



Template-based Automatic Speech Recognition meets Prosody

Dino Seppi, Kris Demuynck, and Dirk Van Compernelle

ESAT, Katholieke Universiteit Leuven, Belgium

dino.seppi@esat.kuleuven.be

Abstract

In this paper, we use prosodic information to improve the accuracy of our template-based automatic speech recognizer. Prosodic information is harvested adopting a data-driven approach. A number of prosodic features is extracted, then combined into major groups, and finally studied separately and together. All acoustic evidence, both segmental and suprasegmental, is modelled non-parametrically. The different sources of information are conveniently combined with segmental conditional random fields. Prosody enhances the accuracy of the state-of-the-art baseline by reducing the word error rate by 7% relative on the *nov92*, 20k trigram, Wall Street Journal task.

Index Terms: ASR, template based, example based, DTW, prosody, segmental conditional random fields.

1. Introduction

A recent trend in automatic speech recognition, ASR, is the shift of the main research focus from the phonemic content to phenomena that do not apparently distinguish words, but reflect the wider phonological and grammatical structure of the message. One such phenomenon, prosody, carries many different aspects of meaning that are crucial to sustain a normal conversation between humans. In the past, these aspects were deemed irrelevant or even misleading for ASR and were therefore normalized away. Hence, short time spectral information, such as Mel Frequency Cepstral coefficients became the *de facto* standard in state-of-the-art front ends. On the contrary, prosody has been successfully exploited for the detection and classification of paralinguistic events, such as emphasis, emotion, gender, etc.

There are probably two main reasons why the interest in prosody for ASR has revived only in the last few years. First, prosody is difficult to model. Even if closely connected with the phonological structure, not all prosodic events are perceptually salient in all conditions; in other words prosodic attributes show a high intrinsic variability. This may also explain the scarcity and heterogeneity of labeled data. Second, prosodic information is difficult to model in Hidden Markov Model (HMM) based ASR which is currently the dominant approach. HMMs model short intervals of speech, typically up to several tenths of milliseconds. On the contrary, prosodic attributes span many frames, up to several seconds of speech. It is therefore not straightforward to integrate acoustic-phonetic features and prosodic attributes within the same statistical framework. A recent and effective way to exploit suprasegmental information (e.g. word duration [1]) for large vocabulary tasks has been re-ranking word hypotheses found in word graphs (WG). Other approaches essentially condition the language and

the acoustic model on prosodic evidence, thereby constraining recognition [2]; yet other approaches model prosody separately to enhance ASR pre- or post-processing [3].

Template-based (or example-based) Automatic Speech Recognition (T-ASR) [4, 5, 6] has been proposed as a model for human speech recognition based on the assumption that humans represent speech, at least partially, by individual memory traces. The new possibilities offered by template based approaches compared to HMM techniques, combined with the hope of attaining human-like speech recognition accuracy, have led to an increased interest in T-ASR in recent years. As demonstrated in [7, 4], competitive T-ASRs are also very useful in combination with HMM-based ASR as the two framework weaknesses (and errors) are to a large extent uncorrelated. The T-ASR framework implemented at ESAT [8] has been embracing these new trend and challenging concepts.

One of the goals of our work is to narrow the gap between the present T-ASR and the most competitive HMM paradigm. An effective way to improve episodic models and incorporate them in hybrid systems relies on the top-down paradigm, i.e. word graph rescoring with additional, complementary sources of information [7]. In this paper we pursue a similar path and focus on exploiting *prosodic* information. The information contained in the suprasegmental structure of speech fits well in the template framework as i) it can be extracted and stored at the suprasegmental level and ii) it does not need to be modeled explicitly. However, since the number of relevant features that we derive is relatively large, an efficient scheme is needed to rank and combine these features. Segmental Conditional Random Fields (S-CRF) [9] proved to be an effective and efficient way to achieve such an optimization.

The main aim of this paper is thus to verify whether this combination, T-ASR, S-CRF, and prosody, is as promising as it sounds and if a significant improvement over a good baseline can be obtained. Prosodic information, its extraction and processing, is presented in detail in Sec. 2. In Sec. 3 we describe the baseline T-ASR and how prosodic information is embedded into that framework. In Sec. 4 we present the experiments carried out to date, and the relative results. We conclude in Sec. 5 with a discussion on the present results and on future work.

2. Prosodic features

Acoustic features describe the segmental information of speech, which is also the primary source for word decoding: words correspond to phone sequences and phones correspond to articulatory postures and gestures. Speech production is at the same time characterized by the signal *amplitude* and the *mode of excitation*. These phenomena generally have a stationary time evolution much longer than the spectral configuration, and are thus referred to as suprasegmental. Prosody is reflected in these types of acoustic suprasegmental quantities and has percepti-

This work has been supported by the European Community under grant RTN-CT-2006-035561 (Sound-to-Sense) and by the Fund for Scientific Research Flanders, FWO, under grant G.0260.07 (TELEX).

ble correlates called (basic) *prosodic attributes*: pitch, loudness, voice quality, duration, pausing, and speaking rate [10]. They can be directly obtained from acoustic features or from their combination, e.g. the perceived loudness of a complex sound might depend on the sound energy, the spectral structure, and its duration.

The prosodic attributes considered in this paper have been applied to speech analytics tasks, such as emotion recognition [11], automatic punctuation [12], or speaker recognition [13], to name just a few. We adopted a systematic data-driven approach by employing the openSMILE toolkit [14] which allowed the following stages of computation: first, acoustic time-varying features (or acoustic contours) were obtained by a sliding window analysis; second, contours were non-linearly filtered by a mean filter; finally, time normalization (where necessary) was achieved by the application of two types of functionals, namely first order statistics (means and standard deviations) and extreme statistics (minima and maxima).

Although prosodic *features* are modelled directly in the T-ASR framework (cf. Sec.3), in the next sections of this paper the analysis and the results are presented by prosodic *attributes*; note that this is of course not the unique way towards a division into classes. Below we draw a more systematic description of each type of attribute and how it is obtained from the prosodic—and in few cases also linguistic—features. For more details about the feature implementation, cf. [10]. The order of the list reflects, approximatively, the complexity of the implementation and the computational demand required for each attribute.

- **Duration**, as obtained from the alignment, models the temporal aspects of phonological units. Therefore duration is usually correlated with the linguistic content of an utterance. For instance, function words are shorter on average, content words are longer. This information can be used, e.g., to discriminate these two classes.
- In this study we model the **speaking rate** *implicitly* by two features: duration and the count of phones per word, the latter being derived from the word dictionary. It is well known that the rate of speech varies considerably across speakers: this phenomenon can be exploited to give more weight to templates coming from homogeneous clusters of speakers. Speaking rate and duration are the only two *scalar* prosodic attributes. As such they do not need time normalization.
- We derived the **loudness** contour $L(t)$ from the normalized short-time energy (intensity, $I(t)$) obtained using a sliding Hamming window. As often done, we modelled the loudness as perceived by human listeners as $L(t) \propto I(t)^{0.3}$. Energy is the only prosodic feature already considered in the baseline front-end through the MIDA features (cf. Sec.3.1).
- The extraction of the acoustic correlate of **pitch**, F0, is usually error-prone: the sensitivity to strong first formants, especially when they coincide with the second or third harmonic, is one of the big problems in pitch determination [15]. Pitch detection based on the Cepstrum avoids this issue by spectral flattening. For the detection of voiced/unvoiced segments the zero-crossing rate of the autocorrelation function is used. Finally pitch values are filtered by moving average smoothing. An additional feature always used in combination with pitch is a flag indicating whether the current phoneme is a voiced one (or not).

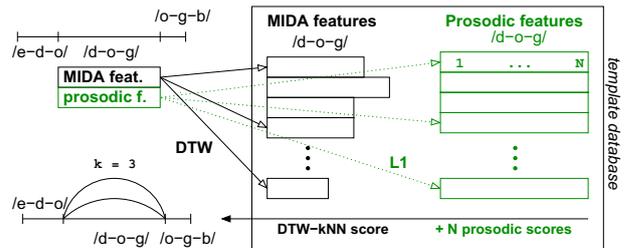


Figure 1: *Rescoring of CD-phone hypothesis /d-o-g/.* The $k = 3$ realizations of /d-o-g/ with smallest euclidean distance from the current hypothesis are selected among those present in the template database. The k dynamic time warping euclidean distances become the new acoustic scores. In green/dotted prosodic information is added: N prosodic scores, one for each entry in the k -NN list, are calculated using Eq.1 and appended to each DTW score.

- The **voice quality** features comprise jitter (the relative cycle-to-cycle variation of fundamental frequency), a flag identifying voiced sounds, the voicing probability (cf. pitch), and the signal zero-crossing rate. They depend in part on other contours such as pitch (jitter) and reflect peculiar voice quality properties such as breathiness, creakiness, or harshness. Therefore they can be used to detect atypical templates in the database.

Eventually, two groups of functionals are applied on a unit basis: first order statistics (mean and standard deviation) and extreme order statistics (minima and maxima). The total dimension of the prosodic vector per template is therefore $N = 3$ (duration, phone count, voiced phone flag) + 5 (loudness, pitch, jitter, zero-crossing rate, voicing probability) \times 4 functionals = 23.

3. Template-based speech recognizer

The T-ASR is a two-pass speech recognition system. In the first step, an HMM-based system generates word graphs (WGs) enriched with phone segmentations. In the second pass, word acoustic scores are obtained by searching in the template database for the best-matching examples and by using the respective euclidean distance, dynamic time warping (henceforth, DTW) scores. At this stage prosodic features can be extracted and prosodic scores can be added to DTW scores.

3.1. Baseline

The conventional HMM system used to create the WGs was obtained using the SPRAAK toolkit [16]. The front-end generates Mel Spectra features post-processed by mutual information discriminant analysis (MIDA) and implements vocal tract length normalization and mean subtraction. The training material for the HMM models used to generate the WGs is the Wall Street Journal (WSJ) continuous speech recognition training corpus. It consists of 81 hours of speech from 284 speakers. The phonetic transcriptions for the training data and the test lexica are drawn from CMUdict 0.6d.

The template database was obtained by aligning the WSJ training data using triphone units. This yielded 2.8 million templates. In total we count 4225 template (or triphone) classes populated by 256 up to 12990 (for an /n/ allophone) different speech realizations.

The key stage of the T-ASR consists in the generation of a k -NN list for each phone within each word arc, by searching for best matching realizations in the template database. Euclidean distances for template matching proved to be a reliable metric probably due to the fact that the templates in the database had undergone data sharpening [17]. The computation of k -NN scores is illustrated in Fig.1 in black.

At this point, averaging the k -NN template scores, as firstly introduced in [8], has major advantages: ensemble scores significantly reduce the decoder complexity and, at the same time, improve the system performance as averaging alleviates the system’s sensitivity to outliers and mislabeling effects. Furthermore, the current, more careful implementation lends itself much better than its predecessors (e.g. [4]) for being used with large databases. All the more so in combination with S-CRF for exploiting different, alternative sources of information. Finally, before Viterbi decoding, word scores are obtained by accumulating phone scores for each word in the WG.

3.2. Integrating prosodic information

A number of prosodic features is extracted during the creation of the template database (Sec.2). These features are stored in the template database in an additional look-up table. Hence, the final information collected in each template is made of: 1) the actual acoustic realization as a sequence of acoustic feature (MIDA) vectors, 2) the transcription, and 3) a prosodic vector of dimension N (cf. Fig.1 in green/dotted).

During recognition prosodic information is extracted for each phone in the WG. The prosodic features are then compared to the prosodic features of the templates in the k -NN list. The idea is to check whether the prosodic characteristics of a phone match with those of its most similar templates stored in memory.

More specifically: first, all features are logarithmically transformed; then for each prosodic feature f , we compute the prosodic *score* as the L_1 distance of $\log f$ from the log of the average of the prosodic features f_i in the k -NN list:

$$s = \left| \log f - \log \frac{1}{k} \sum_{i=1}^k f_i \right| \quad (1)$$

The rationale behind the scores s , one for each of the N features in the prosodic table, is that they will be close to zero when a phone realization is supported by a set of templates with similar values; in the opposite case, the scores will grow and thereby penalize hypothesized phone realizations with aberrant prosodic features. There are two reasons for log-transforming the features: first, they are more conform to the LM log-probabilities and DTW scores, and second, they are invariant to the absolute feature values and focus on the relative differences instead. Finally, the worst scores were limited to $\log 2$ times the logarithm of the expected value, to avoid large penalizing scores. The *word-based* prosodic scores are obtained by accumulating phone scores within each word.

Before word graph decoding, the new word log-likelihoods are obtained as the linear combination of the word-based DTW- and the prosody-based scores; the logarithm of the LM score is one of the scores. As the number of scores (proportional to N) grows, finding the optimal weights becomes daunting. The segmental conditional random fields toolkit, SCARF [9], offers the possibility to combine word scores efficiently. SCARF iteratively calculates the posterior probabilities of the words by the forward-backward algorithm and adjusts the weights (on the

Table 1: *Word error rates [%] applying the single prosodic attributes using different time normalizations (functionals). Baseline T-ASR on the first line (‘-’).*

<i>prosodic attribute</i>	<i>functionals</i>			
	mean & st.dev.		min. & max.	
	<i>dev92</i>	<i>nov92</i>	<i>dev92</i>	<i>nov92</i>
-	7.06	9.11	7.06	9.11
duration	6.93	8.79	6.93	8.79
speaking rate	6.79	8.70	6.79	8.70
loudness	6.94	9.23	7.01	9.16
pitch	6.86	8.88	6.91	8.88
voice quality	6.84	8.86	6.79	8.98

dev92 dataset) so that the product of the word posteriors for the correct path increases at each iteration. Typically, convergence is reached already after 30 iterations.

4. Experiments

The testing material is the *20k* open vocabulary test set (*nov92*, 30 min., 8 speakers) which is evaluated using a trigram language model. Weight optimization is done on the non-verbalized punctuation parts of the *dev92 20k* and *dev92 5k* datasets together (64 min., 10 speakers). Note that this introduces a bias as the *5k* part does not contain out-of-vocabulary words; however the benefit of having extra material for optimization is expected to outweigh the downside of having a small bias. A graph error rate of 1.75% on the *dev92* and 2.68% on the *nov92* dataset allows for a good margin of improvement.

A first set of experiments was done to study the effectiveness of each single prosodic attribute in isolation (cf. Sec.2): Tab.1 presents an overview of the results using the two different groups of functionals. The general trend is that both groups offer almost identical time normalizations and no definitive conclusions can be drawn. One reason for the negligible differences is presumably that the time length of the units under analysis is quite short (101 ms on average) so that the effect of using different functionals is negligible.

The best performing prosodic attribute is *speaking rate*. Note that the speaking rate is implemented implicitly in the S-CRF framework by the pair of features ‘number of phones in the word’ and ‘duration’ which is well known to be important by itself. Nevertheless, all but one of the prosodic attributes show small but significant improvements and generalize quite well over the baseline: from 0.2% up to 0.4% absolute on the *nov92* dataset. The poor results obtained for loudness have a clear explanation: DTW-features, as distances between MIDA feature vectors, do already embed the short time energy (and derivatives).

In a second group of experiments we combined prosodic attributes together following a forward selection algorithm. Starting with the most simple attribute, duration, we tested in turn all other attributes and added the one which led to the best performance improvement on the *dev92* set. Taking into account the findings in Tab.1 two major choices were made: due to the poor results, loudness was not considered; only mean and standard deviations were applied.

As can be seen in Tab.2, all attributes contribute to the best result (from 9.11 to 8.51 on the *nov92* test set). However, the larger progress is realized when voice quality is pooled together with duration; after that, further improvements saturate

Table 2: Combination of prosodic attributes using mean and standard deviation as functionals. Figures are word error rates [%]. Baseline T-ASR results on the first line ('-').

prosodic attributes	# features	dev92	nov92
-	-	7.06	9.11
duration	1	6.93	8.79
+ pitch	4	6.98	8.79
+ speaking rate	2	6.79	8.70
+ voice quality	8	6.77	8.54
+ pitch	10	6.77	8.49
+ speaking rate	9	6.75	8.61
+ pitch	11	6.73	8.51

and gains are no longer statistically significant. Nevertheless, the system does not overfit to the *dev92* data when more attributes are added: even all of them together generalize well on the left out data (*nov92*). Further investigations are at bay to confirm these small but consistent improvements on other, larger datasets.

Finally, Tab.3 compares our T-ASR system to an HMM-based one (cf. Sec.3.1) without and with prosodic information being exploited. We call the latter system a *hybrid episodic-abstract* one [4]: while DTW scores are substituted with HMM-based acoustic models, prosodic features are still modelled non-parametrically, i.e. examples are stored in a (prosodic-only) template database.

5. Conclusions and future work

In this paper we described a template-based speech recognizer that exploits information derived from a number of prosodic sources. The system proved to be effective on a large vocabulary task: a performance gain of 7% relative was observed. Prosodic attributes were analysed separately and together confirming that more complex attributes (*voice quality*) might be beneficial to improve more simple though robust ones (*duration*). Furthermore, the same prosodic non-parametric modelling was successfully applied in combination with classic HMM-based acoustic likelihoods, here with smaller gains.

We expect that further improvements can be obtained by ameliorating and extending the present features, especially F0 extraction methods. An important prosodic attribute that is still missing is *pausing*. However, feature extraction and modelling might be not so straightforward. Pauses are silence intervals in an utterance that may contain breathing or background noise, but can also be long speech segments of almost uniform spectral characteristics, such as 'eh'.

Finally, the refinement of the architecture is also in order: one important improvement recently introduced in our T-ASR framework is the possibility of modelling multiple levels of symbolic representation, such as words (word templates are well suited for the frequent function words) and syllables (less frequent words may benefit from smaller and more generic units). As this study did not exploit this feature—we exclusively focused on the *extraction* of information at the phone level—the next logical step is to explore the multilevel and multi-timescale nature of speech by directly considering also longer word and sub-word units.

Table 3: Prosodic attributes (duration + voice quality) with an hybrid episodic-HMM ASR system. Figures are WER [%].

prosody:	T-ASR		HMM-ASR	
	-	✓	-	✓
<i>dev92</i>	7.06	6.77	6.02	5.98
<i>nov92</i>	9.11	8.54	7.85	7.69

6. References

- [1] D. Seppi, D. Falavigna, G. Stemmer, and R. Gretter, "Word Duration Modeling for Word Graph Rescoring in LVCSR," in *Proc. of Interspeech*, 2007, pp. 1805–1808.
- [2] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S. Kim, J. Cole, and J. Choi, "Prosody dependent speech recognition on radio news," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 232–244, 2006.
- [3] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The prosody module," in *VerbMobil: Foundations of Speech-to-Speech Translations*, W. Wahlster, Ed. Springer, 2000, pp. 106–121.
- [4] S. Demange and D. Van Compernelle, "HEAR: an hybrid episodic-abstract speech recognizer," in *Proc. of Interspeech*, 2009, pp. 3067–3070.
- [5] D. Kanevsky, T. N. Sainath, B. Ramabhadran, and D. Nahamoo, "An analysis of sparseness and regularization in exemplar-based methods for speech classification," in *Proc. of Interspeech*, 2010, pp. 2842–2845.
- [6] P. Nguyen, G. Heigold, and G. Zweig, "Speech recognition with flat direct models," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 994–1006, 2010.
- [7] K. Demuyneck, D. Seppi, D. Van Compernelle, P. Nguyen, and G. Zweig, "Integrating meta-information into exemplar-based speech recognition with segmental conditional random fields," in *Proc. of ICASSP*, 2011, pp. 5048–5051.
- [8] K. Demuyneck, D. Seppi, H. Van hamme, and D. Van Compernelle, "Progress in example based automatic speech recognition," in *Proc. of ICASSP*, 2011, pp. 4692–4695.
- [9] G. Zweig and P. Nguyen, "SCARF: A segmental conditional random field toolkit for speech recognition," in *Proc. of Interspeech*, 2010, pp. 2858–2861.
- [10] A. Kießling, *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Shaker, 1997.
- [11] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge," *Speech Communication*, 2011.
- [12] J. Kim and P. C. Woodland, "The Use of Prosody in a Combined System for Punctuation Generation and Speech Recognition," in *Proc. of Interspeech*, 2001, pp. 2757–2760.
- [13] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Communication*, vol. 46, no. 3-4, pp. 455–472, 2005.
- [14] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. of ACM Multimedia*, 2010.
- [15] W. J. Hess, "Pitch and Voicing Determination of Speech with an Extension Toward Music Signals," in *Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer, 2008, pp. 181–211.
- [16] K. Demuyneck, J. Roelens, D. Van Compernelle, and P. Wambacq, "SPRAAK: An open source speech recognition and automatic annotation kit," in *Proc. of Interspeech*, 2008, p. 495.
- [17] M. De Wachter, K. Demuyneck, and D. Van Compernelle, "Outlier correction for local distance measures in example based speech recognition," in *Proc. of ICASSP*, 2007, pp. 433–436.