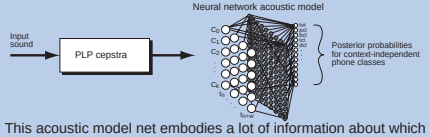
Speech/music discrimination based on posterior probability features

Gethin Williams • Department of Computer Science, University of Sheffield, UK • g.williams@dcs.shef.ac.uk
Dan Ellis • International Computer Science Institute, Berkeley CA USA • dpwe@icsi.berkeley.edu

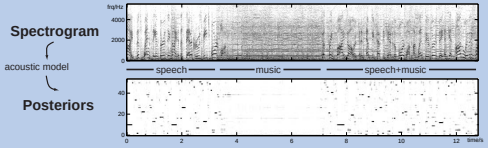
Summary: This work uses the acoustic model from a speech recognizer to estimate the probability that the current signal corresponds to a phoneme. Statistics of these distributions are used to distinguish speech from nonspeech (such as music) with high accuracy.

Introduction

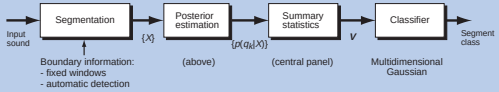
- In a hybrid connectionist-HMM speech recognizer [MorgB95], a neural net estimates the posterior probability of a phone label given a window of acoustic features, $p(q_k|X)$:



- This acoustic model net embodies a lot of information about which feature vectors correspond to speech sounds and phonemic boundaries (in this case, in a Broadcast News corpus [Cook99]).
- Plotting the posterior probabilities of every (context-independent) phone as a function of time reveals clear differences between speech and nonspeech such as music:



- We designed some simple statistics to apply to a segment's worth of posterior probabilities to reveal if the segment was speech or not:



- Segmenting audio and classifying the segments is useful to avoid wasting effort attempting to recognize words in nonspeech [Williams99], as well as for indexing etc.

Why speech features for locating nonspeech?

- The basic features we are using to represent the sound (PLP cepstra) have been specifically developed to represent phonetic variety in speech, not other audio (unlike [ScheirS97]).
- However, they are precisely tuned to the characteristics of speech, and consequently behave very differently when the signal is not.

The Four Segment-Summary Statistics

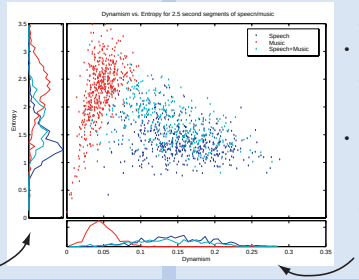
Average per-frame entropy

- **Entropy** is a measure of the randomness or unpredictability of a process. Low entropy signals are regular; high entropy systems are unpredictable.
- Given a discrete probability distribution function, such as the posteriors from the acoustic model, the entropy is defined as:

$$H(x) = -\sum_{\forall x} p(x)\cdot\log_2\left(p(x)\right)$$

This can be interpreted as the number of bits of information carried by each sample of the process.

- When the feature vectors are a good fit to the acoustic model, the model is confident which phone is present and entropy is low.
- Music fails to fit the model, posteriors are equivocal, and the average per-frame entropy over the segment is high.



'Dynamism'

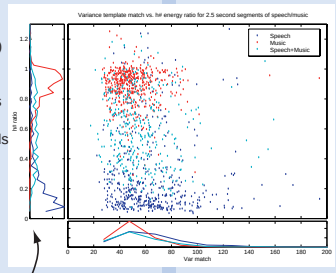
- We can see from the posterior display in the introduction that speech segments have regular, sharp transitions in the posteriors, corresponding to phone transitions in the utterance.
- Music segments, by contrast, give posteriors that change less often and typically have slow transitions because the model cannot decide which phone is present.
- The squared first-order difference, $\Sigma_k \left(p(q_k^n) - p(q_k^{n-1})\right)^2$ (where $p(q_k^n)$ is the probability of phone q_k at time step n) is large across sharp transitions, and thus when averaged across a segment, giving so-called **dynamism**, it is higher for speech than nonspeech.

h# energy ratio

- The acoustic model was trained on a large Broadcast News corpus with one label (h#) used for all silence and nonspeech.
- In examples of clean speech, segments labeled h# will usually be pauses and gaps between words.
- In nonspeech, an arbitrary variety of sounds may fall into the nonspeech class.
- The average energy ratio of the signal in frames labeled h# to the rest of the signal:

$$\frac{\Sigma_n\, p(q^n_{h\#})\cdot e^n \,/\, \Sigma_n\, p(q^n_{h\#})}{\Sigma_n\left(1\text{-}p(q^n_{h\#})\right)\cdot e^n \,/\, \Sigma_n\left(1\text{-}p(q^n_{h\#})\right)}$$

(where $p(q^n_{h\#})$ is the probability of nonspeech at time n and e^n is the signal energy) is close to 0 for clean speech and around 1.0 for nonspeech.



Variance template match

- Some nonspeech segments will, by chance, fall into one of the phone classes.
- When this happens, however, the balance between phones in the segment will typically be very different from a segment of real speech.
- We tried to capture this by defining a **template** of the per-label variance for speech segments, along with the variance of those samples over the training set.
- For an unknown example, the variance-weighted (Mahalanobis) distance to this template is a measure of how speech-like the segment is.
- This was useful for the 15 second segments (see Results), but 2.5 second segments are too short to be balanced.

Results

- The measures were evaluated on the Scheirer/Slaney database of 15 second segments recorded at random from FM radio stations.
- The database contains 80 examples of speech alone, 100 of music of various kinds, and 60 of speech over background music.
- We tested in two conditions: segment statistics calculated over the full 15 second examples, and over 2.5 second segments formed by dividing each example into 6 equal pieces.
- Longer segments have more stable statistics and are easier to classify.
- The data was divided in 4 equal 'cuts', with 3 used to set the decision-model parameters (single-Gaussian models of the feature distributions for speech and nonspeech classes) to test the remaining quarter, repeated for each cut. d' is a measure of class mean separation.
- Classification was via a simple likelihood ratio test between the two distribution models.

	15 second segments				*2.5 second segments*			
Feature	*Speech*	*Music*	*Error*	*d'*	*Speech*	*Music*	*Error*	*d'*
Entropy	75/80	73/80	7.5%	3.3	425/480	402/480	13.9%	1.9
Dynamism	80/80	80/80	0%	4.9	447/480	462/480	5.3%	3.0
h# energy	78/80	79/80	1.9%	6.0	434/480	458/480	7.1%	2.9
Var. tplt.	78/80	80/80	1.3%	4.3	151/480	444/480	38.2%	0.5
4 features	80/80	80/80	0%	9.6	472/480	472/480	1.7%	4.7
3 features	80/80	80/80	0%	7.9	476/480	472/480	1.3%	4.7

- '4 features' classifications are made with all features;
  '3 features' use just Entropy, Dynamism and h# energy ratio.

Acknowledgments & References

This work was funded by the European Union through the SPRACH (20077) and THISL (23495) projects. We are also very grateful to Eric Scheirer, Malcolm Slaney and Interval Research Corporation for making available to us their database of speech/music examples. Thanks to Gary Cook and Tony Robinson of Cambridge University for the RNN acoustic model used in this project.

[MorgB95] N. Morgan and H. Bourlard, "Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach," *Signal Processing Magazine*, pp 25-42, May 1995.

[Cook99] G. Cook, J. Christie, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, A. Robinson and G. Williams, "The SPRACH System for the Transcription of Broadcast News," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Herndon VA, Feb 1999.

[ScheirS97] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. ICASSP*, Munich, 1997.

[Williams99] G. Williams, *Acoustic confidence measures in connectionist speech recognition*, Ph.D. thesis, Dept. of Computer Science, Univ. of Sheffield, 1999.