# MUSIC MODELS FOR MUSIC-SPEECH SEPARATION

*Thad Hughes and Trausti Kristjansson*

Google Research, USA

## ABSTRACT

We consider the task of speech recognition with loud music background interference. We use model-based music-speech separation and train GMM models for music on the audio prior to speech. We show over 8% relative improvement in WER at 10 dB SNR for a real world Voice Search ASR system.

We investigate the relationship between ASR accuracy and the amount of music background used as prologue and the the size of music models.

Our study shows that performance peaks when using a music prologue of around 6 seconds to train the music model. We hypothesize that this is due to the dynamic nature of music and the structure of popular music. Adding more history beyond a certain point does not improve results. Additionally, we show moderately sized 8-component music GMM models suffice to model this amount of music prologue.

***Index Terms—*** ASR, noise robustness, noise reduction, non-stationary noise, music

## 1. INTRODUCTION

Thanks to the rapid adoption of mobile computing devices such as smart phones and tablet computers, automatic speech recognition (ASR) technology is used in an increasingly wide range of environments with diverse background noise characteristics. For example, Google's Voice Search speech recognition system is now used on many types of smart phones [1], and has recently been extended to work inside the Chrome web browser, which runs on desktops and tablet computers.

Voice Search works well on mobile devices when used near-field, i.e. when the device is brought close to the user's mouth, which results in a high signal-to-noise ratio (SNR). However, speech-enabled applications are being increasingly used on tablets and desktops where the the distance from the microphone to the user is greater and the speech-to-interference ratio is worse by 10-20 dB.

Further, the type of noise interfering with the speech signal in far-field scenarios is often non-stationary. Non-stationary noise such as background speech, television, and music is prevalent even in the query data received by Google Voice Search on the mobile phone platform, and as far-field applications of ASR technology grow, the problem will worsen.

Non-stationary noise combined with a lower SNR can cause problems for ASR systems, because the noise is not well-represented by the acoustic model and cannot be easily removed with simpler techniques such as spectral subtraction [2] or Vector Taylor Series (VTS) [3].

In this paper, we describe experiments with using model-based techniques such as Max and Algonquin to remove musical background noise from speech before it is processed by an ASR system. Section 2 describes the algorithms we have applied, section 3 describes the training and evaluation setup, and section 4 describes the way in which parameters were tested and presents the results. Finally, section 5 concludes with a summary and discussion of future work.

## 2. MODEL-BASED MUSIC-SPEECH SEPARATION

A number of model-based noise suppression techniques have been proposed in the literature including Parallel Model Combination (PMC), Vector Taylor Series (VTS), MAX[4], Algonquin[5] and recently Non-negative Matrix Factorization (NMF). In this study we report results for Max and Algonquin. We anticipate that the results will carry over to other methods.

The components of these methods are the speech model $p(x)$, the noise model $n(x)$, and the *interaction model* $p(y|x, n)$. The joint distribution is

$$p(y, x, n) = p(y|x, n)p(x)p(n). \tag{1}$$

The interaction model describes the relationship between the power spectra of the observation $y$, $n$ and $x$. In the exact from it is:

$$|Y_t|^2 = |X_t|^2 + |N_t|^2 + 2\sqrt{|X_t||N_t|}\cos\theta_t \tag{2}$$

Where $X_t$ and $N_t$ are complex Fourier coefficients and $\theta$ is the phase angle between the speech an interference. The phase is not known and can be modeled as a random variable.

In both Max and Algonquin, the speech and interference signals are modeled with Gaussian mixtures.

$$p(x) = \sum_i \pi_i N(x; \mu_i, \Sigma_i) \tag{3}$$

where $x$ is a frame of log-spectrum features, $p(x)$ is the probability of the observation, $\pi_i$ is the mixture weight, $\mu$ is the

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

Gaussian mean and $\Sigma$ is the covariance matrix, which we assume to be diagonal throughout this study.

It is possible to introduce strong temporal models [5] [6]. These models are especially useful when the interference signal is very similar to the target speech signal. In this study, we do not use temporal constraints for the speech or interference models.

### 2.1. Algonquin interaction model

The Algonquin method approximates Eqn. 2 as

$$|Y_t|^2 = |X_t|^2 + |N_t|^2 + e \tag{4}$$

where $e$ is an error term. In the log domain, the error term is assumed to be zero mean Gaussian with variance $\Psi$. The interaction likelihood can therefore be written as:

$$p(y|x,n) = N(y; \log(\exp(x) + \exp(n)), \Psi) \tag{5}$$

The approximation is in the magnitude of the uncertainty due to the unknown phase. Slightly better approximations are possible that take the relative size of this term into account [7].

### 2.2. Max interaction model

The Max method makes a more severe approximation to Eqn. 2. The observation is assumed to be equal to the maximum of the speech or interference signals, i.e.

$$|Y_t|^2 = \max(|X_t|^2, |N_t|^2) \tag{6}$$

The resulting likelihood can be expressed as:

$$p(y|x,n) = \delta(y - \max(x,n)) \tag{7}$$

## 3. EXPERIMENTAL SETUP

The experimental setup is shown in Figure 1. We divide the utterances for each speaker into training and testing sets, using the least-noisy 70% of the data (according to SNR as estimated by the percentile method) for training the speaker's speech GMM. This ensures that the speech GMM is composed of relatively clean speech.

### 3.1. Training speech GMMs

To train the speech model for each speaker (pictured in the left column of Figure 1), we first compute 256-dimensional log-spectral feature vectors for each of the speaker's training utterances, using 25ms frames spaced at 10ms intervals. Next, we separate the speech frames from the non-speech frames using a percentile-based VAD and use only the speech frames to estimate a GMM averaging at least 20 frames per component, with at most 200 components. We also use the least-noisy of the non-speech frames to estimate a smaller 20-component GMM, which we combine with the speech GMM to form a speech and non-noise GMM.

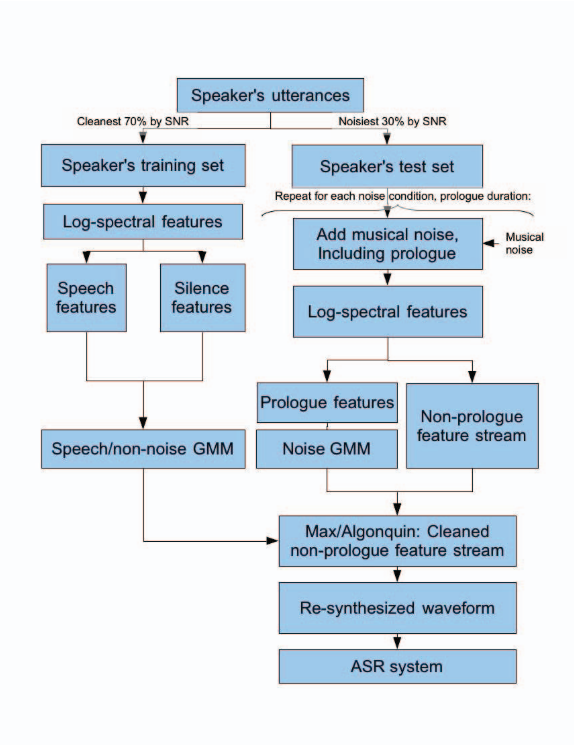

**Fig. 1**: Experimental setup showing the training (left) and testing (right) pipelines. Each speaker's utterances were split 70%/30% into training and testing data. The training data was used to build per-speaker speech models of the log-spectral features, and the testing data was augmented with musical noise, after which a noise model was computed from the music prologue of each testing utterance, and together with the speech model, used to remove the musical noise interfering with the speech.

### 3.2. Training noise models

The remaining 30% of the data for each speaker is held out as test data, as shown in the right column of Figure 1. For each utterance in the test set, we select a random song from a database of 500 popular songs, and mix it with the utterance at the desired SNR. We include the desired amount of musical prologue before the onset of speech in the utterance. We then compute the same 256-dimensional log-spectral feature vectors used to create the speech model, and use the feature frames from the prologue to construct noise GMMs of varying size.

### 3.3. Noise removal and evaluation

We then apply the Max and Algonquin noise reduction techniques using the per-speaker speech model constructed from the speaker's training data, and the per-utterance noise model constructed from each utterance's prologue.

The resulting feature frame sequence is then re-synthesized as a waveform using the overlap-add method and sent to the speech recognizer to test the de-noising quality. All speech recognition was performed with a recent production version of Google's Voice Search speech recognizer. The gender independent acoustic model uses standard 3-state context-dependent (triphone) GMMs trained on a 39-dimensional PLP cepstral coefficients, optimized for mid-field data, and is trained with Linear Discriminant Analysis (LDA), semi-tied covariance (STC) modeling [8], and boosted MMI [9]. The Voice Search language model used for recognition contains more than one million English words.

### 3.4. Dataset characteristics

The entire dataset consists of approximately 38,000 manually-transcribed utterances containing 38 hours of anonymized English-language spoken queries to Google Voice Search. The utterances were spoken by 296 different speakers, and range in length from 0.2 to 12.3 seconds, with a mean of 3.6 seconds. The utterances were recorded and stored in 16-bit, 16kHz uncompressed format.

The dataset contains a varying amount of speech for each speaker. To account for the varying amount of training data for each speaker, we group the speakers into three groups, according to the number of frames used as training data for that speaker. The groups are:

| Group | # training data frames | # speakers |
|---|---|---|
| Small | $< 5,000$ | 62 |
| Medium | $\geq 5,000$ and $< 10,000$ | 145 |
| Large | $\geq 10,000$ | 89 |

As described in section 3.2, the test set was selected from this dataset by holding out the most noisy 30%, leaving 13,000 utterances in the test set.

## 4. RESULTS

In this section, we characterize the results according to several parameters of the interfering noise and the algorithms, as well as the amount of training data used to train the speaker model.

Figure 2 shows the relative reduction in word error rate at different SNRs. Notice that the greatest reduction in WER happens for the noisiest condition. This trend is expected for two reasons. First, the recognizer's acoustic models are trained on mixed conditions and the benefit from noise cleaning are less substantial when the noise level is low and similar to the conditions the recognizer was trained on.

Second, any noise enhancement methods with free parameters will introduce *decision noise* and artifacts. It is common to retrain the system in a "multi-condition" style on the processed audio to alleviate the issue of artifacts. Although retraining should give overall better results, it should not affect the trends seen here and hence we did not retrain the system on processed speech.

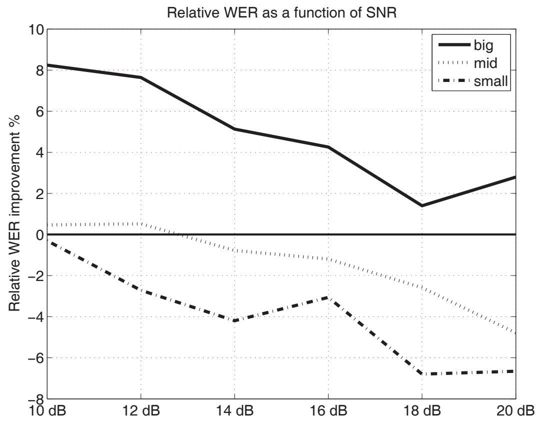

**Fig. 2**: Relative word error rate as a function of SNR.

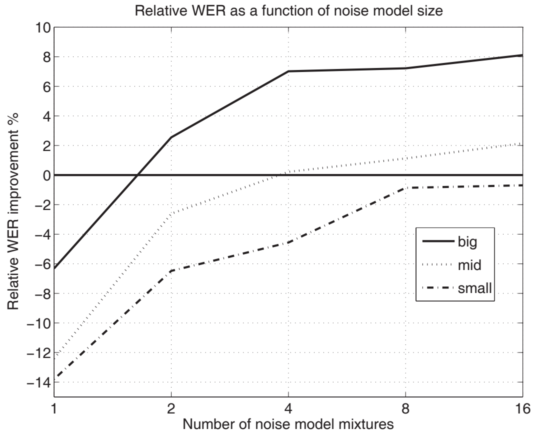

**Fig. 3**: Relative word error rate as a function of number of mixtures in noise model.

Figure 3 shows the relative improvement in word error rate for different noise model sizes for a music prologue of length 8 seconds. This plot is characterized by a steep improvement until *enough* components are used. After this, there is a plateau and only modest gains are achieved by doubling and quadrupling the number of components.

Figure 4 shows the relative reduction in word error rate for different amounts of music prologue used to train the noise GMM, and for different noise GMM sizes. This plot shows that using approximately 6 seconds of music prologue is optimum for removing subsequent musical interference later in the signal. Using less prologue is insufficient to build an adequate noise model, and more prologue is likely detrimental because it includes music too dissimilar to the music interfering with the subsequent speech. Also interesting is that increasing the noise model size from 8 to 16 components degrades performance slightly, perhaps due to over-fitting.

The average reduction in WER over all SNRs for the Max algorithm was 2.7% and for Algonquin algorithm it was 4.7%,

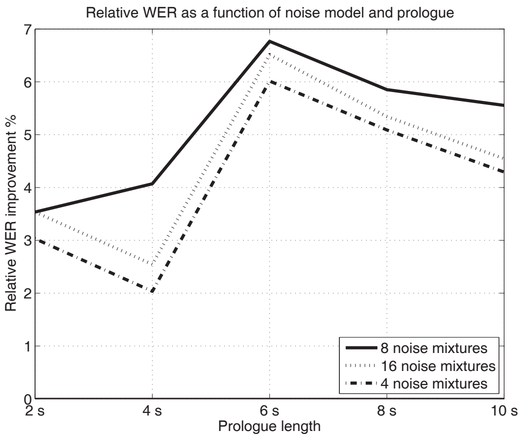

**Fig. 4**: Relative word error rate as a function of the length of the music prologue used to train the noise GMM, as well as the size of the noise GMM, when using Algonquin at an SNR of 16dB.

| # training frames | <5000 | 5000-10000 | >10000 |
|---|---|---|---|
| Rel. WER improvement | -3.9% | -1.4% | 5.0% |

**Table 1**: Average reduction in WER at 10dB for different amounts of speaker training data.

The Algonquin algorithm outperforms the MAX algorithm with almost double the reduction in WER. Additionally, the Max algorithm was more sensitive to over-fitting, i.e. the performance worsened as the noise model increased beyond the optimum number of components.

The amount of training data available for training the speaker models has a dramatic effect on performance. This is shown in Table 1, and is shown in all the plots. With approximately 100 seconds or more training data good results can be achieved. Interestingly, with less than 50 seconds of training data, the speaker models are so poor that processing the audio hurts recognition almost across the board.

## 5. DISCUSSION

In this paper we have shown that we can achieve substantial reduction in WER for speech with music background noise, even with modest music models trained on a small amount of data prior to the speech. It is a common condition for Voice Search and other applications to have access to the noise environment prior to voice input, making this a realistic scenario. We expect other non-speech background noise such as TV or radio noise to have similar characteristics.

A very interesting result is that performance peaks for a prologue of approximately 6 seconds. We hypothesize that this is due to the dynamic nature of music and that the phrasing structure of popular music.

Future work will include exploring better ways to construct speaker models with less training data, such as adaptation of a global speaker model using a small amount of individual speaker data.

## 6. REFERENCES

[1] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, “Google Search by Voice: A case study,” in *in Visions of Speech: Exploring New Voice Apps in Mobile Environments, Call Centers and Clinics*. 2010, A. Neustein, Ed. Springer.

[2] Steven Boll, “Suppression of acoustic noise in speech using spectral subtraction,” in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1979.

[3] Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern, “A vector Taylor series approach for environment-independent speech recognition,” in *Proc. ICASSP-96*, 1996, pp. 733–736.

[4] A. Varga and R. Moore, “Hidden Markov Model Decomposition of Speech and Noise,” in *in IEEE International Conference on Speech and Signal Processing*, 1990, pp. 97–100.

[5] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath, “Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system,” in *in ICSLP*, 2006, pp. 97–100.

[6] John R. Hershey, Steven J. Rennie, Peder A. Olsen, and Trausti T. Kristjansson, “Super-human multi-talker speech recognition : A graphical modeling approach,” in *Computer Speech and Language*, 2009.

[7] Li Deng, Jasha Droppo, and Alex Acero, “Log-Domain Speech Feature Enhancement Using Sequential MAP Noise Estimation and a Phase-sensitive Model of the Acoustic Environment,” in *in ICSLP*, 2002, pp. 1813–1816.

[8] M. Gales, “Semi-tied covariance matrices for hidden Markov models,” in *IEEE Transactions on Speech and Audio Processing*, 1999, pp. 272–281.

[9] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, “Boosted MMI for model and feature space discriminative training,” in *in ICASSP*, 2008.