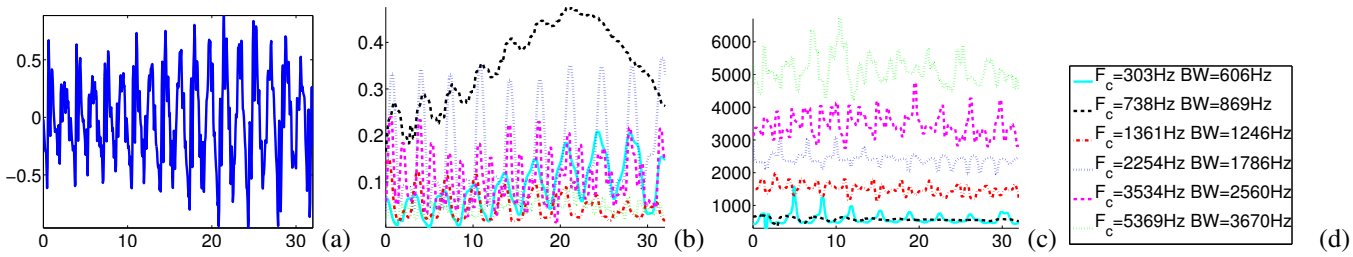

Fig. 1. Short-time non-linear analysis of (a) 32 ms of speech of phoneme /a/ by applying Gabor-ESA with six Mel-spaced, 50% overlapped Gabor filters. (b) Instantaneous amplitudes and (c) Instantaneous Frequencies (time axis in milliseconds) (d) Quantities $\mathrm{F_c}$ and BW denote the central frequencies and bandwidths of the Gabor filterbank, both in Hz.

Finally, Cepstral Mean Substraction (CMS) normalization is applied to compensate for long-term spectral effects, such as those caused by different microphones and audio channels.

3.2. Nonlinear speech processing

Based on the evidence that speech resonances can be modeled with an AM–FM signal

$$r_i(t) \;=\; a_i(t)\cos\Big(\int_0^t \omega_i(\tau)\,d\tau\Big)\;, \qquad (1)$$

a variety of modulation features have been proposed for ASR [5] that are mainly extracted from the first- and second-order statistics of the instantaneous amplitudes $a_i(t)$ and angular frequencies $\omega_i(t)$, as computed by applying one of the Energy Separation Algorithms (ESA) [4] for signal demodulation. In this work, second-order modulation features are described and used for ASR experiments namely (a) Mean Instantaneous Amplitudes (MIA) and (b) Mean Instantanesous Frequencies (MIF) of the bandpasssed speech signals. This set of features provide information about the variation of each speech resonance signal in the time and frequency domain. MIA and MIF features have been successfully applied to a variety of speech and music applications, such as noisy speech detection [7], phoneme classification [4], and music instrument classification [13]. Also, their combination with standard MFCCs led to recognition improvements for a medium-size vocabulary ASR task as reported in [5] .

3.3. MIA and MIF extraction

MIAs and MIFs are the short-time means of the normalized instantaneous amplitude and frequency[1] signals $a_i[n]$, $f_i[n] = \omega_i[n]/2\pi,\ i \in [1, K]$, where i is the filter index. The instantaneous frequency signal values are trimmed within $\pm\,5\%$ of their mean value, so that isolated spikes are ignored. The same short-time analysis window length and step are used as for MFCCs to extract MIAs and MIFs in three steps:

1. The AM–FM composed speech $s(t) = \sum_{i=1}^{K} r_i(t)$ is convoluted with a Gabor filterbank $\{g_i(t)\}$, $i \in [1, K]$. The filterbank is Mel-spaced, spanning the interval $[\,0\,,\ f_s/2]$ Hz; also, K filters, in total, are used with a frequency overlap of 50%.

2. Each signal component $s_i(t)$ is demodulated to its instantaneous AM–FM components $a_i[n]$, $f_i[n]$ using Gabor-ESA [4].

3. The $a_i[n]$, $f_i[n]$ are averaged for the N samples of the analysis frame. MIAs and MIFs are derived as $(1/N)\times \sum_n \log(a_i[n])$ and $(1/N)\times\sum_n f_i[n]$, respectively, and are concatenated to form feature vectors that are augmented with their first- and second-order derivatives.

The above processing is applied to the 95% peak normalized speech signals. Motivated by the non-linear human perception of speech, MIAs are transformed using a logarithm. MIFs are only scaled from the frequency domain to the [0,1] range, by dividing with $f_s/2$. Then, both features are mean and variance normalized to cope with long-term effects. Standardization is applied per utterance, across filters for MIAs in order to keep the relative information that exists between the coefficients, and per filter for MIFs. Figure 1 depicts the non-linear raw features for a frame. Smoothness of the corresponding curves is achieved by computing signal derivatives as convolutions and not as sample differences [4]. As it can be observed, the existence of fine structure of speech formants lies in the fact that their instantaneous frequencies vary in reference with each filter's central frequency. The variation becomes more unstable in the higher frequencies due to the usage of only $K = 6$ mel-spaced filters that span all the frequency bins. Thus, in the next experiments, we also tried to apply a Gabor filtebank with $K = 12$ filters. Overall, two feature sets are produced for experimentation, MIAF-6 and MIAF-12 corresponding to the concatenated MIA and MIF vectors that are extracted for $K = 6,\ 12$ filters. These sets are augmented with the signal's squared-amplitude energy. Finally, first- and second-order derivatives are added to form 39- and 75-dimensional vectors for each set respectively.

[1] Instantaneous frequencies $f_i,\ i \in [1, K]$ are measured in Hz.

4. ACOUSTIC MODELING

Context dependent triphone HMMs are trained with approximately 22.6 hours of speech. The developed HMMs consist of three states with continuous density GMM probabilities having diagonal covariance matrices. Due to the absence of time labels in the "Logotypographia" transcriptions, a set of three-state monophone models is first built using "flat-start" initialization. Then, training of monophones, triphones, state tying, and Gaussian mixture splitting is performed, resulting to a set of tied-state, context-dependent triphones with three states and 16 Gaussians per state. For the state tying process, a decision tree clustering strategy is applied utilizing 82 hand-crafted phonetic questions adequate for the set of 28 phonemes considered for Greek LVCSR. Models for silence and noise are also trained, the former employing short segments at the beginning and end of each utterance, and the latter using data transcribed as "breath", "clear throat", "paff noise", "paper rustle", and "phone ring". Four sets of acoustic models are produced corresponding to the extracted feature sets which are namely (a) the MFCCs, (b) the PLPs, (c) the MIAF-6, and (d) the MIAF-12 as described above.

5. LVCSR EXPERIMENTS AND RESULTS

5.1. Clean speech and far-field simulations

Recognition experiments are based on the "Logotypografia" database which has been developed in the Greek journalism domain for speaker independent, large vocabulary dictation systems and is publicly available for experimentation and benchmarking. Specifically, we used a subset of 27 hours of close-talk recordings and split it to three sets: (a) 22.6 hours of training data (`train`) b) 1.2 hours of development (`devel`) data and (c) 2.3 hours of testing data (`eval`). Each set is speaker independent consisting of 54, 15, and 15 non-overlaping, gender balanced speakers respectively. The close-talk recordings took place in two different acoustic environments: a) a sound proof room and b) a quiet environment with an AKG C410 head-mounted microphone providing SNR levels that distribute in the range of 12-40 dB. For further details on the database see [3].

Simulations of distant-speech in a home environment are also considered for experimentation. It is assumed that the simulated signal $\hat{s}(t)$ is acquired by convolving the source signal $s(t)$ with a room impulse response $r(t)$ that corresponds to a specific acoustic path from the source to a distant microphone. Further details about the estimation of impulse responses can be found in [10]. The simulated signal is also contaminated with additive ambient noise $v(t)$. The simulation formula $\hat{s}(t) = s(t) * r(t) + v(t)$ was applied in each set of the clean recordings as described above. Two sets were produced for experimentation, `reverb1` and `reverbR`. In `reverb1`, conditions are constant, i.e source in position LA

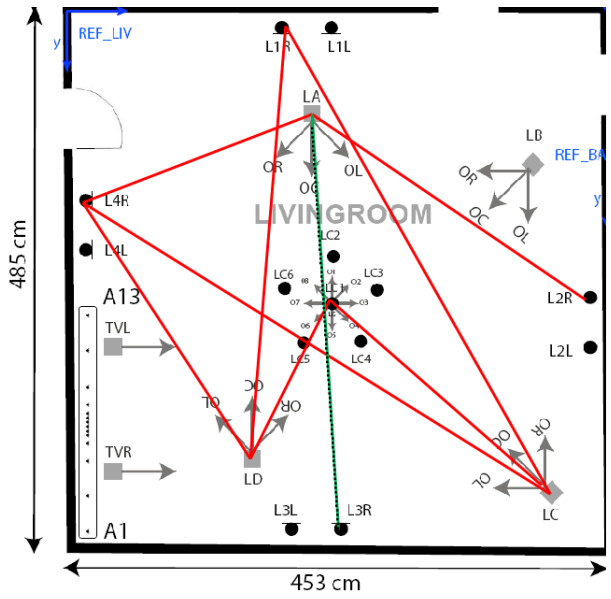

Fig. 2. Simulation map of an apartment living room. Green (dashed) and red (solid) lines correspond to the source-microphone location pairs which were simulated in the `reverb1` and `reverbR` sets, respectively.

(see the map of Fig. 2), microphone LR3 and additive ambient noise with gain 3. In the more challeging `reverbR`, conditions are randomly changed by applying 10 source-microphone impulse responses (LA-L3R, LA-LC1, LA-L4R, LA-L2R, LC-LC1, LC-L4R, LC-L1R, LD-LC1, LD-L1R, LD-L4R) combined with 3 noise levels (gains 3, 6, 9). With these two sets, we can test the ability of the ASR system to recognize distant speech from multiple speakers in multiple positions inside the simulation room as it is depicted in Fig. 2.

5.2. Recognition results

To evaluate the developed Greek LVCSR system, recognition experiments are conducted in matched and mismatched conditions for the `clean`, `reverb1`, and `reverbR` sets, in a speaker independent framework. The decoding parameters such as the word insertion penalty, the weights for the acoustic and language models, and the pruning threshold are optimized on the `devel` set. All decoding experiments are performed using the HTK LVCSR decoder that supports tri-gram language models [12]. The decoding vocabulary contains all the 37k unique words of the corpus.

Speaker independent recognition results are reported in terms of WER, %, in Table 3. Overall, the performance under matched conditions is considered quite satisfactory, especially given the large vocabulary size and the challenging conditions of the simulation sets. The achieved 14.16% with MFCCs on the original set of the "Logotypographia" database is comparable with the 19.27% that is reported in [3] for a larger set of the same database. As expected, there

training conditions	testing conditions								
	clean			reverb1			reverbR		
	MFCCs	PLPs	MIAF-6	MFCCs	PLPs	MIAF-6	MFCCs	PLPs	MIAF-6
clean	**14.16**	**14.24**	**19.84**	89.18	89.22	93.62	90.28	90.05	93.62
reverb1	96.49	95.42	98.02	**33.19**	**32.38**	**38.97**	41.64	44.09	50.68
reverbR	95.16	94.68	97.12	41.92	43.32	51.07	**42.59**	**43.22**	**49.54**

Table 3. WER, %, of the Greek LVCSR system on the eval set in matched (in bold) and mismatched train/test conditions.

is a performance degradation in the distant speech data that is more pronounced in the more challenging reverbR scenario. Moreover, when acoustic models are trained and tested in mismatched conditions, the WER increases significantly compared to the matched condition results. Such degradation is less prominent between the two noisy conditions, compared to the degradation between the noisy and clean conditions. Comparing the three employed feature sets, MFCCs performed on average slightly better in matched and mismatched conditions. It is worth noting that the MIAF-12 set achieved 21.45%, 38.03% and 48.32% for matched conditions which is slightly better than MIAF-6 in noise but degraded in clean speech. On average, MIAF-12 has the same performance but adds complexity due to the 75-dimensional feature vectors.

The performance of MIAF features is degraded in all conditions, especially for the distant-speech sets but this performance can be explained by the fact that the proposed modulation features were originally conceived for ASR in [5] as capturing the second-order non-linear structure of speech formants, whereas the linear speech model and its corresponding features (e.g., MFCCs) capture the first-order linear structure of speech. This leads to considering an ASR system that provides fusion of linear and non-linear features. A first attempt to this direction has been made in clean data and results showed that the combination of MFCCs with MIAF-6 in a multi-stream HMM framework with stream weight optimization, achieved a 2% relative reduction in WER.

6. CONCLUSIONS

We have designed and developed the components of a Greek Large Vocabulary Continuous Speech Recognition framework. The obtained improvements in language modeling in terms of perplexity and WER, and the incorporation of standard as well as nonlinear acoustic features led to reasonable recognition performance for the challenging large vocabulary database in which we experimented. Moreover, a second contribution was the extraction and the incorporation of second-order modulation features which may potentially lead to improvements when fused with first-order features stemming from the linear model of speech. Finally, beyond the baseline implementation, a first step has been done in exploring how standard (MFCCs and PLPs) and modulation features behave in a LVCSR framework when the input speech is distant like in real life home applications.

Acknowledgment

The authors wish to thank M. Omologo, P. Svaizer, and L. Cristoforetti of Fondazione Bruno Kessler Italy, for providing the simulated data for distant speech recognition and Figure 2.

7. REFERENCES

[1] I. Chatzi, N. Fakotakis, and G. Kokkinakis, "Greek speech database for creation of voice driven teleservices," in *Proc. Eurospeech*, 1997.

[2] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers," *IEEE Transactions on Speech and Audio Processing*, pp. 281–289, Jul. 1996.

[3] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, and V. Diakoloukas, "Large vocabulary continuous speech recognition in Greek: Corpus and an automatic dictation system," in *Proc. Interspeech*, 2003.

[4] D. Dimitriadis and P. Maragos, "Continuous energy demodulation methods and application to speech analysis," *Speech Communication*, vol. 48, no. 7, pp. 819–837, 2006.

[5] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters*, 2005.

[6] D. Dimitriadis, A. Metallinou, I. Konstantinou, G. Goumas, P. Maragos, and N. Koziris, "GRIDNEWS: A distributed automatic Greek broadcast transcription system," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[7] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2024–2038, 2006.

[8] I. Mporas, T. Ganchev, T. Kostoulas, K. Kermanidis, and N. Fakotakis, "Automatic speech recognition system for home appliances control," in *Proc. 13th Panhellenic Conference on Informatics (PCI)*, 2009, pp. 114–117.

[9] D. Oikonomidis and V. Digalakis, "Stem-based maximum entropy language models for inflectional languages," in *Proc. Eurospeech*, 2003.

[10] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, "Impulse response estimation for robust speech recognition in a reverberant environment," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2012.

[11] R. Rosenfield and P. Clarkson, "Statistical language modeling using the CMU-Cambridge toolkit," in *Proc. Eurospeech*, 1997.

[12] S. Young *et al.*, *HTK – Hidden Markov Model Toolkit*, Manual, 2006. [Online]. Available: http://htk.eng.cam.ac.uk/

[13] A. Zlatintsi and P. Maragos, "AM-FM modulation features for music instrument signal analysis and recognition," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2012.