ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

# Acoustic Features for Classification Based Speech Separation

*Yuxuan Wang*[1], *Kun Han*[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive Science, The Ohio State University, USA
{wangyuxu,hank,dwang}@cse.ohio-state.edu

## Abstract

Speech separation can be effectively formulated as a binary classification problem. A classification based system produces a binary mask using acoustic features in each time-frequency unit. So far, only pitch and amplitude modulation spectrogram have been used as unit level features. In this paper, we study other acoustic features and show that they can significantly improve both voiced and unvoiced speech separation performance. To further explore complementarity in terms of discriminative power, we propose a group Lasso approach for feature combination. The final combined feature set yields promising results in both matched and unmatched test conditions.

**Index Terms**: Speech separation, binary classification, feature combination, group Lasso

## 1. Introduction

Speech separation refers to the problem of separating target speech from its background interference. For monaural mixtures, one can only utilize intrinsic properties of speech or interference to do the separation, which remains a very challenging problem in signal and speech processing. In this paper, we consider monaural speech separation from nonspeech interference.

In the last decade, computational auditory scene analysis (CASA), which does separation based on perceptual principles, has shown considerable promise. The estimation of the ideal binary mask (IBM) is suggested as a primary computational goal of CASA [14]. The IBM is a time-frequency (T-F) binary mask, constructed from premixed target and interference. A mask value 1 for a T-F unit indicates that the signal-to-noise ratio (SNR) within the unit exceeds a threshold (target-dominant), and 0 otherwise (interference-dominant). In this paper, we use a 0 dB threshold in all the experiments.

It is natural to cast IBM estimation as a binary classification problem. Recent studies have applied this formulation and achieved good speech separation results in both anechoic and reverberant environments [12, 7, 9, 3]. In [5, 7], pitch-based features are used and show good performance, however, they cannot deal with unvoiced speech which lacks harmonic structure. In [9, 3], amplitude modulation spectrogram (AMS) is used, which is able to separate unvoiced speech as AMS can be extracted from both voiced and unvoiced speech. However, in [3] it is shown that the generalization of AMS is not good.

Our previous study [3] suggests that support vector machines (SVMs) are better than Gaussian mixture models (GMMs) for T-F unit classification. But on the unit level feature side, only pitch and AMS have been studied so far. In this paper, we aim to enlarge the possible feature repository for better IBM estimation. In order to explore complementary features in terms of discriminative power, we also propose a group Lasso [16] approach to combine features in a principled way.

This paper is organized as follows. In the next section, we present an overview of the proposed system. Section 3 describes the proposed feature combination approach. The experimental results are reported in Section 4. The last section concludes this paper.

## 2. System Overview and Feature Extraction

A sound mixture with 16 kHz sampling frequency is first fed into a 64-channel gammatone filterbank from 50 Hz to 8000 Hz. The output in each channel is then divided into 20-ms frames with 10-ms overlapping between consecutive frames. This procedure produces a T-F representation called a cochleagram. Features are extracted from each T-F unit in the cochleagram, and we train a Gaussian-kernel SVM for each subband channel separately.

We study existing speaker and speech recognition features on the task of speech separation, including gammatone frequency cepstral coefficients (GFCC), mel-frequency cepstral coefficients (MFCC), relative spectral transform (RASTA) and perceptual linear prediction (PLP). These acoustic features are usually derived at the frame level. To get features for the T-F unit $u_{c,m}$ in channel $c$ and at frame $m$, we take the filtered output $x_c(t)$ in channel $c$. Treating $x_c(t)$ as the input, conventional frame level acoustic feature extraction is carried out and the feature vector at frame $m$ is taken as the feature representation for $u_{c,m}$. This simple procedure enables us to extract unit features based on acoustic features originally derived in a frame-by-frame manner, as well as features involving neighboring frames, as done in RASTA.

We follow common practice to extract 15-D AMS, 31-D MFCC, 13-D PLP and 13-D RASTA-PLP features, and we describe the extraction of 6-D pitch-based features and 31-D GFCC as below.

### 2.1. Pitch-based Features

To get pitch-based features for $u_{c,m}$, we first calculate the normalized autocorrelation function at each time lag $\tau$, denoted by $A(c, m, \tau)$. Its value at the pitch period $\tau_m$ indicates the dominance of voiced speech [5].

The average instantaneous frequency $\bar{f}(c, m)$ estimated from the zero-crossing rate of $A(c, m, \tau)$ are used to derive the second and third features. If the T-F unit $u_{c,m}$ is target-dominant, the product of $\bar{f}(c, m)$ and $\tau_m$ gives a harmonic number. Hence, we set the second feature to be the nearest integer of $\bar{f}(c, m)\tau_m$ and the third feature to be the distance between the product and its nearest integer.

The next three features are derived in the same way except that they are extracted from the envelopes of filter responses. The resulting 6-D feature vector is:

$$\mathbf{x}_{c,m} = \begin{pmatrix} A(c, m, \tau_m) \\ [\bar{f}(c, m)\tau_m] \\ |\bar{f}(c, m)\tau_m - [\bar{f}(c, m)\tau_m]| \\ A_E(c, m, \tau_m) \\ [\bar{f}_E(c, m)\tau_m] \\ |\bar{f}_E(c, m)\tau_m - [\bar{f}_E(c, m)\tau_m]| \end{pmatrix} \quad (1)$$

where $[\cdot]$ denotes the round operation, and subscript $E$ indicates envelope.

In training, we use ground truth pitch extracted from clean speech by PRAAT [1]. In testing, we use pitch estimated by a recently proposed multipitch tracker [8].

### 2.2. Gammatone Frequency Cepstral Coefficient

GFCC is an effective speaker feature [13]. To get GFCC, a signal is decomposed by a 64-channel gammatone filterbank first. Then, we decimate the filter response to an effective sampling rate of 100 Hz, resulting in a 10-ms frame shift. The magnitudes of the decimated filter outputs are then loudness-compressed by a cubic root operation. Finally, discrete cosine transform is applied to the compressed signal and the first 31 coefficients are preserved to yield GFCC.

## 3. Feature Combination

It is shown in speech recognition that complementarity exists between basic acoustic features [2]. Large performance boosts could be observed even if individual features do not perform well. Our goal is to select a set of complementary features in a principled way. The complementarity should be related to the discrimination of target-dominance and interference-dominance. Here, by complementarity, we mean that each feature type provides complementary information to boost classification and thus their combination (concatenation in paper) should outperform an individual type.

This problem is essentially a group variable selection problem, which is to find important groups of explanatory factors for prediction. Group Lasso [16] approaches this problem by incorporating a mixed-norm regularization over regression coefficients. Since our classification is binary, we employ the logistic regression extension of group Lasso [11]. The estimator is

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \sum_i \log\Big(1 + \exp(-y_i(\boldsymbol{\beta}^T \mathbf{x}_i + a))\Big) + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_{\mathcal{I}_g}\|_2 \quad (2)$$

where $\mathbf{x}_i$ and $y_i$ are the $i$th training sample and label scaled to $\{-1, 1\}$, respectively. $a$ is a parameter (intercept). $\|\cdot\|_2$ is the $\ell_2$ norm. $\boldsymbol{\beta}$ consists of $G$ predefined non-overlapping groups and $\mathcal{I}_g$ is the index set of the $g$th group. The log loss in the minimization concerns discrimination. The second term in the minimization is an $\ell_1/\ell_2$ mixed-norm regularization, which imposes an $\ell_1$ regularization between groups and an $\ell_2$ regularization within each group. It is well known that the $\ell_1$ norm induces sparsity, therefore the $\ell_1/\ell_2$ regularization results in group sparsity hence group level feature selection. The level of sparsity of the resulting model can be adjusted by varying the regularization parameter $\lambda$.

To do feature combination, each feature type is defined as a group and they are concatenated together, e.g., AMS (all 15 feature elements) is defined as the first group, PLP as the second, and so on. Then, for a fixed $\lambda$, we minimize the above objective to obtain $\hat{\boldsymbol{\beta}}_\lambda$. Groups having large regression coefficients shall be included in the complementary feature set.

The feature combination is conducted on AMS, PLP, RASTA-PLP, GFCC and MFCC. All features are normalized prior to using group Lasso. With best $\lambda$ (determined by cross-validation), we found that AMS, RASTA-PLP and MFCC have significantly larger regression coefficients than the others. Hence, AMS+RASTA-PLP+MFCC is set as our complementary feature set, denoted as COMP. Here "+" means concatenation.

## 4. Evaluation and Comparison

For the training set, we randomly mix 50 IEEE utterances [6] recorded by a female speaker with three types of noises: N1 – bird chirps, N2 – crow noise, and N3 – cocktail party noise [5], at 0 dB. For the test set, we choose 20 new IEEE utterances. Two test conditions are considered. In the matched test condition, the test utterances are mixed with the trained noises N1-N3. In the

Table 1: Classification performance in the matched test condition. Boldface indicates best

| Feature | Overall | | | Voiced | | | Unvoiced | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| | HIT | FA | HIT−FA | HIT | FA | HIT−FA | HIT | FA | HIT−FA | |
| AMS | 70% | 6% | 64% | 76% | 8% | 68% | 49% | 4% | 45% | 84.6% |
| PLP | 79% | 9% | 70% | 83% | 10% | 73% | 65% | 6% | 59% | 86.5% |
| RASTA-PLP | 74% | 7% | 67% | 79% | 9% | 70% | 56% | 4% | 52% | 85.9% |
| GFCC | 87% | 8% | 79% | 89% | 9% | 80% | 77% | 6% | 71% | 90.1% |
| MFCC | 82% | 7% | 75% | 86% | 8% | 78% | 69% | 5% | 64% | 88.8% |
| PITCH | N/A | N/A | N/A | 77% | 16% | 61% | N/A | N/A | N/A | N/A |
| COMP | 86% | 5% | **81%** | 89% | 6% | **83%** | 75% | 3% | **72%** | **91.8%** |

Table 2: Classification performance in the unmatched test condition

| Feature | Overall | | | Voiced | | | Unvoiced | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| | HIT | FA | HIT−FA | HIT | FA | HIT−FA | HIT | FA | HIT−FA | |
| AMS | 60% | 23% | 37% | 64% | 22% | 42% | 44% | 25% | 19% | 72.7% |
| PLP | 71% | 30% | 41% | 73% | 30% | 43% | 65% | 32% | 33% | 70.0% |
| RASTA-PLP | 69% | 12% | 57% | 71% | 13% | 58% | 60% | 9% | 51% | **83.8%** |
| GFCC | 77% | 33% | 44% | 76% | 32% | 44% | 77% | 34% | 43% | 69.4% |
| MFCC | 74% | 29% | 45% | 75% | 29% | 46% | 70% | 29% | 41% | 71.7% |
| PITCH | N/A | N/A | N/A | 76% | 20% | 56% | N/A | N/A | N/A | N/A |
| COMP | 80% | 20% | **60%** | 80% | 21% | **59%** | 80% | 20% | **60%** | 80.0% |

Table 3: Classification performance in the matched test condition with PITCH combined

| Feature | Overall | | | Voiced | | | Unvoiced | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| | HIT | FA | HIT−FA | HIT | FA | HIT−FA | HIT | FA | HIT−FA | |
| AMS+PITCH | 72% | 12% | 60% | 81% | 10% | 71% | 42% | 14% | 28% | 82.1% |
| PLP+PITCH | 77% | 10% | 67% | 83% | 9% | 74% | 57% | 12% | 45% | 85.0% |
| RASTA-PLP+PITCH | 76% | 10% | 66% | 84% | 9% | 75% | 47% | 10% | 37% | 84.6% |
| GFCC+PITCH | 83% | 10% | 73% | 88% | 8% | 80% | 65% | 12% | 53% | 87.2% |
| MFCC+PITCH | 79% | 9% | 70% | 85% | 8% | 77% | 60% | 11% | 49% | 86.1% |
| COMP+PITCH | 82% | 7% | **75%** | 87% | 7% | **80%** | 70% | 9% | **61%** | **89.0%** |

Table 4: Classification performance in the unmatched test condition with PITCH combined

| Feature | Overall | | | Voiced | | | Unvoiced | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| | HIT | FA | HIT−FA | HIT | FA | HIT−FA | HIT | FA | HIT−FA | |
| AMS+PITCH | 65% | 12% | 53% | 73% | 12% | 61% | 31% | 11% | 20% | 82.5% |
| PLP+PITCH | 71% | 13% | 58% | 75% | 13% | 62% | 53% | 10% | 43% | 83.5% |
| RASTA-PLP+PITCH | 72% | 9% | 63% | 77% | 11% | 66% | 51% | 7% | 44% | 86.1% |
| GFCC+PITCH | 77% | 22% | 55% | 80% | 20% | 60% | 66% | 25% | 41% | 77.7% |
| MFCC+PITCH | 72% | 12% | 60% | 76% | 13% | 63% | 56% | 12% | 44% | 83.7% |
| COMP+PITCH | 79% | 11% | **68%** | 80% | 12% | **68%** | 72% | 12% | **60%** | **86.9%** |

unmatched test condition, the test utterances are mixed with three unseen noises: N4 – crowd noise at a playground, N5 – traffic noise, and N6 – electric fan noise. All test mixtures are mixed at 0 dB.

We employ classification accuracy as well as hit minus false-alarm (HIT−FA) rate as the major evaluation criteria in this paper. Here, the HIT rate is the percent of correctly classified target-dominant T-F units (1s) in the IBM. The FA rate is the percent of wrongly classified interference-dominant (0s) T-F units in the IBM. The HIT−FA rate is proposed in [9] and shown to be highly correlated with human speech intelligibility.

From Table 1, we can see that in the matched test condition all the features perform relatively well in terms of FA rate. The performance gaps mainly stem from the HIT rate. AMS clearly underperforms the other features as its HIT rate is significantly lower. GFCC, on the other hand, performs very well with 79% overall HIT−FA. Unvoiced speech is difficult to separate due to weak energy and the lack of harmonicity. GFCC again is significantly better than the other single features, achieving 71% HIT−FA in the unvoiced interval. The good performance of GFCC is probably due to its effectiveness as a speaker identification feature [13]. While individually AMS, RASTA-PLP, and MFCC do not perform on par with GFCC, their combination COMP performs better in terms of both HIT−FA and classification accuracy.

The unseen noises cause mismatch between training and test set. Not surprisingly, both accuracy and HIT−FA rates of all the features significantly degrade in the un-

Table 5: HIT−FA and SNR comparisons with [9]

| System | Overall HIT−FA | | SNR (dB) | |
|---|---|---|---|---|
| | Matched | Unmatched | Matched | Unmatched |
| Proposed | **75%** | **68%** | 15.6 | **10.5** |
| Kim et al. [9] | 67% | 39% | **16.0** | 8.7 |

matched test condition. From Table 2 we can see that the degradation mainly comes from substantially increased FA rates. Among all the features, PITCH achieves the minimal performance degradation (5%) as it represents intrinsic properties of speech. Interestingly, in this test condition, RASTA-PLP becomes the best one in the single feature category. As shown in [4], RASTA-PLP is essentially a modulation-frequency filter, which retains slow modulations corresponding to speech. Again, the complementary feature set COMP performs the best in this condition in terms of HIT−FA.

Considering the stable performance of PITCH, we further concatenate PITCH with other features in hope to create a richer and more robust representation. For unvoiced frames, we simply set PITCH to be all zeros in the concatenation. Table 3 and 4 show accuracies and HIT−FA rates in two test conditions. In the matched test condition, the combination does not lead to improvement due to pitch estimation error. However, even with estimated pitch, the performance of all the features gains large improvements by the combination in the unmatched test condition. This demonstrates the good generalization ability of PITCH. The final combined feature set COMP+PITCH achieves impressive HIT−FA rates in both voiced and unvoiced intervals.

Finally, we compare the proposed system, i.e., subband SVMs trained using COMP+PITCH, with a recent classification based separation system [9]. We report HIT−FA and SNR results in Table 5. As we can see, the proposed system significantly outperforms Kim et al.'s system in terms of HIT−FA and accuracy (not shown) for both matched and unmatched test conditions. Kim et al.'s system has been shown to improve speech intelligibility in noise; hence we expect the proposed system to provide further improvements due to the significant HIT−FA improvements. Kim et al.'s system produces marginally higher SNR in the matched test condition, mainly due to higher HIT rate. But the FA rate is also much higher than that of the proposed system, which is more detrimental to intelligibility [10]. The proposed system produces about 1.8 dB higher SNR in the unmatched test condition.

## 5. Conclusions

We have shown that by turning frame level speech and speaker features into unit level features, we can significantly expand the feature repository for classification based speech separation. Compared with previously used AMS feature (e.g. [9, 3]), the newly included features significantly improve both voiced and unvoiced speech separation. As single features, GFCC achieves excellent classification performance in the matched test condition, and RASTA-PLP in the unmatched test condition.

We have also identified a discriminative complementary feature set using group Lasso, and showed that PITCH generalizes very well. The final feature set COMP+PITCH shows promising separation results in both matched and unmatched test conditions. We have also demonstrated the effectiveness of the final feature set in a variety of other acoustic conditions. Due to limited space, the interested reader is referred to our technical report [15].

**Acknowledgements.** This research was supported in part by an AFOSR grant (FA9550-08-1-0155) and an STTR grant from AFOSR.

## 6. References

[1] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer (Version 4.3.14)*, 2005. [Online]. Available: http://www.fon.hum.uva.nl/praat

[2] G. Garau and S. Renals, "Combining spectral representations for large-vocabulary continuous speech recognition," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 16, no. 3, pp. 508–518, 2008.

[3] K. Han and D. Wang, "An SVM based classification approach to speech separation," in *ICASSP*, 2011, pp. 5212–5215.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 2, no. 4, pp. 578–589, 1994.

[5] G. Hu, "Monaural speech organization and segregation," PhD Dissertation, The Ohio State University, Biophysics Program, 2006.

[6] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, 1969.

[7] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 17, no. 4, pp. 625–638, 2009.

[8] ——, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 19, pp. 1091–1102, 2011.

[9] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, pp. 1486–1494, 2009.

[10] N. Li and P. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, 2008.

[11] L. Meier, S. V. D. Geer, and P. Bühlmann, "The group Lasso for logistic regression," *J. Roy. Stat. Soc. Ser. B*, vol. 70, no. 1, pp. 53–71, 2008.

[12] M. Seltzer, B. Raj, and R. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[13] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *ICASSP*, 2008, pp. 1589–1592.

[14] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Divenyi P., Ed. Kluwer Academic, Norwell MA., 2005, pp. 181–197.

[15] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech separation," Ohio State University Dept. of CSE, Tech. Rep. TR37, 2011.

[16] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.