COMBINING ARTICULATORY AND ACOUSTIC INFORMATION FOR SPEECH RECOGNITION IN NOISY AND REVERBERANT ENVIRONMENTS

Katrin Kirchhoff

AG Angewandte Informatik
Technische Fakultät, Universität Bielefeld
Postfach 100 131, 33501 Bielefeld, Germany
katrin@techfak.uni-bielefeld.de

ABSTRACT

Robust speech recognition under varying acoustic conditions may be achieved by exploiting multiple sources of information in the speech signal. In addition to an acoustic signal representation, we use an articulatory representation consisting of pseudo-articulatory features as an additional information source. Hybrid ANN/HMM recognizers using either of these representations are evaluated on a continuous numbers recognition task (OGI Numbers95) under clean, reverberant and noisy conditions. An error analysis of preliminary recognition results shows that the different representations produce qualitatively different errors, which suggests a combination of both representations. We investigate various combination possibilities at the phoneme estimation level and show that significant improvements can been achieved under all three acoustic conditions.

1. INTRODUCTION

Whereas most speech recognition systems use a cepstral or spectral representation of the speech signal, there have also been attempts at using articulatory information. This includes parameters derived from actually observed articulatory trajectories [12, 15], as well as heuristically defined articulatory features which are inferred from the speech signal using statistical classifiers [3, 2, 4]. These attempts have been motivated by two major assumptions: first, coarticulation can be modelled more easily in the production-based domain than in the acoustic domain. Second, it is assumed that articulatory parameters are more robust towards cross-speaker variation and signal distortions such as additive noise. A third assumption can be made, namely that acoustic and articulatory representations of speech are mutually complementary information sources whose combination in a speech recognition system might be beneficial.

Previously, articulatory-based speech recognizers have primarily been developed for clean speech; the potential of an articulatory representation of the speech signal for noisy test conditions, by contrast, has not been explored. Moreover, there have barely been attempts at systematically combining articulatory recognizers with standard acoustic recognizers. This paper investigates the second and third of the above assumptions by reporting speech recognition experiments on a variety of acoustic test conditions (clean, reverberant, and additive pink noise) for individual acoustic and articulatory speech recognizers, as well as for a combined system using both representations. In the following we will first describe the baseline recognition systems. Section 3 presents an error analysis of initial recognition results; Section 4 discusses various classifier combination schemes and word recognition results using a combined system.

2. BASELINE SYSTEMS AND SPEECH MATERIAL

2.1. Speech Corpus

Training and recognition were carried out on the OGI Numbers95 corpus [1], which consists of continuously spoken numbers recorded over both analogue and digital telephone lines from a broad set of speakers. The results reported here were obtained on the "core-subset". The training set consists of 3590 utterances, 347 of which were used for cross-validation during MLP training. The test set consists of 1206 utterances. The recognition lexicon contains 34 words. Six different test sets reflecting three acoustic conditions were used. The first is the clean test set[1]. The second test set is a digitally reverberated version of the clean test set, using an impulse response measured in an echoic room (0.5 seconds reverberation time). Finally, four test sets were generated by adding pink noise to the clean test set at various signal-to-noise ratios: 30 dB, 20 dB, 10 dB, and 0 dB.

2.2. Acoustic Baseline Systems

All systems used for the experiments reported in this paper are hybrid ANN/HMM recognizers using 29 context-independent phones. Phone probabilities are estimated by three-layer MLPs using online error back propagation and the softmax activation function. Systems vary with respect to the acoustic preprocessing and the size of the hidden layers. All systems use the same back-off bigram and the same recognition lexicon. It should be noted, however, that the recognition lexicon was developed for the acoustic baseline systems and was not optimized for the articulatory systems. All systems were trained iteratively; at each step, training labels were re-generated from a forced alignment of the previously trained models with the speech files.

Two acoustic baseline systems were employed for the experi-

[1] "Clean" means that no artificial noise was added. However, this test set does contain natural background noises.

Feature Group	Features
Voicing	+voice, -voice, silence
Manner	stop, vowel, fricative, approximant, nasal, lateral, silence
Place	dental, labial, coronal, palatal, velar, glottal, high, mid, low, silence
Front-Back	front, back, nil, silence
Lip Rounding	+round, -round, nil, silence

Table 1: Initial articulatory feature set

ments reported in this paper: the first system (baseline I), uses 8 log-RASTA-PLP [6] coefficients and their first derivatives, computed every 10 ms with a window of 25 ms. The input to the MLP consists of nine frames and the hidden layer has 400 units. Baseline I is used for the clean test set. The second system (baseline II) also uses an input window of nine frames and the number of hidden units (HUs) is 560. The acoustic parameterization consists of 15 modulation spectrogram features [5]. These features are derived from a critical-band-like filterbank with subsequent computation of normalized amplitude envelopes in each channel. These are then filtered to estimate the spectral energy of modulations between 2 and 16 Hz. This front end deemphasizes fine-grained phonetic detail like onsets and formant trajectories and enhances spectral changes which roughly correspond to the syllabic rate of speech. Modulation spectrogram features have been demonstrated to be very robust in noisy and reverberant environments [5, 14]. Baseline II is used for the reverberant and noisy test sets.

2.3. Articulatory Systems

For both acoustic baseline systems, corresponding articulatory systems were built using a set of heuristically defined articulatory features describing manner and place of articulation. The entire set of features is divided into subsets according to orthogonal articulatory dimensions (see Table 1). For each subset, a separate MLP was trained on the acoustic parameterization whose output units correspond to the articulatory classes in that subset. The context window varies between 5 and 9 frames; the hidden layer sizes ranges from 50 to 100 HUs. These were chosen to maximize the recognition accuracy while minimizing the number of parameters. The posterior feature probabilities output from each network are concatenated and passed on to a higher-level integrative MLP which maps them to the desired phone probabilities. This MLP uses a context of 9 frames and has 380 HUs.

In order to make the systems comparable in terms of the number of parameters in the phone estimation network, the initial set of articulatory features was subject to an information-theoretic feature selection algorithm [9]. This procedure successively eliminates irrelevant and/or redundant features from the initial set while minimizing the relative entropy between the phoneme distribution given the original feature set and the phoneme distribution resulting from the reduced feature set. Approximations to the true conditional distribution are computed on the training set. In preliminary experiments, this algorithm proved superior to principal components analysis and allowed us to reduce the initial feature set to 18 features. Voicing features and all silence features were eliminated completely; furthermore, the features

System	WER	INS	DEL	SUB
clean AC	8.4	2.0	1.7	4.7
clean AF	8.9	1.5	2.0	5.4
reverb AC	22.1	1.8	5.9	14.4
reverb AF	23.7	3.1	4.7	16.0
30 dB noise AC	**15.5**	2.8	2.2	10.5
30 dB noise AF	**17.4**	2.4	3.4	11.6
20 dB noise AC	20.3	4.9	2.7	12.7
20 dB noise AF	21.7	4.3	3.6	13.9
10 dB noise AC	31.3	10.3	3.2	17.8
10 dB noise AF	30.0	6.1	5.7	18.3
0 dB noise AC	**50.8**	18.0	4.9	27.9
0 dB noise AF	**43.6**	7.1	10.2	26.3

Table 2: Baseline word error rates (in %), AC = acoustic, AF = articulatory

approximant, *dental* and *front-back-nil* are missing from the reduced feature set.

2.4. Baseline Recognition Results

Tables 2 shows the baseline word error rates for clean, reverberant and noisy speech. Results for the acoustic (AC) and the articulatory feature (AF) system are comparable; the only statistically significant differences are those between the bold-printed numbers. in Table 2[2].

3. ERROR ANALYSIS

Although both systems yield similar word error rates, this is no indication of their performance at the level of subword unit classification. In order to ascertain whether the different systems made different errors at this level, the frame-level phone confusion matrices were analyzed. Figures 1 to 3 show graphic representations of the diagonals of the confusion matrices for all acoustic test sets. These reveal qualitative differences be-

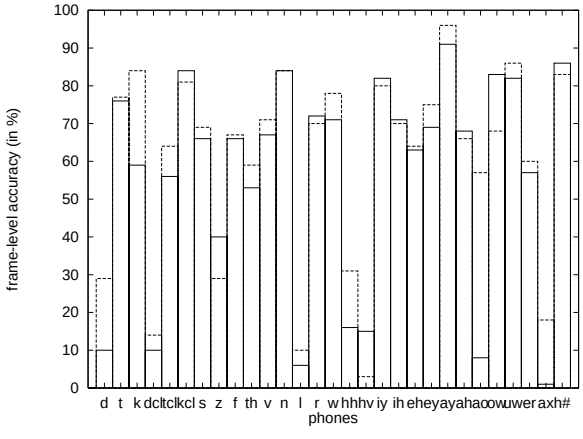

Figure 1: Frame-level phone accuracies, clean speech. Solid lines represent the acoustic system, dashed lines the articulatory system.

[2]Statistical significance is based on a difference of proportions test. Results at a level ≤ 0.05 were considered significant.

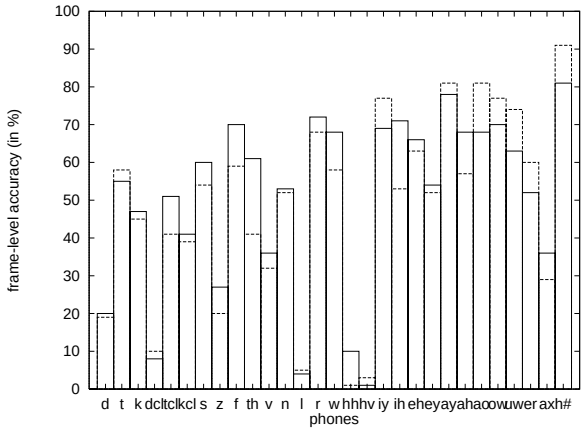

Figure 2: Frame-level phone accuracies, reverberant speech. Solid lines represent the acoustic system, dashed lines the articulatory system.

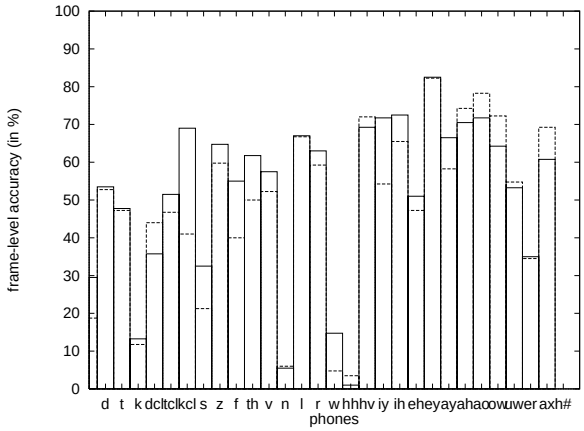

Figure 3: Frame-level phone accuracies, averaged over all noisy test sets. Solid lines represent the acoustic system, dashed lines the articulatory system.

tween the articulatory and acoustic recognition systems. In the RASTA-based systems, consonantal segments, especially voiceless stops and fricatives are classified more accurately in the articulatory systems whereas the acoustic system does better on vocalic segments. As far as the modulation spectrogram based systems are concerned, consonants are consistently modelled better by the acoustic systems, whereas certain vowels (especially /ao,ow,uw,ax/) and silence are better distinguished by the articulatory systems. Most of these class-specific differences in recognition accuracy are highly statistically significant.

The 0 dB noise test case deserves special attention because here the articulatory system shows a markedly better performance than the acoustic system. A closer look at the frame-level phone accuracies showed that the main portion of the error reduction in articulatory system was due to a better separation of silence, voiceless fricatives, and voiceless plosives. Thus, since the acoustic and articulatory classifiers produce characteristically different errors, it might be beneficial to combine them.

System	FER	WER	INS	DEL	SUB
clean	22.53	**7.3**	1.2	1.6	4.4
reverb	30.25	**20.3**	3.6	3.1	13.6
30 dB noise	26.71	15.0	2.6	2.1	10.3
20 dB noise	32.13	**18.4**	2.8	2.8	12.7
10 dB noise	40.96	**27.9**	6.2	4.3	17.4
0 dB noise	52.62	**41.0**	5.9	10.8	24.3

Table 3: Frame and word error rates (in %) for combined system, product rule combination

4. CLASSIFIER COMBINATION

4.1. Combination Rules

Classifier combination is widely used in the machine learning community [7, 13] and has more recently been applied to speech recognition [10, 14]. Since the MLPs outputs can be interpreted as Bayesian a posteriori probabilities, the phoneme classifiers in our hybrid systems easily lend themselves to combination by means of standard linear probability combination rules. The two most widely used combination rules [8] are the product rule and the sum rule. Given N classifiers $c_1, ..., c_N$ and K classes $\omega_1, ...\omega_K$, the product rule computes

$$P(\omega_k|x_1, ...x_N) = \frac{1}{P(\omega_k)^{N-1}} \prod_{n=1}^{N} P(\omega_k|x_n) \qquad (1)$$

where x_n is the feature vector input to classifier n and $P(\omega_k)$ is the a priori probability for class k. This rule rests on the assumption that the input representations given the classes are statistically independent and that classes have equal priors. Under the assumption of equal priors, the sum rules computes the average of the classifier output probabilities:

$$P(\omega_k|x_1, ..., x_N) = \frac{1}{N} \sum_{n=1}^{N} P(\omega_k|x_n) \qquad (2)$$

It has been observed [8] that sum rule combination effects a dampening of estimation errors of the individual classifiers, whereas errors are amplified when outputs are fused by the product rule. Thus, sum rule combination is potentially more robust to noisy input, which might be an advantage in acoustically instable environments.

4.2. Combination Recognition Results

Both of the above rules have been applied to combination under all acoustic conditions. The frame-level accuracy rates for the combined output and the corresponding word error rates are shown in Table 3 for the product rule combination scheme and in Table 4 for sum rule combination.

The word error rates in bold print in Table 3 are significantly better than the corresponding best system in Table 2. As far as the frame error rate is concerned, the sum rule scheme achieves a better result than the product rule combination method in noisy (20 dB, 10 dB, and 0 dB SNR) test cases and in the

System	FER	WER	INS	DEL	SUB
clean	21.76	8.0	0.9	2.3	4.7
reverb	31.46	20.9	1.2	5.3	14.4
30 dB noise	27.08	15.9	2.1	2.9	11.0
20 dB noise	31.96	20.3	2.6	4.0	13.7
10 dB noise	40.38	28.7	3.7	7.1	17.9
0 dB noise	52.16	43.8	9.3	8.1	26.5

Table 4: Frame and word error rates (in %) for combined system, sum rule combination

RASTA-based system; these differences are statistically significant. However, the product rule combination scheme always produces an equivalent or better word error rate.

5. DISCUSSION AND CONCLUSIONS

We have presented an approach to robust processing of speech using information from both the acoustic and the articulatory domain. Whereas the acoustic systems perform slightly better in the case of clean speech and noise at high signal-to-noise ratios, the articulatory systems show a distinct advantage in the presence of noise at low signal-to-noise ratios. Furthermore, it was shown that although individual recognizers based on the different representations yield similar word error rates, they provide different information at the frame-level. Simultaneous exploitation of the two information sources by means of a linear combination of the phoneme classifier outputs further improves word recognition. Obviously, the additional level of pre-phoneme classification in terms of articulatory features seems to help identify segments which are highly confusable on the basis of the acoustic representation alone. These findings suggest a novel approach to the dichotomy between acoustic or perception-based representations and production-based representations, viz. the combination of both in acoustic contexts where they can be shown to provide different information. It should be possible to eliminate the full articulatory feature representation of the signal. Instead, individual articulatory classifiers might be used which target that subset of the subword classes which receives a poor classification in the acoustic feature space (cf. also [11]).

Of the two combination rules which we investigated, the sum rule shows a tendency to achieve better frame-level accuracy rates in the RASTA-based system and in the noisy test cases, which might support previous observations that sum rule combination is more robust towards estimation errors produced by the individual classifiers. The product rule, by contrast, achieves a lower word error rate. This is probably not due to independence of the input representations; the articulatory representation has in all cases been derived from the acoustic representation and is unlikely to be completely independent from it. A more plausible reason is that product rule combination produces phoneme distributions which interact more favourably with the structure of the word recognition lexicon (pronunciation variants, minimum durations, etc.). This suggests that a combination scheme should be applied which is designed to minimize word-error rate directly.

Acknowledgements
This work was supported by the German Research Association (DFG) within the graduate research program “Task-oriented communication” and was carried out at the International Computer Science Institute, Berkeley, USA. Thanks are due to Nikki Mirghafori and Brian Kingsbury for providing the acoustic baseline systems, to Brian Kingsbury for providing the noisy test sets, and to Steve Greenberg, Nelson Morgan, and Jeff Bilmes for fruitful discussions about this work.

6. REFERENCES

1. R. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at CSLU. *Eurospeech*, 1:821–824, 1995.
2. L. Deng and D. Sun. A statistical approach to ASR using atomic units constructed from overlapping articulatory features. *JASA*, 95:2702–2719, 1994.
3. K. Elenius and G. Tacacs. Phoneme recognition with an artificial neural network. *Eurospeech*, pages 121–124, 1991.
4. K. Erler and G.H. Freeman. An HMM-based speech recognizer using overlapping articulatory features. *JASA*, pages 2500–2513, 1996.
5. S. Greenberg and B.E.D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. *ICASSP*, 2:1647–1650, 1997.
6. H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2:578–589, 1994.
7. T.K. Ho, J.J. Hu, and S.N. Srihari. Decision combination in multiple classifier system. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 16:66–75, 1994.
8. J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, 1998.
9. D. Koller and M. Sahami. Toward optimal feature selection. In L. Saitta, editor, *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kaufmann, 1996.
10. B. Mak. Combining ANNs to improve phone recognition. *ICASSP*, 4:3253–3256, 1997.
11. P. Niyogi and P. Ramesh. Incorporating voice onset time to improve letter recognition accuracies. *ICASSP*, 1:721–724, 1998.
12. J. Papcun, T.R. Hochberg, F. Thomas, J. Larouche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data. *JASA*, 92:688–700, 1992.
13. K. Woods. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:405–410, 1997.
14. S.-L. Wu, B.E.D. Kingsbury, N. Morgan, and S. Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. *ICASSP*, pages 721–724, 1998.
15. J. Zacks and T.R. Thomas. A new neural network for articulatory speech recognition and its application to vowel identification. *Computer, Speech and Language*, 8:189–209, 1994.