A VECTOR TAYLOR SERIES APPROACH FOR ENVIRONMENT-INDEPENDENT SPEECH RECOGNITION

Pedro J, Moreno, Bhiksha Raj and Richad M. Stern
Department of Electrical and Computer Engineering
&School of Computer Science
Carnegie Mellon University

Presented by
Howard

A MODEL OF THE ENVIRONMENT (2-1)

$$\mathrm{Z}(\omega) = X(\omega)\,|H(\omega)|^2 + N(\omega)$$

- Where $\mathrm{Z}(\omega)$ represents the power spectrum of the degraded speech, $\mathrm{X}(\omega)$ is the power spectrum of the clean speech, $H(\omega)$ is the transfer function of the linear filter, and $N(\omega)$ is the power spectrum of the additive noise.
- In the log-Spectral domain this relation can be expressed as:

$$z = x + q + \log(1 + e^{n-x-q})$$

of in more general term:

$$z = x + f(x, n, q)$$

where is an unknown parameter that represents the effects of linear filtering in the log-spectra domain.

- We also assume that the PDF of the log-spectra of the speech signal can be well represented by a summation of multivariate Gaussian distributions:

$$p(x) = \sum_{k=0}^{M-1} P[k] N_x(\mu_{x,k}, \Sigma_{x,k})$$

A MODEL OF THE ENVIRONMENT (2-2)

- Furthermore, we assume that the statistics of noise can be well represented by a single Gaussian $N_n(\mu_n, \Sigma_n)$.
- The problem of compensation is two fold. First, the parameters q, , μ_n ,and Σ_n need to be determined. Second, the distribution of Z given the PDF of χ and the parameters q, μ_n, and Σ_n has to be computed. Because of the non-linearity of the function $f(n, x, q)$, both problems are non-trivial. Only for very simple expressions of the function $f(n, x, q)$can $p(z)$ be computed analytically.

  But function like $\log(1+e^{n-x-q})$ is not possible to compute $p(z)$ analytically.
- While $p(z)$ could be computed by Monte-carlo methods, this approach is computationally expensive and requires previous knowledge of the parameters q ,μ_n and $\sum_n$. VTS provides a framework that enables an analytical solution to both problems.

DESCRIPTION OF THE VTS ALGORITHMS

- The key of the new VTS algorithm is to approximate the generic vector function $f(n, x, q)$ with a vector Taylor series approximation:

$$f(x,n,q) \cong f(x_0,n_0,q_0) + \frac{d}{dx} f(x_0,n_0,q_0)\{x - x_0\} + \frac{d}{dn} f(x_0,n_0,q_0)\{n - n_0\} + \frac{d}{dq} f(x_0,n_0,q_0)\{q - q_0\} + \cdots$$

where $f(x_0, n_0, q_0)$ is the vector function evaluated at a particular vector point.

Similarly, $\frac{d}{dx} f(x_0, n_0, q_0)$ represents the matrix derivative of the vector function at a particular vector point.

- The Taylor expansion is exact everywhere when the order of the Taylor series is infinite. However, when x has a Gaussian distribution, the function can be expanded around the mean of and the expansion needs to be good only within a relatively narrow region around the mean. We take advantage of this fact to truncate the Taylor series after just a few terms.

HMM ADAPTATION USING A PHASE-SENSITIVE ACOUSTIC DISTORTION MODEL FOR ENVIRONMENT-ROBUST SPEECH REOGNITION

Jinyu Li, Li Deng, Dong Yu, Yifan Gong, and Alex Acero
Microsoft Corporation, One Microsoft Way, Redmond

Presented by Howard

Introduction(2-1)

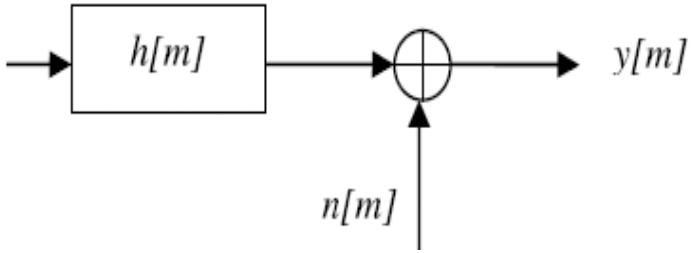
- In recent years, a popular approach to joint compensation of additive convolution distortions (JAC) in the model domain has been proposed.
- Common among these studies is the use of vector Taylor series (VTS) approximation to linearize the model for closed form HMM adaptation formulas and for noise/channel parameter estimation
- All of the JAC/VTS work for HMM adaptation, the environment-distortion model makes the assumption of instantaneous phase synchrony (phase-insensitive) between the clean and the mixing noise.
- This assumption is relaxed in the work reported in “Enhancement of logspectra of speech using phase-sensitive model of acoustic environment”, where a new phase term was introduced to account for the random nature of the phase asynchrony

Introduction (2-2)

- The JAC/VTS approach implements in model-domain, the phase-sensitive implements in feature-domain.
- The JAC/VTC gets a better recognition result.
- The research in this paper extends and integrates these two set of work.
- The new algorithm implements environment robustness via HMM adaptation taking into account phase asynchrony between clean and mixing noise.

Phase-JAC/VTC adaptation algorithm

- With DFT the follc relations can be e frequency domair



θ_k

$$Y[k] = X[k]\,H[k] + N[k]$$

- where k is the frequency-bin index in DFT given a fixed-length time window.
- The power spectrum of the distorted speech can then be obtained as:

$$|Y[k]|^2 = |X[k]|^2\,|H[k]|^2 + |N[k]|^2 + 2|X[k]|\,|H[k]|\,|N[k]|\cos\theta_k$$

- Where denotes the (random) angle between the tow complex variables $N[k]$ and $(X[k]H[k])$

Algorithm for HMM Adaptation Given the Joint Noise and Channel Estimates(4-1)

- By applying a set of Mel-scale filters (L in total) to the power spectrum, we have the l-th Mel filter-bank energies for distorted speech, clean speech, noise and channel:

$$|\tilde{Y}^{(l)}|^2 = \sum_k W_k^{(l)} |Y[k]|^2$$

$$|\tilde{X}^{(l)}|^2 = \sum_k W_k^{(l)} |X[k]|^2$$

$$|\tilde{N}^{(l)}|^2 = \sum_k W_k^{(l)} |N[k]|^2$$

$$|\tilde{H}^{(l)}|^2 = \frac{\sum_k W_k^{(l)} |X[k]|^2 |H[k]|^2}{|\tilde{X}^{(l)}|^2}$$

Where the l-th filter is characterized by the transfer function $W_k^{(l)} \geq 0 (\sum_k W_k^{(l)} = 1)$

The phase factor $\alpha^{(l)}$ of the l-th Mel filter-bank:

$$\alpha^{(l)} = \frac{\sum_k W_k^{(l)} |X[k]||H[k]||N[k]| \cos\theta_k}{|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|}$$

Algorithm for HMM Adaptation Given the Joint Noise and Channel Estimates(4-2)

- Then, the following relation is obtained in the Mel filter-bank domain for the l-th Mel filter-bank output: $|\tilde{Y}^{(l)}|^2 = |\tilde{X}^{(l)}|^2 |\tilde{H}^{(l)}|^2 + |\tilde{N}^{(l)}|^2 + 2\tilde{\alpha}^{(l)}|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|.$
- The phase-factor vector for all the L Mel filter-banks is defined as:

$$\alpha = \left[\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(l)}, \ldots \alpha^{(L)}\right]^T$$

- By taking logarithm and multiplying the non-square discrete cosine transform (DCT) matrix C to both sides of the form above for all the L Mel filter-banks , the following nonlinaer distortion model is obtained in cepstral domain:

$$y = x + h + C\log(1 + \exp(C^{-1}(n-x-h)) + 2\alpha \bullet \exp(C^{-1}(n-x-h)/2))$$

$$= x + h + g_\alpha(x,h,n)$$

where $g_\alpha(x,h,n) = C\log(1 + \exp(C^{-1}(n-x-h)) + 2\alpha \bullet \exp(C^{-1}(n-x-h)/2))$

C^{-1} is the (pseudo) invese DCT matrix. y, x, n and h are the vector-valued distorted specch, clean speech, noise and channel respectively, all in the MFCC domain.

Algorithm for HMM Adaptation Given the Joint Noise and Channel Estimates(4-3)

- Using the first-order VTS approximation with respect to x, n and h, we have

$$y = \mu_x + \mu_h + g(\mu_x, \mu_h, \mu_n) + G(x - \mu_x) + G(h - \mu_h) + (I - G)(n - \mu_n) \qquad (13)$$

where $\left.\frac{\partial y}{\partial x}\right|_{\mu_x, \mu_n, \mu_h} = \left.\frac{\partial y}{\partial h}\right|_{\mu_x, \mu_n, \mu_h} = G$, $\frac{\partial y}{\partial n} = I - G$

$$G = I - C \cdot diag\left(\frac{\exp(C^{-1}(\mu_n - \mu_x - \mu_h)) + \alpha \bullet \exp(C^{-1}(\mu_n - \mu_x - \mu_h)/2)}{1 + \exp(C^{-1}(\mu_n - \mu_x - \mu_h)) + 2\alpha \bullet \exp(C^{-1}(\mu_n - \mu_x - \mu_h)/2)}\right) \cdot C^{-1}$$

$diag(.)$ stands for the diagonal matrix with its diagonal component value equal to the value of the vector in the argument.

- For the given noise mean vector μ_n and channel mean vector μ_h , the value of $G(.)$ depends on mean vector μ_x . Specifically, for the k-th Gaussian in the j-th state, the element of $G(.)$ matrix becomes:

$$G_\alpha(j,k) = I - C \cdot diag\left(\frac{\exp(C^{-1}(\mu_n - \mu_{x,jk} - \mu_h)) + \alpha \bullet \exp(C^{-1}(\mu_n - \mu_{x,jk} - \mu_h)/2)}{1 + \exp(C^{-1}(\mu_n - \mu_{x,jk} - \mu_h)) + 2\alpha \bullet \exp(C^{-1}(\mu_n - \mu_{x,jk} - \mu_h)/2)}\right) \cdot C^{-1}$$

- Then, the Gaussian mean vectors (the k-th Gaussian in the j-th state) in the adapted HMM for the degraded speech can be obtained by taking expectation of both side of Eq. (13):

$$\mu_{y,jk,\alpha} \approx \mu_{x,jk} + \mu_h + g_\alpha(\mu_{x,jk}, \mu_h, \mu_n)$$

which is applied only to the static portion of the MFCC vector.

Algorithm for HMM Adaptation Given the Joint Noise and Channel Estimates(4-4)

- The covariance matrix $\Sigma_{y,jk,\alpha}$ in the adapted HMM can be estimated as a weighted sum of $\Sigma_{x,jk}$, the covariance matrix of the clean HMM, and Σ_n, the covariance matrix of noise, by taking variance "operation" on both sides of Eq. (13):

$$\Sigma_{y,jk,\alpha} \approx G_\alpha(j,k)\Sigma_{x,jk}G(j,k)^T + (I - G_\alpha(j,k))\Sigma_n(I - G(j,k))^T$$

$$\mu_{\Delta y,jk,\alpha} \approx G_\alpha(j,k)\mu_{\Delta x,jk} + (I - G_\alpha(j,k))\mu_{\Delta n},$$

$$\mu_{\Delta\Delta y,jk,\alpha} \approx G_\alpha(j,k)\mu_{\Delta\Delta x,jk} + (I - G_\alpha(j,k))\mu_{\Delta\Delta n},$$

$$\Sigma_{\Delta y,jk,\alpha} \approx G_\alpha(j,k)\Sigma_{\Delta x,jk}G_\alpha(j,k)^T + (I - G_\alpha(j,k))\Sigma_{\Delta n}(I - G_\alpha(j,k))^T$$

$$\Sigma_{\Delta\Delta y,jk,\alpha} \approx G_\alpha(j,k)\Sigma_{\Delta\Delta x,jk}G_\alpha(j,k)^T + (I - G_\alpha(j,k))\Sigma_{\Delta\Delta n}(I - G_\alpha(j,k))^T$$

Algorithm for Re-estimation of Noise and Channel

$$\mu_h = \mu_{h,0} + \left\{ \sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) G_\alpha(j,k)^T \Sigma_{y,jk,\alpha}^{-1} G_\alpha(j,k) \right\}^{-1} \cdot$$

$$\left\{ \sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) G_\alpha(j,k)^T \Sigma_{y,jk,\alpha}^{-1} \left[y_t - \mu_{x,jk} - \mu_{h,0} - g_\alpha(\mu_{x,jk}, \mu_{h,0}, \mu_{n,0}) \right] \right\}$$

$$\mu_n = \mu_{n,0} + \left\{ \sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{y,jk,\alpha}^{-1} (I - G_\alpha(j,k)) \right\}^{-1} \cdot$$

$$\left\{ \sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{y,jk,\alpha}^{-1} \left[y_t - \mu_{x,jk} - \mu_{h,0} - g_\alpha(\mu_{x,jk}, \mu_{h,0}, \mu_{n,0}) \right] \right\}$$

$$\mu_{\Delta n} = \mu_{\Delta n,0} + \left\{ \sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{\Delta y,jk,\alpha}^{-1} (I - G_\alpha(j,k)) \right\}^{-1} \cdot$$

$$\left\{ \sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{\Delta y,jk,\alpha}^{-1} \left[\Delta y_t - G_\alpha \mu_{\Delta x,jk} - (I - G_\alpha(j,k)) \mu_{\Delta n,0} \right] \right\}$$

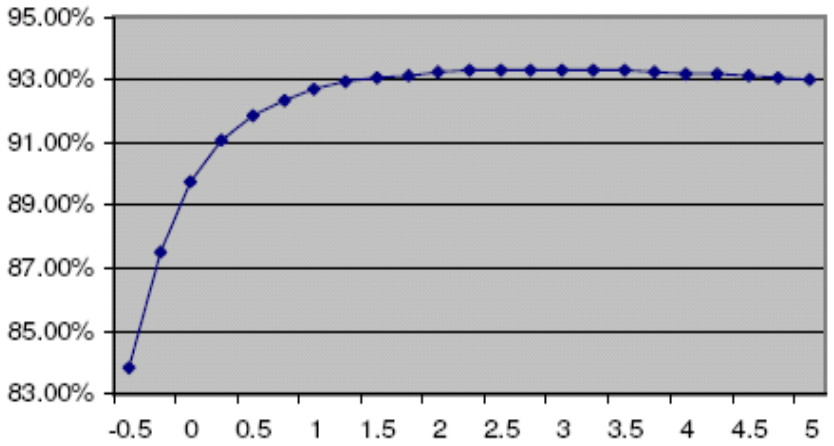
$$\mu_{\Delta\Delta n} = \mu_{\Delta n,0} + \left\{ \sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{\Delta\Delta y,jk,\alpha}^{-1} (I - G_\alpha(j,k)) \right\}^{-1} \cdot$$

$$\left\{ \sum_t \sum_{j\in\Omega_s} \sum_{k\in\Omega_m} \gamma_t(j,k) (I - G_\alpha(j,k))^T \Sigma_{\Delta\Delta y,jk,\alpha}^{-1} \left[\Delta\Delta y_t - G_\alpha \mu_{\Delta\Delta x,jk} - (I - G_\alpha(j,k)) \mu_{\Delta\Delta n,0} \right] \right\}$$

Algorithm steps

1. Read in s distorted speech utterance
2. Set the channel mean vector to all zeros
3. Initialize the noise mean vector and diagonal covariance matrix using the first and last N frames from the utterance using sample estimates
4. Adapt the HMM parameters
5. Decode the utterance with the adapted HMM parameters
6. re-estimate all the noise and channel parameters
7. Adapt the HMM parameters
8. Use the final adapted model to obtain the utterance ouput transcription
9. Goto step1.

Experiment



ACOUSTIC MODELING BY PHONEME TEMPLATES AND MODIFIED NOE-PASS DP DECODING FOR CONTINUOUS SPEECH RECOGNITION

V.ramasubramanian, Kaustubh Kulkarni, Bernhard Kaemmerer
Professional Speech Processing – India, Siemens Corporate Technology, Bangalore, India

Presented by
Howard

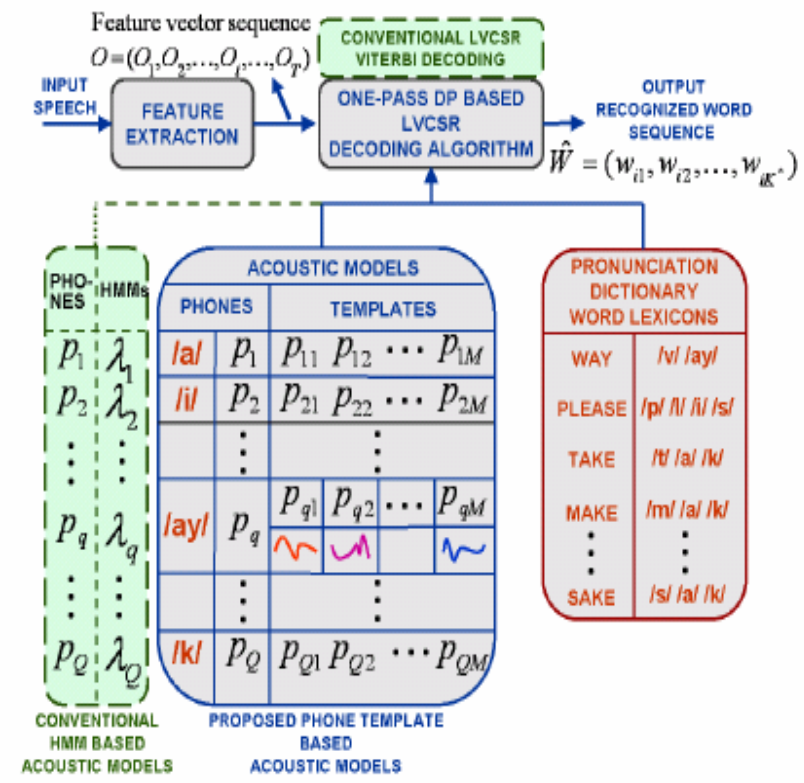
Introduction(2-1)

- Current continuous speech recognition is largely based on HMM based acoustic modeling of phones and triphones. However, various shortcomings of HMMs have long been felt now, mainly with respect to its inability to account for inter-frame correlations and the difficulties in reliably estimating very large number of context-dependent HMM parameters from limited training data.
- In this paper, the main approach is to use a template database of continuous speech which is annotated phonetically with various acoustic as well as non-verbal attributes.
- Proposing acoustic modeling by use of multiple templates of a monophone or triphones drawn from training data.

Introduction (2-2)

- HMM phone model is replaced with a set of multiple templates.
- During decoding, this system uses a token passing strategy to search this continuous database
- Decoding is done by a modified one-pass dynamic programming algorithm requiring more complex recursions when compared to the conventional one-pass DP algorithm used for connected word recognition.
- The pronunciation dictionary (word lexicon) is specified as a linear baseform of phones triphones as in conventional CSR.
- This algorithm conforms to the basic definition of CSR and dose not give any particular emphasis to the natural ordering of the templates in the training data

Proposed phoneme template modeling



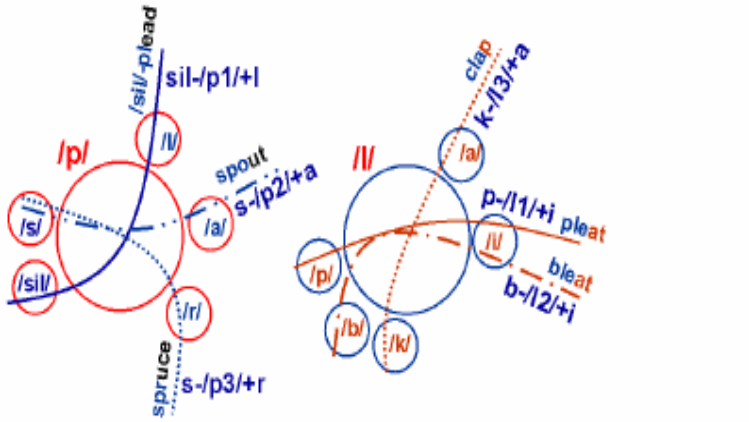
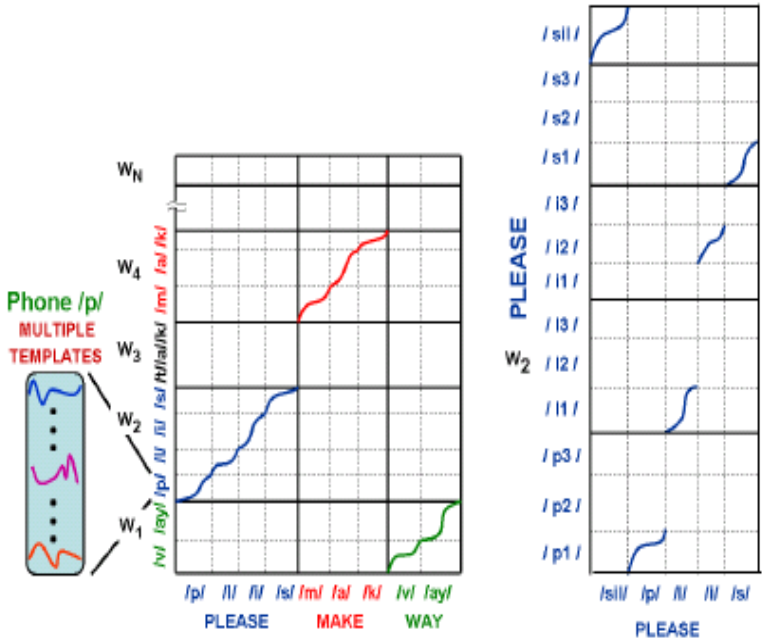


Fig. 2. *Multiple phoneme template trajectories of /p/ and /l/*

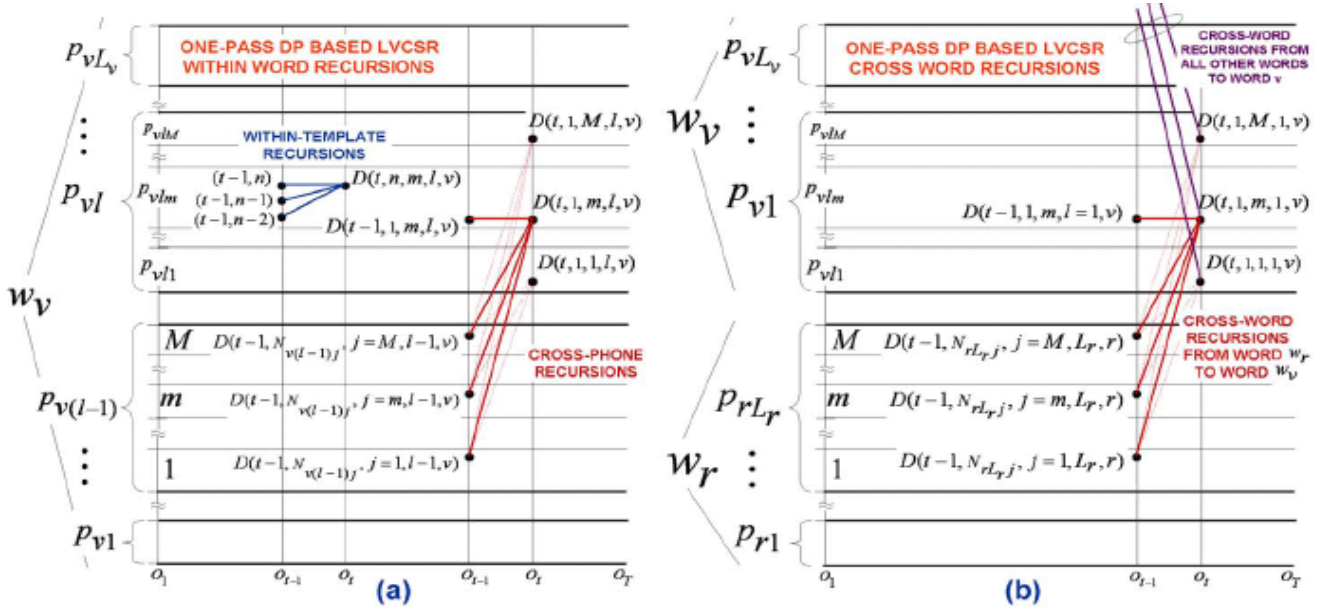
Table 1. *Phoneme templates of /i/ and /s/ with triphone contexts*

Phone	From word	Phone	From word
$n-\mathbf{i}_1+d$	*need*	$i-\mathbf{s}_1+m$	*prism*
$l-\mathbf{i}_2+s$	*lease*	$a-\mathbf{s}_2+p$	*clasp*
$r-\mathbf{i}_3+z$	*freeze*	$i-\mathbf{s}_3+sil$	*grease-sil*

Proposed decoding algorithm



Proposed one-pass DP based CSR



- Within-word recursions

  a. within-phoneme-template recursion

$$D(t,n,m,l,v) = d(t,n,m,l,v) + \min_{j=(n,n-1,n-2)\&(j>0)} [D(t-1,j,m,l,v)]$$

  b. cross-phone recursion

$$D(t,n=1,m,l,v) = d(t,n=1,m,l,v) + \min[D(t-1,n=1,m,l,v), \min_{j=1,\ldots,M} [D(t-1,N_{v(l-1)j},j,l-1,v)]]$$

- Cross-word transitions

$$D(t,n=1,m,l,v) = d(t,n=1,m,l,v) + \min[D(t-1,n=1,m,l,v), \min_{j=1,\ldots,M} [D(t-1,N_{v(l-1)j},j,l-1,v)]]$$

- Termination and backtracking

$$D^* = \min_{v=1,\ldots,V} \min_{m=1,\ldots,M} D(T,N_{vL_vm},m,L_v,v)$$

Experiments

- The experiments done here are primarily intended to bring out the acoustic modeling efficacy of phoneme templates in various contextual settings, rather than on the largeness of the continuous speech recognition tasks or the use of language models, efficient search et.

Table 2. *Experiments (1-4) & their phoneme template definitions*

Template Type	Templates drawn from	
	Outside (sa1,sa2)	Inside (sa1, sa2)
Monophone	1. Mono-non-sa1-sa2	3. Mono-in-sa1-sa2
Triphone	2. Tri-non-sa1-sa2	4. Tri-in-sa1-sa2