A LINK BETWEEN CEPSTRAL SHRINKING AND THE WEIGHTED PRODUCT RULE IN AUDIO-VISUAL SPEECH RECOGNITION

Simon Lucey, Sridha Sridharan and Vinod Chandran

Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
slucey@ieee.org, s.sridharan@qut.edu.au, v.chandran@qut.edu.au

ABSTRACT

The weighted product rule has been shown empirically to be of great benefit in audio-visual speech recognition (AVSR), for isolated word recognition tasks. A firm theoretical basis for the selection of effective weights is of considerable interest to the audio-visual speech processing community. In this paper a clear link is established between the selection of effective weightings and the approximately isotropic shrinkage that the distribution of acoustic cepstral features undergo in the presence of additive noise. An elucidation of the theoretical relationship between the cepstral shrinkage and the variance of the HMM audio log-likelihoods is then explored.

1. INTRODUCTION

The effective integration of the acoustic and visual modalities, for the task of isolated word recognition, is of particular interest to the audio-visual speech community. A late integration strategy, of integrating the confidence scores of the *independently* trained acoustic and visual classifiers, has been shown [6] empirically to work well for isolated word recognition. The problem of how to effectively combine classifiers, representing the acoustic and visual speech modalities, is the topic of ongoing research.

The effective combination of an ensemble of classifiers is a topic of particular importance to the pattern recognition community. There is currently empirical evidence [1] that using an ensemble of classifiers for a recognition problem can give superior performance over those classifiers individually under favorable circumstances. Care must be taken however, as some combinations of classifiers can performs worse than those classifiers individually an effect known as *catastrophic fusion*.

In this paper, the benefits of the weighted product rule, specifically for the task of combining classifiers trained on acoustic and visual speech, are elucidated from a theoretical perspective. A firm link between the selection of effective weightings and the approximately isotropic shrinkage that the distribution of acoustic cepstral features undergo in the presence of additive noise is made. From this link insights are gained into how the weighted product rule lessens the compounding effects of classifier confidence errors, upon combination.

2. AUDIO-VISUAL DATABASE AND MODELLING

The M2VTS database [2] was used for experiments in this paper. It consisted of, 37 subjects (male and female) speaking four repetitions (shots) of ten French digits from *zero* to *nine*. The mouth region of interest (ROI) chosen was based on the subject's eye separation distance d_{eye}, with the ROI being defined as an $(3d_{eye}) \times (4d_{eye})$ box positioned at at the mouth center, for every frame of each video sequence.

Visual features were extracted from the mouth ROI through a combination of linear discriminant analysis (LDA) and principal component analysis (PCA) [3]. PCA was first used to gain a subspace that preserved the 50 highest modes of variation in the mouth ROI, this was done to remove low energy noise that may otherwise affect the ability of LDA to create an effective discriminant subspace. LDA was then employed to create a discriminant nine dimensional subspace, based on the ten word classes. Shots 1-3 of the M2VTS database were used for the generation of the subspaces. To remove unwanted subject variation, the mean of each visual sequence was removed before calculating the discriminant subspace. For acoustic features we used standard HTK [4] mel-frequency cepstral coefficients (MFCC) with cepstral mean subtraction. Delta features were appended to both acoustic and visual features. Acoustic features were sampled every 10ms while visual features were sampled at 40ms intervals.

Separate hidden Markov models (HMM) were used to model the acoustic and visual utterances using HTK ver 2.2 [4]. For the acoustic and visual modalities, an utterance was modelled using a 3 state, left to right, HMM with 3 mixtures per state and diagonal covariance matrices. The likelihood scores $p(\mathbf{O}|\boldsymbol{\lambda}_i)$ from each HMM $\boldsymbol{\lambda}_i$ were used to gain the a posteriori probability estimates, assuming equal priors, using Bayes rule,

$$\hat{Pr}(\omega_i|\mathbf{O}) = \frac{p(\mathbf{O}|\boldsymbol{\lambda}_i)}{\sum_{n=1}^{N} p(\mathbf{O}|\boldsymbol{\lambda}_i)} \quad (1)$$

Shots 1-3 of the M2VTS database was used for training the HMMs with shot 4 being used for testing.

3. THE WEIGHTED PRODUCT RULE

Excellent results, in AVSR have been received through integrating the confidence scores received from the acoustic

and visual classifiers via the weighted product rule. The weighted product rule can be expressed as,

$$\zeta(\omega_i|\mathbf{O}^{\{av\}}) = \hat{Pr}(\omega_i|\mathbf{O}^{\{a\}})^{\alpha} \times \hat{Pr}(\omega_i|\mathbf{O}^{\{v\}})^{(1-\alpha)} \quad (2)$$

where $\hat{Pr}(\omega_i|\mathbf{O}^{\{m\}})$ is the a posteriori *estimate* of utterance $\mathbf{O}^{\{m\}}$ coming from word class ω_i for modality $\{m = a$ or $v\}$. It must be emphasised that $\zeta(\omega_i|\mathbf{o})$ is a confidence score (not necessarily between zero and one), *not* a probability, but is equivalent to the audio-visual a posteriori probability estimate $\hat{Pr}(\omega_i|\mathbf{O}^{\{av\}})$ in terms of the class decision boundaries it realises.

Bayesian theory dictates [1] that the weighted product rule should be optimal when $\alpha = 0.5$ (i.e. normal product rule), if one is combining *error free* a posteriori class probabilities. In practice however, one can rarely use this weighting due to the differing decision boundaries realised from the *mismatch* between train and test utterances. This mismatch results in a confidence error,

$$\hat{Pr}(\omega_i|\mathbf{O}^{\{m\}}) = Pr(\omega_i|\mathbf{O}^{\{m\}}) + \epsilon_i(\mathbf{O}^{\{m\}}) \quad (3)$$

In practice one can only ever apply the weighted product rule to the a posteriori probability estimates, where the compounding effect of these confidence errors must be taken into account when selecting a suitable combination strategy.

There are two options available to us to try and lessen the compounding effect of these confidence errors. Ideally, one can try and adapt the classifier to the test utterances, thus removing the confidence error and allowing for optimal combination through the normal product rule. Generally, this is impossible in practice as it requires a violation of causality (i.e. access to the test utterances before testing). Alternatively, one can try and dampen the effects of these confidence errors, upon combination, through the judicious choice of combination strategies. This approach has a lot more appeal in practice, as it can be implemented without violating causality. For the specific task of AVSR we shall show that the weighted product rule can act in both an adapting and dampening capacity, whilst not violating causality.

3.1. Calculating a suitable α

The weighting factor α is used to lessen the impact of confidence errors introduced as a result of train/test mismatches. The formulation of the α weighting factor is given by,

$$\alpha = \frac{\beta_a}{\beta_a + \beta_v} \quad (4)$$

where the β_a and β_v values reflect the relative train/test mismatch in the acoustic and visual classifiers respectively. The β_a and β_v values may act in an adapting or dampening capacity depending on the type of mismatch. Both β_a and β_v values lie between zero and one, with an β value of one signifying there is no train/test mismatch in that modality.

Irrespective of what capacity the α factor is acting in, an optimal weighting factor α^* can be found through an exhaustive search of values between zero and one, to improve overall recognition in the presence of a train/test mismatch. Again, this type of approach for finding the optimal α^* is of limited use in a practical AVSP system as it requires a violation of causality. Causal estimates of α^* can however be made for differing acoustic noise conditions based on a qualitative knowledge of how additive noise affects the likelihood scores of the acoustic classifier.

4. ADAPTATION THROUGH THE WEIGHTED PRODUCT RULE

In general other combination strategies, than the weighted product rule, have been shown to be of benefit in independent classifier combination (eg. sum, median and majority vote rules [1]), as they are less sensitive to the compounding effects of confidence errors upon combination [1]. However, in particular instances the weighted product rule can have superior performance to other rules when a mismatch is encountered due to the weighted product rule's ability to *adapt* to the changed test set. For example, in a high dimensional (D) observation space one can define a multi-class (N) set of Gaussian likelihood functions that have class separability due to their class covariance difference not mean difference. Assuming equal priors, one can define the a posteriori probabilities estimates as,

$$\hat{Pr}(\omega_i|\mathbf{o}) = \frac{\mathcal{N}(\mathbf{0}, \sigma^2_{trn}\mathbf{\Sigma}_i)|_{\mathbf{o}}}{\sum_{n=1}^{N}\mathcal{N}(\mathbf{0}, \sigma^2_{trn}\mathbf{\Sigma}_n)|_{\mathbf{o}}} \quad (5)$$

such that all classes have zero mean but different covariance matrices $\mathbf{\Sigma}_i$. One could place an additional constraints that,

$$\frac{1}{D}tr(\mathbf{\Sigma}_i) = 1, \forall i \quad (6)$$

and,

$$det(\mathbf{\Sigma}_i) = det(\mathbf{\Sigma}_j), \forall i, j \quad (7)$$

so as to ensure all class distinction comes from the orientation *not* the homoscedastic variance σ^2_{trn} of the covariance matrices. An observation set is said to be *homoscedastic* if all eigenvalues λ_d describing the Gaussian distribution of the observation set have the same magnitude such that,

$$\sigma^2 = \lambda_d, \forall d \quad (8)$$

A heteroscedastic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ (i.e. a distribution whose eigenvalues have different magnitudes) can be approximated by a homoscedastic distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ in a maximum likelihood sense by,

$$\sigma^2 = \frac{1}{D}\sum_{d=1}^{D}\lambda_d \quad (9)$$

the variance σ^2 calculated from such an approximation is referred to as the homoscedastic variance. The homoscedastic variance of a Gaussian distribution has a clear interpretation as the average variance of the distribution. If a train/test mismatch occurs that changes the test set homoscedastic variance to σ^2_{tst}, a confidence error shall occur when using the a posteriori probability estimates gained from the test set. The confidence error free a posteriori probabilities will be,

$$Pr(\omega_i|\mathbf{o}) = \frac{\mathcal{N}(\mathbf{0}, \sigma^2_{tst}\mathbf{\Sigma}_i)|_{\mathbf{o}}}{\sum_{n=1}^{N}\mathcal{N}(\mathbf{0}, \sigma^2_{tst}\mathbf{\Sigma}_n)|_{\mathbf{o}}} \quad (10)$$

This error free a posteriori probability can be placed in terms of the estimated a posteriori probabilities and an exponential weighting,

$$\begin{aligned} Pr(\omega_i|\mathbf{o}) &= \frac{\hat{Pr}(\omega_i|\mathbf{o})^\beta}{\sum_{n=1}^{N}\hat{Pr}(\omega_i|\mathbf{o})^\beta} \\ &= \frac{\left[\mathcal{N}(\mathbf{0},\sigma_{trn}^2\mathbf{\Sigma}_i)|_{\mathbf{o}}\right]^\beta}{\sum_{n=1}^{N}\left[\mathcal{N}(\mathbf{0},\sigma_{trn}^2\mathbf{\Sigma}_n)|_{\mathbf{o}}\right]^\beta} \\ &= \frac{\exp(-\frac{\beta}{2\sigma_{trn}^2}\mathbf{o'\Sigma_i^{-1}o})}{\sum_{n=1}^{N}\exp(-\frac{\beta}{2\sigma_{trn}^2}\mathbf{o'\Sigma_n^{-1}o})} \end{aligned} \quad (11)$$

where,

$$\beta = \frac{\sigma_{trn}^2}{\sigma_{tst}^2} \quad (12)$$

The form given in Equation 11 can easily be applied to the weighted product rule described in Equation 2 for improved classifier combination performance. It must be emphasised that the use of the weighted product rule is not just dampening the compounding effects of confidence errors in the acoustic and visual modalities. It is also functioning as a peculiar form of classifier adaptation, primarily on the acoustic classifier.

Finally it can be shown that the the variance of the log-likelihoods, if $\mathbf{\Sigma} = E\{\mathbf{\Sigma}_i\}$, will be,

$$Var\{\log\left[\mathcal{N}(\mathbf{0},\sigma_{tst}^2\mathbf{\Sigma})|_{\mathbf{o}}\right]\} = \frac{D}{2}\frac{\sigma_{tst}^4}{\sigma_{trn}^4} \quad (13)$$

At first glance it may seem that such a train/test mismatch is very unlikely in a practical scenario and is not generalised enough in nature to be worthy of much discussion. However, in acoustic speech processing it has been reported [5] that the homoscedastic variance of the observation space shrinks when additive noise is employed during the extraction of cepstral based features as commonly used in most acoustic speech recognition applications. Admittedly for this assumption to be effective the MFCC observation space of the entire M2VTS database has to be described by a homoscedastic Gaussian distribution. In reality one knows that the true distribution of the MFCC observation space is usually not described so simplistically. However, it has been shown [6, 7] that the observation space of many speech applications employing an MFCC representation can be adequately described by a mixture of Gaussians. With this in mind, such a simplistic approximation may be of benefit when combining classifier outputs in speech applications.

5. CAUSAL ADAPTATION

For most AVSR applications additive acoustic noise is the most common form of *varying* train/test mismatch, making the acoustic modality's β_a mismatch value of greater significance than the visual modality's β_v mismatch value, which is normally fixed and typically acts in a dampening capacity, due to the nature of most mismatches in the visual modality. Looking at Equations 4 and 12, it can be seen if the varying form of train/test mismatch stems from acoustic noise, then the optimal weighting factor α^* will be proportional to β_a. Dupont and Luettin [7] reported such a relationship, but the link between cepstral shrinkage and the weighting factor α was never made. This proportionality can be equated to Equation 12 as,

$$\beta_a = \frac{\alpha_{trn}^*}{\alpha_{tst}^*} \quad (14)$$

where α_{trn}^* and α_{tst}^* are the optimal weighting factors for the train set and test set conditions respectively.

Unfortunately, the ability to find the homoscedastic variance σ_{tst}^2 of the test set requires a violation of causality. However, using Equation 13, under the assumption that the acoustic HMM classifier can be approximated by similar Gaussians, one can make a causal approximation of the acoustic mismatch value β_a from the log-likelihoods of the acoustic HMM classifier,

$$\beta_a \approx \frac{\sqrt{Var\{\mathbf{ll}_{tst}\}}}{\sqrt{Var\{\mathbf{ll}_{trn}\}}} \quad (15)$$

which can be used to estimate β_a for a specific acoustic noise context where,

$$\mathbf{ll} = [\log p(\mathbf{O}|\boldsymbol{\lambda}_1), \ldots, \log p(\mathbf{O}|\boldsymbol{\lambda}_N)] \quad (16)$$

the vector $\mathbf{ll}$ contains the log-likelihoods of N class digits taken from the acoustic HMM classifier. In this approach there is *no* violation of causality as $\mathbf{ll}_{tst}$ is found after classification. A comparison between the non-casual approximation of β_a in Equation 12 and the actual value calculated in Equation 15 can be seen in Figure 1 over a gamut of acoustic noise levels.

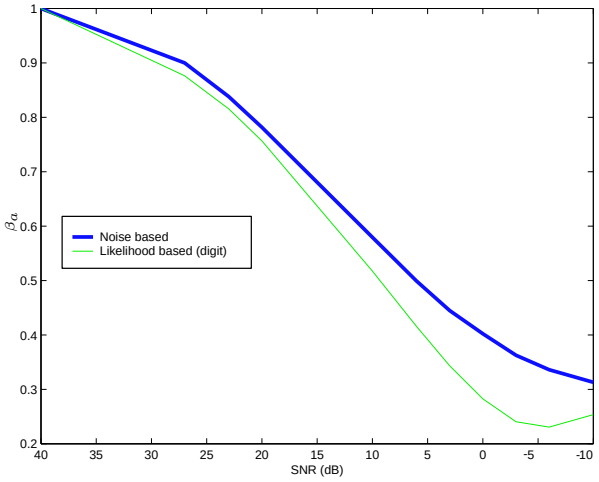

Fig. 1. Comparing the log-likelihood approximations of β_a using audio log-likelihoods taken from the speech and speaker recognition HMM classifiers.

The log-likelihood approximation of β_a in Figure 1 performs quite well, with the approximation only starting to fail badly after approximately 0 dBs of acoustic noise.

It must be noted that the visual modality's classifier also contains a mismatch in practice. This mismatch is *not* due to the addition of any external noise in the visual modality, rather it is associated with the visual HMM classifier being

undertrained. If there was no train/test mismatch in the visual modality then $\beta_v = 1$ could be assumed, however in the presence of a train/test mismatch this assumption does not hold. An empirical value for β_v can be found by assuming in clean acoustic conditions that $\beta_a = 1$ provided the acoustic classifier is not undertrained. Given an optimal weighting α^*_{cln} found under clean acoustic conditions one can then find β_v as,

$$\beta_v = \frac{1 - \alpha^*_{cln}}{\alpha^*_{cln}} \tag{17}$$

Using this estimate of β_v one can gauge the effectiveness of modelling weighting factor α^* based on the cepstral shrinking nature of acoustic speech in additive noise. Figure 2 contains word error rates (WER) calculating α^* by a,

exhaustive search technique: where α is varied between zero and one, with an exhaustive search conducted to find the best weighting α^* in terms of recognition performance.

noise based technique: where the homoscedastic variance $\sigma^2_{MFCC}(\mathsf{NOISE})$ of the acoustic MFCC features with deltas is used to approximate α^*. Equations 4 and 12 are used to calculate α^*, where $\sigma^2_{trn} = \sigma^2_{MFCC}(40dB)$ and $\sigma^2_{tst} = \sigma^2_{MFCC}(\mathsf{NOISE})$. The notation of $\sigma^2_{MFCC}(\mathsf{NOISE})$ refers to the homoscedastic variance of MFCC features with a certain amount of additive acoustic noise.

likelihood based technique: where the standard deviation of acoustic log-likelihoods from the acoustic classifier is used to approximate α^* using Equation 4 and 15. For Equation 15 the reference clean variance $Var\{\mathbf{ll}_{trn}\}$ is obtained from the train set before testing.

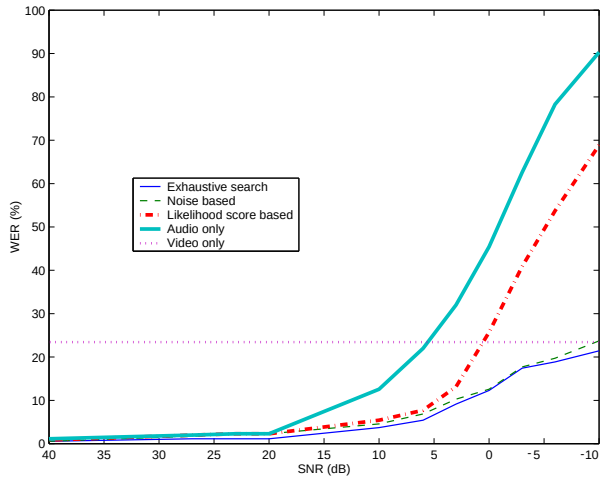

Fig. 2. Evaluation of techniques for approximating α^* in (a) speech recognition and (b) speaker recognition using the weighted product rule.

In Figure 2 it can be seen that the relationship between levels of additive acoustic noise and the optimal weighting factor α^* seems to hold, as similar error rates are received for the exhaustive search and noise based techniques. Both the exhaustive search and noise based techniques receive error rates below the catastrophic fusion boundary. The causal likelihood based technique fares quite well in small amounts of acoustic noise. However, the technique starts to fail in the presence of high amounts of acoustic noise (i.e 0 to -10 dBs). This can be attributed firstly to the variance estimate $Var\{\mathbf{ll}_{tst}\}$ being unreliable due to sample error, as there are only N log-likelihoods being used to gain the estimate. Secondly, it was shown in Figure 1 that the approximation of β_a tends to fail after 0 dBs of additive acoustic noise.

6. CONCLUSIONS

The mechanism the weighted product rule employs for achieving effective classifier combination performance, specifically for the task of AVSR, has been elucidated. It has been postulated that the weighted product rule acts in two capacities. Firstly, in a dampening capacity, similar to the sum and majority vote rules, where the *overall* effect of confidence errors upon combination is addressed. Secondly, in an adapting capacity, where the some of the *individual* classifier confidence errors are addressed for particular types of train/test mismatches (i.e. those seen from cepstral shrinking). Using this understanding one can gain far more accurate estimates of an effective weighting α^* based on knowledge of how train/test mismatches manifest in either speech modality.

A causal approach for estimating α^* was developed based on the log-likelihood scores from the acoustic HMM classifier. This approach was able to provide WERs below the catastrophic fusion boundary for virtually all noise levels.

7. REFERENCES

[1] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," *IEEE Transacations on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, March 1998.

[2] S. Pigeon, "The M2VTS database," Laboratoire de Telecommunications et Teledection, Place du Levant, 2-B-1348 Louvain-La-Neuve, Belgium, 1996.

[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press Inc., 24-28 Oval Road, London NW1 7DX, 2nd edition, 1990.

[4] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 2.2)*, Entropic Ltd., 1999.

[5] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58–70, September 1996.

[6] A. Adjoudani and C. Benoit, "Audio-visual speech recognition compared across two architectures," in *EUROSPEECH'95*, Madrid Spain, September 1995, pp. 1563–1566.

[7] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, September 2000.