Using Syllables in a Hybrid HMM-ANN Recognition System

Alfred Hauenstein *
Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, 81730 München, Germany
email: Alfred.Hauenstein@mchp.siemens.de

Abstract

An approach to speech recognition using syllables as basic modelling units is compared to a state-of-the-art system employing phonemes. The technological framework is a hybrid HMM-ANN[1] recognition system applied on small to medium vocabulary recognition tasks.

Although the number of units to be classified nearly doubles, it is shown that the syllable can outperform the phoneme slightly but significantly in terms of unit classification capability, measured as frame error rate. Comparing the overall system performance (measured in word error rate) the phoneme-based system still performs obviously better for continuous speech tasks, while the syllable-based system is superior for isolated word recognition tasks on cross-database tests. This suggests the need for further work on the understanding of the interaction of knowledge sources on the frame-, word-, and sentence-level in current recognition systems.

1 Introduction and motivation

While automatic speech recognition (ASR) systems are becoming mature for certain restricted tasks (e.g. phone mail access, control of desktop applications, office dictation), they still perform poorly on more difficult tasks, especially under real-world conditions such as with background noise, reverberation and spontaneous speech. One idea arising again recently is the proposal of a paradigm shift from phoneme-based to syllable-based recognition systems [5], [9].

Some of the potential advantages of syllable-based ASR are:

- The human auditory system integrates time spans of about 200 ms of speech [1], which corresponds roughly to the duration of syllables [5]. Thus the very robust human perception may be modelled more accurately by use of syllables instead of phonemes.
- The relative duration of syllables is less dependent on variations in speaking rate than the relative durations of phonemes [5]. Therefore the mismatch between the observation window for classification (feature vectors including multiple frames and Δ-components) and the duration of the unit classified is reduced for speakers whose speaking rate varies from the average.
- It was shown that time spans of 250 ms are suitable for methods of cepstral mean subtraction in order to suppress convolutional noise [6]. In this time span a stationary noise signal can be discriminated from a non-stationary speech signal. Shorter time spans (about 100 ms) may capture the stationary part of a vowel only and inhibit any distinction between speech and noise. This shows the potential advantage of a syllable-based approach, using windows of 200-250ms for unit classification.

The mentioned arguments make us believe that a syllable-based ASR system can be more robust than a phoneme-based one especially when dealing with spontaneous speech. As hybrid HMM-ANN systems easily incorporate temporal context for classification, we decided to compare the phoneme-based and the syllable-based approach in the framework of a hybrid HMM-MLP[2] speech recognition system developed at International Computer Science Institute (ICSI) [2], [10].

2 Recognition task

The recognition task chosen is referred to as "Numbers"; it is provided by OGI[3] [3]. The database contains spontaneously uttered telephone speech (analog and digital channels) in American English sampled at 8 kHz. Most utterances are continuous speech, but there are some isolated words as well. The vocabulary comprises 92 different words: digits (0-9, "oh"), cardinal numbers (e.g. "seventy", "thousand") ordinal numbers (e.g. "hundredth"), and non-numbers (e.g. "area", "code", "excuse", "it's").

1. Hidden Markov Model (HMM)-Artificial Neural Network (ANN)

2. Multi-Layer-Perceptron (MLP): a special ANN architecture.

3. Center for Spoken Language Understanding (CSLU) at Oregon Graduate Institute (OGI), Portland, Oregon, U.S.A.

* The work described was carried out during the authors visit at International Computer Science Institute, Berkeley, CA., U.S.A.

The bigram language model used for recognition has a test set perplexity of 12.97. The training set (both for acoustics and language modelling) contains 4.71 h of speech; the test set contains 2.23 h.

3 Phoneme-based baseline system

In order to obtain a baseline for comparisons with a syllable-based system, a "standard" phoneme-based recognizer is trained and tested.

3.1 Experimental outline

For the baseline system we extract RASTA-PLP features employing a frame rate of 12.5 ms and a window size of 25 ms. 17 features per frame are extracted (8 RASTA, 8 ΔRASTA, and 1 Δ-energy) [7].

The total phoneme set consists of 56 context independent phonemes, of which 49 are actually used for the Numbers task. Phonemes are not divided into sub-parts. Classification (i.e probability estimation for the phonemes) is done using a single-hidden-layer MLP. 9 consecutive frames of the feature vector are presented to the input layer, so the total input window length is 125 ms[4]. The number of input units N_I computes to: $N_I = 9 \times 17 = 153$. The neural net has $N_H = 1200$ hidden units (HUs) and $N_O = 49$ actually used output units. Thus, the total number of parameters of the MLP trained is $N_P = (N_I + N_O) \times N_H = 242,400$.

In an embedded training procedure we train alternatingly the MLP and the pronounciation lexicon. MLP training is started from randomized weights in order to get comparable results with the syllable-based training. For lexicon training the phoneme durations of all words are trained from the automatic segmentation of the training set. A single pronounciation lexicon is employed, which does not necessarily contain the canonical, but instead the most probable pronounciation found during training [12]. The use of multiple pronounciations did not significantly reduce the word error rate. Furthermore it is more straightforward to map the phoneme labels of a single pronounciation lexicon onto syllable labels needed for bootstrapping the syllable-based recognition system.

3.2 Evaluation

For performance evaluation two different measures are used. First, we compute a cross-validation (CV) accuracy on an independent CV data set. Thus the number of correctly classified frames with respect to the phoneme set used is obtained[5]. Therefore, the CV accuracy is a direct measurement of the classification step. Note, that the CV accuracy depends on the inventory of the recognizer (phonemes vs. syllables). The percentage of falsely classified frames is denoted as the "frame error rate".

Second, the word error rate is computed. This measures the overall system performance. The results of the phoneme-based baseline system are shown in table 1. The frame error rate is 17.55%, while the word error rate is 8.7%.

input window	N_P: No. of parameters	frame error rate	word error rate: total (sub/del/ins)
9 frames = 125 ms	242 400	17.55 %	8.7 % (4.8/2.6/1.3)

Table 1: Phoneme-based baseline

4 Syllable-based recognition system

Syllable-based ASR systems are common for Asian languages like Cantonese, Mandarin, and Japanese. Although the syllable-based approach was proposed for European languages like English and German, too ([4], [8], [11], and more recently: [9]), phoneme-based systems are state-of-the-art. To our knowledge, a syllable-based approach for speech recognition of English has never been published in the context of a hybrid HMM-ANN system.

4.1 Properties of the syllable-based recognizer

In the following we mention only where the syllable-based recognizer differs from the phoneme-based system, presented in section 3.

Starting from the phoneme-based single pronounciation lexicon, a program for syllabification is used, which accesses a database proposing syllable boundaries for pronounciation lexica. Thus, a syllable set of 96 different syllables is derived for the Numbers corpus.

Preliminary tests employing different feature sets (e.g. without Δfeatures or with additional use of a total-frame-energy-component) for the syllable-based recognizer showed no improvements, so that the feature set was kept the same like for the phoneme-based baseline.

Two different sizes of input windows of the MLP are investigated. For the first set of experiments we keep the same window size as used for the phoneme-based system (9 frames = 125 ms). Since a neural net employing 1000 HUs is used, the total number of parameters $N_P = (153 + 96) \times 1000 = 249,000$ is nearly the same as for the phoneme-based approach. Therefore, a fair performance comparison can be done.

In a second experiment, the input window is extended to 15 frames, equaling 200 ms. This value is derived from the temporal structure of syllables as presented in section 1 [5]. A longer time span was not considered since it

4. The window length comprises adjacent frames sampled together; it does not take into account Δcomponents since these are computed using decreasing polynomials, which means that the actual time span is not clearly defined.

5. The frame to be classified is the center of the input window.

would introduce problems recognizing very short mono-syllabic words (all digits, besides "seven" and "zero", are mono-syllabic). Since the number of input units grows to $N_I = 15 \times 17 = 255$, we reduced the number of hidden units to $N_H = 700$, in order to get a comparable amount of parameters $N_P = (255 + 96) \times 700 = 245,700$.

Comparing the architecture of the phoneme-based baseline system and the syllable-based recognizer shows two differences: use of a different unit used for classification resulting in an approximately two-fold larger number unit inventory, and use of different input-window sizes for the classification step.

4.2 Evaluation

Experimental results are shown in table 2. It is obvious that for the standard 9-frame input the frame error rate of the syllable-based system (20.95%) is worse than for the phoneme-based system (17.55%). As expected, the frame error rate improves drastically to 16.24% by extending the acoustical context to 15 frames. As a consequence, the phoneme-based system is outperformed slightly but significantly[6] (relative drop of frame level error rate of 7.5%).

For the word error rate we found the phoneme-based system still performs much better (relatively 38% less errors than the 15-frame input system), while the 15-frame input system performs better than the 9-frame input syllable-based system. But the improvement when using 15-frame input instead of 9-frame input is not as high as expected from the results for frame error rate. It is important to notice that especially the number of deletions for the syllable-based system is much worse than for the phoneme-based system. This problem could not be handled satisfactorily by adapting the heuristic parameters that are responsible for smoothing language model and acoustical probabilities. This might be the reason for the comparably small improvement in word error rate of the syllable-based 15-frame input system over the syllable-based 9-frame input recognizer.

input window	N_P: No. of parameters	frame error rate	word error rate: total (sub/del/ins)
9 frames = 125 ms	249 000	20.95 %	15.2 % (5.7/7.1/2.4)
15 frames = 200 ms	245 700	16.24%	14.0 % (4.8/7.2/2.0)

Table 2: Syllable-based recognizer

5 Cross-database tests

One of the proposed advantages of a syllable-based speech recognition system is the robustness against changes of speaking style (e.g. speaking rate and spontaneous vs. controlled speech) and channel distortions (additive or multiplicative noise), especially if these changes were not seen during training but do occur in test. Therefore, a set of cross-database tests is performed. This means, lexica and MLPs presented in previous sections are applied for recognition on different databases without any adaptation.

The vocabulary of the databases tested needed to be a real subset of the training database in order to cover all syllables of the testset. Thus, the vocabulary for all tests comprises the digits (11 words vocabulary: 0-9, "oh") only. Word error rate is used for performance comparison. The results are summed up in table 3.

Test database	Bellcore-digits		Voicemail digits	TI-digits
	clean	10 SNR		
	2200 isolated digits		2200 isolated digits	8700 utt. (cont. digits)
9-frame-input / phoneme-based	4.5 % (4.5 / 0.0 / 0.0)	25.1 % (25.1 / 0.0 / 0.0)	2.1 % (2.1 / 0.0 / 0.0)	3.3 % (1.1 / 1.7 / 0.5)
15-frame input / syllable-based	2.5 % (2.5 / 0.0 / 0.0)	20.7 % (20.7 / 0.0 / 0.0)	1.9 % (1.9 / 0.0 / 0.0)	8.0 % (0.9 / 5.9 / 1.3)

Table 3: Word error rates for cross-database tests (trained on "Numbers"-task)

The following databases (all: American English) were tested:

- Bellcore digits[7]: isolated digits; telephone speech sampled at 8 kHz; 2200 utterances by 200 speakers. This database was used two-fold. First, recognition was performed on the data as provided: "clean" speech (besides "usual" telephone channel distortion). Second, recognition tests were performed after adding car noise with 10 dB SNR [7]. Especially under noisy conditions the syllable-based system performs significantly better (relative drop of word error rate of 17.5%).
- Voicemail digits[8]: same recording conditions, vocabulary and size as Bellcore digits. The syllable-based system perform slightly but not significantly better.
- TI continuous digit string (testset): 8700 utterances of digit strings; downsampled from 20 to 8 kHz, a sim-

6. significance is tested for a 0.001 level employing a binomial significance test.

7. provided by Bellcore to ICSI for research purposes

8. provided by Siemens AG to ICSI for research purposes

ple language model penalizing all word transitions identically is used. Here, for the only continuous speech cross-database test, the syllable-based recognizer is significantly poorer. Again, like for the test on the Numbers-corpus, the very high number of deletions is crucial for the performance of the syllable-based system.

6 Conclusions and further work

Comparing the results of the phoneme-based and the syllable-based recognizer, it can be seen that a syllable-based system using an extended input window of 200 ms (compared to 125 ms for phonemes) shows significantly better basic classification performance (measured in frame error rate) for the basic units employed. This becomes even more noteworthy when taking into account that the number of units to be classified nearly doubles (from 49 to 96) and therefore the chance of confusions increases, while the total numbers of parameters of the systems are comparable.

When measuring word error rates for cross-database isolated word recognition tasks the syllable-based system performs better than the phoneme-based one. On the other hand, when comparing overall system performance of continuous speech tasks the phoneme-based system still shows obviously lower error rates. Especially the high percentage of deletions of the syllable-based systems is of particular concern.

Consequently, we hypothesize that the higher word error rate is due to two causes, which directs the way for upcoming work. First, the interaction of knowledge sources on the frame-, word-, and sentence-level in current recognition systems needs to be studied in more detail in order to come up with a theoretical framework instead of heuristics[9]. Second, a more detailed look on the syllabification process seems to be useful. This is of particular interest since the recognition rates for isolated digits tasks, where most syllables are unique to one word, showed good word recognition performance. Additionally, this might be due to the relative high proportion of digits in the training set, which might suggest that syllable-based systems need a high number of training events per unit.

7 Literature

[1] H. Bourlard, "Towards Increasing Speech Recognition Error Rates", Proc. Eurospeech 1995, pp. 883-894.

[2] H. Bourlard, N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academics Publishers, 1994, U.S.A.

[3] R.A. Cole, M. Noel, T. Lander, T. Durham, "New Telephone Speech Corpora at CSLU", Proc. Eurospeech, 1995, pp. 821-824.

[4] O. Fujimura, "Syllable as a unit of speech recognition." IEEE Trans. on ASSP (Acoustics, Speech, and Signal processing), vol.ASSP-23, no. 1, 1975, pp. 82-87.

[5] S. Greenberg, "Understanding Speech Understanding: Towards a Unified Theory of Speech Perception", Proc. ESCA Workshop on The Auditory Basis of Speech Perception, Keele Univ., U.K., July 1996.

[6] A. Hauenstein, E. Marschall, "Methods for Improved Speech Recognition over Telephone Lines", Proc. ICASSP 1995, pp. 425-428.

[7] H. Hermansky, N. Morgan, "RASTA Processing of Speech", IEEE Trans. on SAP (Speech and Audio Processing), vol. 2, no. 4, Oct. 1994, pp. 578-589.

[8] M.J. Hunt, M. Lennig, P. Mermelstein, "Experiments in Syllable-based Recognition of Continuous Speech", Proc. ICASSP 1980, pp. 880-883.

[9] K. Kirchhoff, "Phonologically Structured HMMs for Speech Recognition", Proc. 2nd meeting of the ACL SIG in Computational Phonology, Santa Cruz, U.S.A., June 1996, pp. 45-50.

[10] N. Morgan, H. Bourlard, "Neural Networks for Statistical Recognition of Continuous Speech", Proc. IEEE, May 1995, pp. 742-770.

[11] G. Ruske, B. Plannerer, T. Schultz, "Stochastic modelling of syllable-based units for continuous speech recognition", Proc. ICSLP 1992, pp. 1503 - 1506.

[12] C. Wooters, A. Stolcke, "Multiple-pronunciation Lexical Modeling in a Speaker-independent Speech Understanding System", Proc. ICSLP 1994, pp. 1363-1366.

Acknowledgments

The work described in this paper was carried out during the authors visit at International Computer Science Institute (ICSI), Berkeley, CA, U.S.A. We would like to thank everyone who made this stay possible and enjoyable.

First of all, acknowledgments go to Steven Greenberg at ICSI, who inspired this work by the untiring collection of shortcomings in current speech recognition systems and proposing more perceptually oriented approaches. Many thanks to Su-Lin Wu at ICSI who was a great help for many big and small problems and who provided the syllabification program. And, of course, many thanks to Nelson Morgan, the fearless leader of ICSI's speech recognition group, for many helpful comments.

Last, but not least, many thanks to all my colleagues at Siemens, who made my visit at ICSI possible by taking up my otherwise unattended duties.

9. see [1] for a discussion of this problem.