ROBUST AND COMPACT MULTILINGUAL WORD RECOGNIZERS USING FEATURES EXTRACTED FROM A PHONEME SIMILARITY FRONT-END

Philippe Morin, Ted H. Applebaum, Robert Boman, Yi Zhao and Jean-Claude Junqua

Panasonic Technologies, Inc. / Speech Technology Laboratory, 3888 State Street, Suite #202

Santa Barbara, California, 93105, U.S.A.

ABSTRACT

In this paper we characterize the sensitivity of two speaker-dependent isolated word recognizers toward several kinds of variability and distortions; namely noise, channels, distance to microphone and target language. Both recognizers use a phoneme similarity acoustic front-end as a rich representation for speech from which reliable features are extracted. A cross-correlation test showed that a phoneme similarity front-end is more robust to variability and distortions (especially intra-speaker variability) than a LPC cepstral front-end. The first recognizer (Condor) uses a frame-based approach while the second (Pasha) uses the phoneme similarity information contained in a small number of speech segments. The two recognition methods are presented with a special emphasis on the robustness improvements and computational trade-offs that have been made. Experimental results are reported for car noise at different speeds, speakerphone versus handset input in an office environment and several target languages. Recognition accuracy greater than 94% was achieved in a car environment at 60 mph (Condor) and recognition accuracy greater than 95% was achieved for speakerphone input at a distance of 50 cm. in an office environment.

1. INTRODUCTION

Phonemes are essentially discriminated by spectral trajectories which extend well beyond the 10 to 20 milliseconds encompassed in short term spectral analysis methods. Concatenating consecutive analysis vectors into a time-spectral pattern (TSP) vector capture these trajectories. To use these long feature vectors in speech recognition, however, requires an efficient representation; one such representation is phoneme similarities. Phoneme similarities have been used for speech recognition in frame-by-frame dynamic programming matching procedures [1,2,3,5], a continuous density HMM [1] and a matching procedure based on regions of high phoneme similarity [3,4,5]. As phoneme similarities are computed over several consecutive frames of speech, they capture both static and dynamic spectral characteristics. They have been shown to be relatively insensitive to variations between speakers [2] for recognition of isolated words.

2. PHONEME SIMILARITIES

Phoneme similarities can be computed from any fixed frame-rate acoustic analysis, such as LPC cepstral coefficients or filter-bank energies. A fixed number of consecutive analysis vectors are concatenated to form time-spectral pattern vectors. If the time-spectral pattern vectors in each phoneme class p are adequately described by normal distributions with separate means (μ_p) but common covariance (W), a time-spectral pattern vector x can be classified in the phoneme class p which maximizes the linear classification function L_p:

$$L_p = (2\mu_p W^{-1})x - (\mu_p W^{-1}\mu_p)$$

Phoneme similarity values (S_p) are in the form of posterior probabilities, and are computed from the linear classification function by exponentiation and scaling of the vector of P linear classification functions to unit norm.

$$S_p = e^{\alpha L_p} \Big/ \sqrt{\sum_{i=1}^{P} e^{2\alpha L_i}}$$

Increasing the constant exponential factor (α) emphasizes the similarity values of the dominant phoneme, and inhibits the lesser phonemes in the vector of similarity values. The result is that the time series of phoneme similarity values may be approximated as a near-zero background level punctuated by regions of high phoneme similarity as shown in Figure 1.

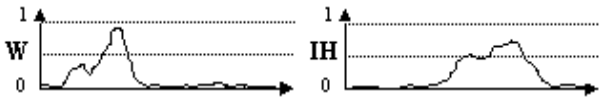

Figure 1: Phoneme similarity values versus time for two major discriminating phonemes in the name "Will".

A training procedure for phoneme similarity reference models, based on the TIMIT database, was presented in [5].

3. ROBUSTNESS

3.1. Comparison of LPC Cepstrum Features and Similarity Features

A cross correlation test was used to quantify the benefit of a similarity feature based approach over raw cepstrum features. To compare each approach, pairs of utterances were aligned using dynamic time warping. Between the pair, the cross correlation coefficient was computed for each feature as given by the following equation:

$$r = \frac{n\sum x_i y_i - (\sum x_i \sum y_i)}{\sqrt{n\sum x_i^2 - (\Sigma x_i)^2}\sqrt{n\sum y_i^2 - (\Sigma y_i)^2}}$$

For the cepstrum, the correlation coefficients were averaged over its 12 features. For the similarities, the 12 phonemes with the greatest coefficients were averaged. Figure 2 shows that similarity features are consistently more repeatable across

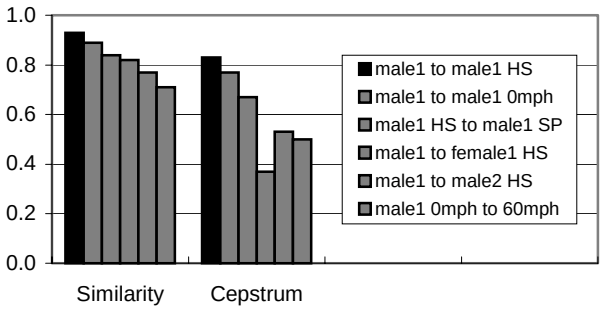

Figure 2: Similarity vs. LPC Cepstrum Cross-Correlation.

matched conditions and various mismatched conditions, such as different talker, different speed of car noise or handset (HS) versus speakerphone (SP). On the telephone speech, each value was computed using 100 utterances (about 85 seconds) versus 25 utterances (about 15 seconds) on the car speech. Phoneme models were tuned to a car environment using multi-style training (see below) but not to the handset/speakerphone environment.

3.1. Multi-style Training

Training utterances may be artificially modified by additive noise or linear filtering, and re-used as additional training repetitions. This simulates more varied testing conditions at training time and results in useful statistics on the variability of features used for recognition. This multi-style training technique is used 1) for the training of phoneme reference models and 2) for the training of word models. For the cross-correlation experiment described above, clean data and clean data plus car noise at 5dB SNR was used. Multi-style training in the case of a phoneme similarity front-end with additive noise was shown in [5] to decrease isolated word recognition error rates rate by 45 to 51% for multi-style training of both phoneme and word models.

3.2. Speech Equalization

Speech equalization is a noise-masking technique that was recently developed. It aims at decreasing the mismatch between training and testing utterances especially in the case of mismatch channels such as handset and speakerphone microphones. The equalization is done in the time domain and is therefore fairly inexpensive. It is driven by three targets: a channel spectral shape, a background noise level and a speech-to-noise ratio (SNR). After applying a channel-dependent filter, the equalization procedure adds noise to the input signals to keep the background level and the SNR near their respective targets in real-time. In the case of the handset/speakerphone problem, the method is used to transform handset data into speakerphone data.

3.3. Automatic Endpoint Detection

To better cope with speech, channel and environment variability, a loose endpoint detection strategy was elected. In effect the recognizers use other sources of knowledge (e.g. similarity background noise) or spotting techniques to better refine those endpoints. To deal with fairly noisy environments such as the car, energy based endpoint detectors are not robust enough. In such cases, spectral subtraction and band-pass filtering (300-3400Hz frequency band) was used to generate more accurate endpoints.

4. SPEECH RECOGNITION METHODS

Phoneme similarities is a rich and redundant form of speech representation that may be used for speech recognition in many ways. However robust feature extraction techniques are generally required when targeting low-end hardware platforms. Such techniques include:

- aligning time series of multiple high phoneme similarities and delta similarities frame-by-frame (Condor system described below).
- aligning reliable regions of high phoneme similarity (TC stage in Alibaba [4]).
- comparing the number of high phoneme similarity regions found in a fixed number of segments in the utterance (RC stage in Alibaba [4]).
- comparing the typical similarity level found in a fixed number of segments in the utterance (Pasha system described below).

4.1. Similarity Frame-Based Method

The Condor system performs frame-by-frame alignment of the time series of phoneme similarity vectors for an unknown utterance to a frame-based word reference. Each word reference frame contains the N similarity values and N delta similarity values of largest magnitude, and the corresponding 2N phoneme identifiers; typically, N is set to six out of a total of 55 phonemes. Word references are created from one or several training utterances by aligning the training utterances with

dynamic time warping and averaging the frame data. The local distance for the alignment is a convex combination of the cosine of the angle between the test and reference similarity vectors and the cosine of the angle between the test and reference delta similarity vectors [2].

Condor has relatively large word reference models, and high computational complexity. However the high redundancy in the speech representation due to storing multiple phoneme similarities at small time increments affords more robust performance in severe environments, such are the car.

4.2. Similarity Segment-Based Method

Unlike Condor, the Pasha system uses a segment-based approach to represent words. It is a robust extension of Alibaba's fast-match stage (RC stage) presented in [4]. The recognition strategy in RC was based on the number of high similarity regions found is each segment. While it performed well as a fast-match stage (top 10 candidates), it was not found accurate and robust enough as a recognizer especially in the case of noise and channel mismatch. The degradation was shown to come from 1) the segmentation method (segments were identified by dividing the utterance into S segments of equal duration) and 2) the use of thresholds in the detection of high similarity regions within the similarity time series. In Pasha, the segmentation was greatly improved and instead of a discrete approach a continuous approach was implemented to account primarily for the information held within the high similarity regions.

In Pasha, an input utterance is first divided into S segments such that the phoneme similarity density over all phonemes is equal in each segment. In this process, the similarity background is subtracted out to better cope with potential automatic endpoint errors. Then the average Root Sum Square (RSS) value of each phoneme segment is computed along with its variance (used as a weighting factor) over the training utterances.

This method allows for a static alignment of the test utterance that is word-independent. Pasha has small word reference models and low computational complexity, while giving adequate recognition accuracy in home or office environments.

	Condor	Pasha
Time Increment	Frame (50/sec.)	Segment (3/word)
Alignment Method	Dynamic Time Warping	None (Static Alignment)
Local Distance	Correlation Cosine	Weighted Euclidean Distance
Word Model Parameters	1200/sec.	330 /word (55 phonemes)
Compacted Word Model Size	800 bytes/sec.	110 bytes/word

Table 1: Comparison of recognition methods used in Condor and Pasha systems.

Condor and Pasha make good use of phoneme similarities in the sense that they both achieve high resolution. In Condor the high resolution is in the time domain while in Pasha the high resolution is in the phoneme domain.

5. EXPERIMENTS

Recognition accuracy has been evaluated for each recognizer in separate conditions.

5.1. Car Environment (Condor)

Ten subjects recorded 25 common English names in one of two compact cars with a microphone mounted on the sun-visor. The

Testing Condition (Speed)	Recognition Accuracy
Parked	99.2
30 MPH	98.8
60 MPH	94.6

Table 2: Recognition accuracy in a car using Condor on 25 names with automatic endpoints.

Condor system was trained with two repetitions of the vocabulary while parked. Testing was done with one repetition recorded while parked, one at 30 miles-per-hour, and two at 60 miles-per-hour (see Table 2). To account for the noise variability, multi-style training and spectral noise subtraction was used.

5.2. Office Telephone (Pasha)

Ten subjects recorded 100 common English names (50 first names and 50 first and last names) in an office environment. Two training and one testing repetition were recorded on the handset of a Panasonic DBS key telephone. Additional testing repetitions were recorded on the speakerphone microphone while the speakerphone was at 50 or 100 centimeters from the talker. Table 3 shows the improvements in accuracy gained with the word modeling in Pasha over the original RC modeling in Alibaba for handset and speakerphone channels. In this experiment, the same front-end (i.e. segmentation by equal phoneme similarity density, multi-style training, speech equalization) was used.

Testing Condition	Word Modeling	
	RC stage (counting high similarity regions)	Pasha (summing squared similarity values)
Handset	94.6	98.3
Speakerphone (50 cm)	86.4	95.2
Speakerphone (100 cm)	82.0	93.0

Table 3: Speaker-dependent recognition accuracy on 100 common English names with automatic endpoints in an office environment.

5.3. Robustness to Different Languages

For each of three languages, two native talkers recorded 100 common names in their language (50 first names and 50 first

and last names) in an office environment. The training and testing conditions were similar to the experiment described in section 5.2. Table 4 shows that a phoneme similarity front-end using only English phoneme models performs fairly well across languages. Further improvements could be achieved. In the case of the porting to a different language, specific phoneme models can easily be built for that language. In the case where the vocabulary itself is mixed, a multi-lingual phoneme set could be used.

Target Language	Handset Recognition Accuracy
Chinese	97.9
French	97.4
German	97.9

Table 4: Speaker-dependent recognition accuracy for three target languages using the Pasha recognizer on 100 common names in an office environment with automatic endpoints.

6. IMPLEMENTATION ISSUES

Both recognition methods have been implemented to run in real-time on small hardware using a TMS320C203 fixed-point DSP processor [6]. Speech endpoint detection, LPC analysis and phoneme similarity computations are all done frame-synchronously. In the Condor system the DTW matching procedure is also done frame-synchronously every 20ms. As Pasha's segmentation depends on the word endpoints, recognition processing is not frame-synchronous. However, as its recognition complexity is small, recognition delay for an approximately 100-word vocabulary is not noticeable.

6.1. Data Compression

Product-type applications are often very much driven by hardware costs which requires the choice of recognition features that remain robust even when compressed and/or quantized. Similarity values have been found to preserve most of their information when quantized to four bits. Furthermore Pasha's word representation has been shown to perform equally well when the information is stored on two bytes per phoneme resulting in 110-byte word templates.

6.2. Phoneme Inventory Size

Our default implementations of Condor and Pasha use 55 phoneme units, corresponding to all TIMIT segments except the closures, which were merged into their neighboring stop burst segments. Based on hardware constraints and application requirements, similarity front-ends can be downscaled to use fewer phoneme units. Experiments with the Pasha recognizer showed an average decrease in recognition accuracy of just 0.6% among all three testing conditions (Table 5) when only using 37 phoneme units. The reduction from 55 to 37 phoneme units represents a positive trade-off since the matching time and memory required for storing the word models is proportional to the phoneme set size.

Testing Condition	37 Phonemes	55 Phonemes
Handset	97.9	98.3
Speakerphone (50 cm)	94.7	95.2
Speakerphone (100 cm)	92.1	93.0

Table 5: Speaker-dependent recognition accuracy using the Pasha recognizer on 100 common English names for different phoneme sets with automatic endpoints.

7. CONCLUSIONS

Phoneme similarities have been shown to be more robust than LPC cepstral parameters to speaker or channel variation. In addition robustness to noise and channel degradation can be further enhanced by inexpensive techniques such as multi-style training and speech equalization.

Multiple phoneme similarities present redundant information and can be the basis for different word representation methods, having different size, complexity and accuracy trade-offs. However the identification (e.g. high similarity regions, top N high similarity values per frame) and extraction (e.g. discrete vs. continuous extraction of high similarity regions as in the RC stage and Pasha respectively) of reliable features is critical when dealing with adverse conditions.

Recognition systems based on phoneme similarities can be implemented in small hardware and achieve high recognition accuracy in a variety of real-world operating conditions.

8. REFERENCES

1. Kimura, T., Endo, M. Hiraoka, S. and Niyada, K. "Speaker Independent Speech Recognition Using Continuous Matching of Parameters in Time-Spectral Form Based on Statistical Measures", *Proc. ICSLP*:169-172, 1992.

2. Hoshimi, M., Miyata, M., Hiraoka, S. and Niyada, K. "Speaker Independent Speech Recognition Method Using Training Speech from a Small Number of Speakers", *IEEE Proc. ICASSP*: 469-472, 1992.

3. Applebaum, T. H., Hanson, B. A. and Morin, P., "Recognition Strategies for Lombard Speech", *Proc. ESCA-NATO Tutorial and Research Workshop on 'Speech Under Stress'*, Lisbon, Portugal, 1995.

4. Morin, P. and Applebaum, T. H., "Word Hypothesizer Based on Reliably Detected Phoneme Similarity Regions", *Proc. Eurospeech*: 897-900, 1995.

5. Applebaum, T. H., Morin, P. and Hanson, B. A., "A Phoneme-Similarity Based ASR Front-End", *IEEE Proc. ICASSP*: 33-36, 1996.

6. Texas Instruments, "TMS320C2xx User's Guide", SPRU127B, 1997.