SCHEMA-BASED MODELING OF PHONEMIC RESTORATION

Soundararajan Srinivasan

Biomedical Engineering Center
The Ohio State University
Columbus, OH 43210, USA
srinivasan.36@osu.edu

DeLiang Wang

Department of Computer and Information
Science & Center for Cognitive Science
The Ohio State University
Columbus, OH 43210, USA
dwang@cis.ohio.state.edu

ABSTRACT

Phonemic restoration refers to the synthesis of missing phonemes in speech when sufficient lexical context is present. Current models for phonemic restoration however, make no use of any lexical knowledge. Such models are inherently inadequate for restoring unvoiced phonemes and may be limited in their ability to restore voiced phonemes too. We present a predominantly top-down model for phonemic restoration. The model uses a missing data speech recognition system to recognize speech utterances as word sequences and activates word templates corresponding to the words containing the masked phonemes. An activated template is dynamically time warped to the noisy word and is then used to restore the speech frames corresponding to the masked phoneme, thereby synthesizing it. The model is able to restore both voiced and unvoiced phonemes. Systematic testing shows that this model performs better than the Kalman-filter based model.

1. INTRODUCTION

Auditory scene analysis [1] brings to the fore the ubiquitous presence of noise in the everyday auditory environment. To listen in these conditions, the auditory system must be robust to such noisy intrusions. Though primitive auditory scene analysis is known to be an innate mechanism for separating speech from interfering sound sources; schema-based stream segregation and grouping supplements the process and sometimes provides the only basis for auditory organization. At the beginning of the 20^{th} century, Bagley [2] reported a series of results, which we now know as phonemic restoration. Phonemic restoration is the perceptual synthesis of missing phonemes when masked by appropriate sounds and when contextual knowledge about the misssing phonemes is available. In 1970, Warren found that when a masking sound replaced the first "s" of the word "legislatures" in the sentence, "The state governors met with their respective legislatures convening in the capital city," listeners had the impression of hearing the phoneme [3]. They were also unable to localize the masking sound within the sentences accurately. Subsequent studies have shown that phonemic restoration depends largely on the linguistic skills of the listeners and the characteristics of the masking sound [4, 5]. Phonemic restoration is a case of auditory induction, a subjective illusion of auditory continuity in noise [6].

Phonemic restoration is a natural way to bring in features of schema-based processing like memory and attention into computational auditory scene analysis (CASA). In speech separation, it may also provide a strong grouping cue for integrating unvoiced consonants. Additionally, phonemic restoration can help restore lost packets in speech transmission systems. Previous attempts to model phonemic restoration have been only partly successful. Cooke and Brown [7] utilize Bregman's four rules for auditory induction [1] as dynamic programming cost functions for restoration. The restoration itself is defined to be a weighted linear interpolation. In its use of temporal continuity for restoration, it is similar to the work of Masuda-Katsuse and Kawahara [8], who use Kalman filtering to predict spectral trajectories in those time-frequency regions that are dominated by noise. The biggest problem for a filtering/inter-polation
system for predicting missing speech segments occurs when temporal continuity of speech frames becomes weak or even absent. This typically occurs with unvoiced speech. In absence of
co-articulation cues, it is impossible to restore the missing portions; in such cases knowledge must be employed. Ellis [9] uses the knowledge of stored acoustic waveforms in the form of a speech recognizer to recognize (hypothesize) the information in the missing regions. The information from the recognizer corresponding to the hypothesis is then projected back to the signal space via feature reconstruction and inverse transformation. Though this idea is promising, few implementation results are presented.

We present a model of phonemic restoration which employs lexical knowledge in the form of a speech recognizer and a sub-lexical representation in word templates performing the role of speech schemas. A hidden Markov model (HMM) based missing data speech recognizer [10] is used to recognize the input sounds as words based on the reliable portions of the speech signal. The word template corresponding to the recognized word is then used to "induce" relevant acoustic signal in the spectro-temporal regions occupied by noise. The templates are formed by averaging (along a dynamic time warped path) tokens of each word.

Section 2 decribes the details of our model. The model has been tested on both voiced and unvoiced phonemes and the test results are presented in section 3. Finally, conclusions and future work is addressed in section 4.

2. MODEL DESCRIPTION

We present a multi-state model for phonemic restoration as shown in Fig. 1. Utterances with words containing masked phonemes are first converted into a spectrogram by Fourier analysis. The missing data ASR recognizes the utterance as word sequences. Word templates corresponding to the noisy words are then chosen based on the results of recognition. A template thus activated, is used to syn-

In *Proc. Eurospeech 2003,* pp. 2053-2056, 2003

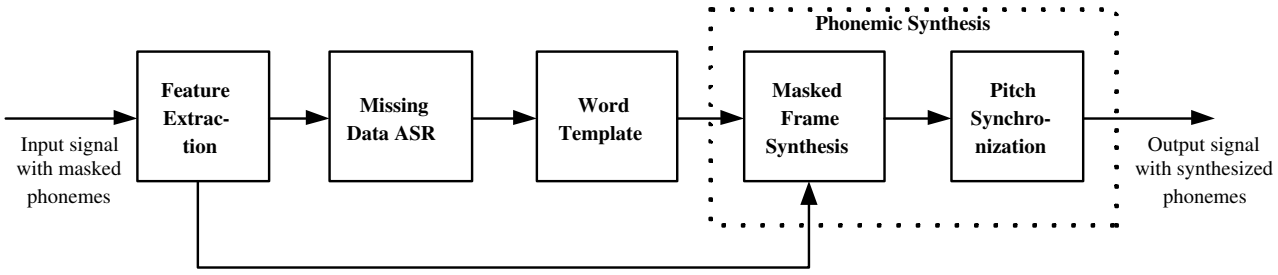

Fig. 1. Block Diagram of the proposed system. The input signal with masked phonmes is converted into the spectral domain and is fed to the missing data ASR, which activates trained word templates corresponding to the words whose phonemes are masked. The masked frames are synthesized by dynamically time warping the templates to the noisy words. These frames are then pitch synchronized with the rest of the utterance.

thesize the frames of the masked phoneme by using dynamic time warping (DTW). These synthesized frames are pitch synchronized to maintain the overall intonational structure of the utterance.

2.1. Feature Extraction

Speech signal with phonemes masked by noise is fed to the feature extraction stage which outputs 513 DFT coefficients, calculated every frame. Each frame is 20ms long with 10 ms overlap between consecutive frames. Frames are extracted by applying a running Hamming window to the signal. The magnitude spectrum thus extracted is converted to the dB scale and is fed to the missing data ASR (see section 2.2) for recognition. Additionally, it is also sent to the synthesis stage (section 2.4) after undergoing diagonalization via the discrete cosine transform [11].

2.2. The Missing Data Speech Recognizer

Traditional ASR systems do not work well in the face of noisy intrusions and other distortions. The missing data ASR [10] makes use of the spectro-temporal redundancy in speech to make optimal decisions about lexical output units. Given a speech observation vector x, the problem of word recognition is to maximize the posterior $P(\omega_i|x)$, where ω_i is a valid word sequence according to the grammar for the recognition task. When parts of x are masked by noise or other distortions, x can be partitioned into its reliable and unreliable constituents as x_r and x_u, where $x = x_r \cup x_u$. One can then seek the Bayesian decision rule given the reliable features. In the marginalization method of [10], the posterior probability using only the reliable features is computed by integrating over the unreliable constituents. Since the feature vector x represents the observed spectral energy and sound sources being additive, the unreliable parts can be constrained as $0 \leq x_u \leq x$. This bounded marginalization method is shown in [10] to have a better recognition score than the simple marginalization method.

We use the 10 state continuous density HMM as suggested by Cooke *et al.* [10]. The task domain is connected digits' recognition. Thirteen (1-9, a silence, very short pause between words, zero and oh) word level models are trained. All except the short pause model have 10 states, whose output distribution is modeled as a mixture of 10 gaussians. The short pause model has only three states. The TIDigits [12] database's male speaker data is used for both training and testing. A HMM toolkit, HTK [13] is used for training and a modified decoder is used for testing.

2.3. Word Template Training by Dynamic Time Warping

Cooke *et al.* [10] suggest that for resotoration, one can use the maximum likelihood estimate of the output distribution of the winning states. Winning states are obtained during recognition by Viterbi decoding of the hidden state sequence. We find that such a restoration is hardly optimal and degrades with increasing number of frames that need to be restored. This is not suprising, the missing data ASR has only 10 states to model each acoustic token and hence state based imputation is an ill-posed one-to-many projection problem.

On the other hand, template based speech recognizers use spectral templates to model each word. These templates could be used as a base fro restoration. We use 100 tokens of isolated word utterances from the training portion of the TIDigits corpus to train each speaker-independent (SI) word template. Assuming all tokens are consistent, we find their cepstral average. For this purpose, these tokens are time normalized by DTW. The distortion measure used in the dynamic programming cost function is the cepstral distance. The local constraint used is the Itakura constrainst [14]. Isolated word utterances corresponding to one test speaker in the test database are used to train a speaker-dependent (SD) template. Utterances of this speaker would be used for testing. Together the two sets of templates form word schemas.

2.4. Phonemic Synthesis

A maximum of 2 phonemes are masked in each utterance of the test speaker by overlaying with white noise. Any transistions into and out of the phoneme are masked too. The phonemes are randomly chosen. Masking yielded a local SNR of -1dB on average. To test the full potential of the proposed model, we use an *a priori* binary mask to differentiate the reliable data from the unreliable ones. The signal and the mask is sent to the missing data ASR which provides the most likely word sequence. Additionally the ASR provides time end points of the recognized words in the signal. We then choose the word templates corresponding to the noisy word and warp them to the speech segment corresponding to the noisy word by DTW. The frames of the template corresponding to the masked frames then replace the masked frames. Our restoration approach is thus knowledge based. To compensate for co-articulation, the imputed frames are manipulated by pitch synchronization techniques (which use interpolated pitch information), PSOLA [15] and LPC-PSOLA [16]. Praat [17] and a local spectral smoother is used for synchronization. The LPC-PSOLA

technique improves the listening experience compared to PSOLA, but is not better than PSOLA as measured by the objective criteria discussed in section 3. Consequently only the results of synchronization using the PSOLA technique is used in the asessement of the results.

3. RESULTS

Informal listening to the restored signal shows that masked voiced and unvoiced phonemes are clearly restored. To measure the performance of the proposed model objectively, two measures are used. The cepstral distance measures the log spectral distance between the original clean signal and the phonemically restored signal:

$$d_C = \sqrt{\left[(C_{1,0} - C_{2,0})^2 + 2\sum_{n=1}^{K}(C_{1,n} - C_{2,n})^2\right]}, \quad (1)$$

where $C_{1,n}$ are the cepstral coefficients derived from AR coefficients of the original signal and $C_{2,n}$ are the corresponding coefficients of the phonemically restored signal. We set $K = 20$. Additionally, the cosh distance [18] between the power spectra of the two signals is computed from (3). Let ps_1 and ps_2 denote the power spectra of the original signal and the phonemically restored signal respectively. The cosh distance is defined as

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left\{cosh\left(\log\left(\frac{ps_1}{ps_2}\right)\right) - 1\right\}d\theta. \quad (2)$$

The distance can be calculated conveniently in its discrete form as

$$\frac{1}{2N}\sum_{n=1}^{N}\left(\frac{ps_1(\omega_n)}{ps_2(\omega_n)} + \frac{ps_2(\omega_n)}{ps_1(\omega_n)} - 2\right). \quad (3)$$

Both measures possess desirable properties of a metric, including symmetry and positive definiteness. The rms log spectra models the log speech spectra very well, but are hard to compute. The cepstral distance and the cosh distance are much easier to compute. Additionally the cepstral distance bounds the rms log spectral distance from below and the cosh distance from above [18].

Three classes of phonemes are considered for restoration: vowels, voiced and unvoiced consonants. The vowels possess strong temporal continuity. The spectral continuity of some voiced consonants, e.g. /l/, changes smoothly but faster than vowels. Unvoiced consonants, especially stops, do not have good temporal continuity [19]. Two isolated word utterances from each of 50 randomly chosen speakers in the training database of the TIDigits corpus are used to train each speaker-independent template. The 2 isolated word utterances (for each word) of the test speaker are used to train each speaker-dependent template. The remaining 55 utterances of the test speaker form the test set.

The results indicate that the model is able to restore all classes of phonemes, with a spectral quality very similar to the original clean signal. Fig. 2 shows the performance of our model as measured by the objective criteria. The results shown are the average of all signals in each class in the test set. If a phoneme is perfectly restored, the distances of the restored signal from the original clean signal are 0. Low values of the distance measures after the restoration of voiced phonemes indicate stet synthesis. The restoration of the unvoiced consonants, especially with the use of speaker-dependent templates, is also good. The data exclude those signals which are incorrectly recognized by the missing data ASR; recognition accuracy is 87.5%. As evident from the figure, the overall performance of the model with speaker independent template is close to that with speaker dependent template. Improved listening experience though is observed with the use of speaker dependent template.

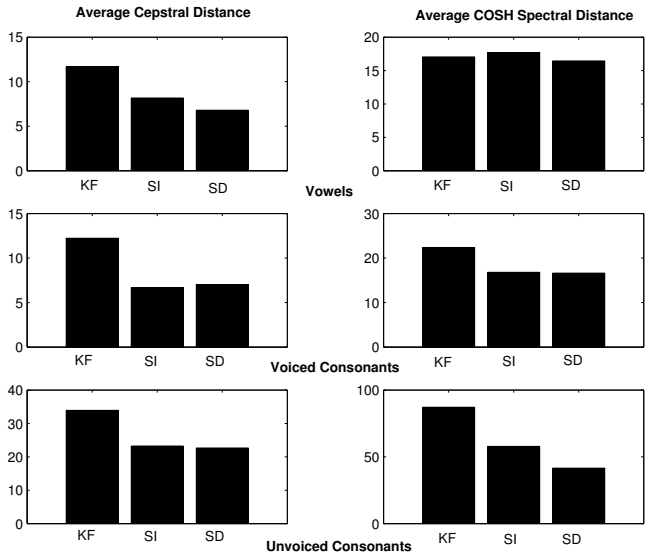

Fig. 2. Performance of the proposed method for phonemic restoration. The above figure shows the performance of the proposed model in restoring the phonemes. SD refers to the performance of our model with speaker dependent templates and SI refers to the performance with speaker independent templates. The top row shows the results corresponding to the restoration of vowels, the middle row the restoration of voiced consonants and the bottom row the restoration of unvoiced consonants. The figures on the left show the average cepstral distance of the restored signal from the original clean signal and the figures on the right show the corresponding average cosh spectral distance. The small distances illustrate good spectral restoration. For comparison, the results of the Kalman filter model (KF) described in Section 4, is also shown . Notice that in restoring unvoiced consonants, our model is substantially better than the Kalman filter model. Restoration of voiced consonants is also better. The performance of our model is similar to that of the Kalman filter in restoring vowels.

4. COMPARISON WITH A KALMAN FILTER MODEL

We compare the performance of our model with the Kalman filter based model of Masuda-Katsuse and Kawahara [8]. They regard cepstral coefficients in each order as a time series that follows a second order auto-regressive (AR) model, which can be predicted and tracked by a Kalman filter. The parameters of the AR model are updated at each frame by a maximum likelihood estimate conditioned on the present and past observed cepstral values. The variance of the noise in the observation model is estimated to be proportional to the reliability of the results from a previous simultaneous grouping process for the speech signal. For the purpose of comparison with our model, we assume prior knowledge of this

variable. This is similar to the assumption of *a priori* mask in our model (see Section 2.4). Additionally, we perform one step backward Kalman smoothing. This improves the performance slightly. Fig. 2 also shows the performance of the Kalman filter for various classes of restored phonemes.

In summary, under both objective criteria discussed in Section 3, our method outperforms the Kalman filtering model significantly. Note that vowels are effectively restored by the Kalman filter. Unvoiced consonants have weak temporal continuity with neighboring phonemes and need knowledge for their restoration. Hence, our method performs substantially better in restoring them. The rapid change in the spectrum causes inaccurate estimation of the AR parameters and hence the tracking by the Kalman filter breaks down. The performance of our method in restoring voiced consonants is also superior to that of the Kalman filter.

5. CONCLUSION

We have presented a top-down based model of phonemic restoration, which performs better than the Kalman filtering model. As stated earlier, the problem for any filtering method is that temporal continuity in speech is not always present. Thus, their performance is best for voiced phonemes (especially vowels) and worst for unvoiced consonants. Hence for phoneme reconstruction, one needs learned schemas. Such schemas represent prior information for restoration.

The model can be used in conjunction with CASA systems to recover masked features and to group unvoiced speech with voiced speech. It can used in restoring lost packets in mobile and internet telephony applications. Our model currently does not consider any bottom-up cues for phonemic restoration. How to generate a binary mask for the missing data recognition also needs to be addressed. Future work would attempt to alleviate both the problems by integrating the model with existing CASA systems (e.g. [20]). Our model is based on recognition and hence fails when recognition fails. Combining recognition with top-down restoration and bottom-up cues should help address this problem.

ACKNOWLEDGMENTS. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027). We thank M.Cooke, H. Kawahara and I. Masuda-Katsuse for their assistance in helping us implement their models.

6. REFERENCES

[1] A. S. Bregman, *Auditory scene analysis*, The MIT Press, Cambridge, MA, 1990.

[2] W. C. Bagley, "The apperception of the spoken sentence: A study in the psychology of language," *American Journal of Psychology*, vol. 12, pp. 80–130, 1900-1901.

[3] R. M. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, pp. 392–393, 1970.

[4] R. M. Warren and G.L. Sherman, "Phonemic restorations based on subsequent context," *Perception and Psychophysics*, vol. 16, pp. 150–156, 1974.

[5] A. G. Samuel, "The role of bottom-up confirmatation in the phonemic restoration illusion," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 7, pp. 1124–1131, 1981.

[6] R. M. Warren, C. J. Obusek, and J. M. Ackroff, "Auditory induction: Perception synthesis of absent sounds," *Science*, vol. 176, pp. 1149–1151, 1972.

[7] M. P. Cooke and G. J. Brown, "Computational auditory scence analysis: Exploiting principles of perceived continuity," *Speech Communication*, vol. 13, pp. 391–399, 1993.

[8] I. Masuda-Katsuse and H. Kawahara, "Dynamic sound stream formation based on continuity of spectral change," *Speech Communication*, vol. 27, pp. 235–259, 1999.

[9] D. P. W. Ellis, "Using knowledge to orgnaize sound: The prediction-driven approach to computational auditory scene analysis, and its application to speech/non-speech mixtures," *Speech Communication*, vol. 27, pp. 281–298, 1999.

[10] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[11] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*, Prentice-Hall, Inc., Upper Saddle River, NJ, 2nd edition, 1999.

[12] R. G. Leonard, "A database for speaker-independent digit recognition," *Proc. ICASSP '84*, pp. 111–114, 1984.

[13] S. Young, D. Kershaw, J. Odell, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0)*, Microsoft Corporation, 2000.

[14] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 2nd edition, 1999.

[15] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[16] E. Moulines and F. Charpentier, "Diphone synthesis using a mulitpulse lpc technique," *Proc. of the FASE Int. Conf.,Edinburgh*, pp. 47–55, 1988.

[17] P. Boersma and D. Weenink, "Praat: doing Phonetics by Computer, Version 4.0.26," *http://www.fon.hum.uva.nl/praat*, 2002.

[18] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 5, pp. 380–391, 1976.

[19] K. N. Stevens, *Acoustic phonetics*, The MIT Press, Cambridge, MA, 1998.

[20] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.