POLYPHONIC INSTRUMENT IDENTIFICATION USING INDEPENDENT SUBSPACE ANALYSIS

Pamornpol Jinachitra

CCRMA, Stanford University
Stanford, CA94305, USA
pj97@ccrma.stanford.edu

ABSTRACT

A system which tries to identify the musical instruments playing concurrently in a mixture is investigated in this paper. The features used in classification are derived from the Independent Subspace Analysis (ISA) which somewhat decomposes each source, and the mixture, into its statistically "independent" components. Without re-grouping or actually separating the sources, they offer physiologically-motivated classification of instruments, assuming the decomposition is robust to the mixing process. The system is evaluated on two-tonal instrument mixtures from a set of five instruments and a phrase of real song from CD.

1. INTRODUCTION

The problem of sound source identification is not only an academic curiosity on how the human brains work and how to make a computer system which can do the same. A desire for automatic classification of audio materials according to instruments makes the problem a practical one. A number of techniques and features have been experimented with in order to identify the musical instrument from an isolated tone [1] [2]. There have been, however, far fewer works which consider identification problems from polyphonic signals. Among them, Eggink and Brown [3] used energy bands as features, omitting from use in classification the bands which tend to have overlapping spectra. They obtained on average 49% identification rate using five instruments over one pitch range. Time-domain template matching and features related to note onset and spectral distribution were investigated in [4] and [5] respectively for two-tonal mixtures from a set of three different instruments.

In this paper, ISA is used to decompose a mixture into its "statistically independent" components and spectral bases, hopefully spanning the subspace of each original source in the mixture. The system does not rely on pitch estimation in contrary to previous systems. Physiologically intuitive features can also be derived from the learned bases for classification which are usually lost when more than one sources are active simultaneously.

2. INSTRUMENT IDENTIFICATION USING ISA

Recently, the use of reduced-rank spectral decomposition into its "independent" subspaces showed a promising way to separate the time-varying spectral contents of a single channel audio mixture into small meaningful components in a data-driven manner [6]. In short, the reduced-rank ISA is a decomposition of the magnitude spectrogram $\mathbf{X}$ of a sound mixture into its "independent" components according to the linear model $\mathbf{X}=\mathbf{AS}$, keeping only the components with significant amount of energy. The columns of $\mathbf{A}$ are the spectral bases which span $\mathbf{X}$ while the rows of $\mathbf{S}$ contain their corresponding (temporal) coefficients of the linear summation. The matrices $\mathbf{A}$ and $\mathbf{S}$ as such can be learned using any of ISA's predecessor, Independent Component Analysis (ICA) algorithms. Each source can be spanned by more than one of the bases. A suitable clustering based on components similarity can be used to group the bases that make up the same source and approximately reconstruct the source via inverse-STFT. Smaragdis also showed in [7] the consistency between the notion of mutual independence often used in the cost function of ICA and the grouping of auditory cues of the same source making it even more intuitive.

2.1. ISA of a single instrument sound

When ISA is applied to a spectrogram of a tone produced by an instrument, the result is the decomposition into components roughly distinguishable as sustain, the note-attack and/or other small spectral variations as shown in Fig. 1 for a piano tone. Naturally, the most energetic component corresponds to the sustained note's spectrum with a lot to offer for identification. Despite having lesser energy, other components may also be useful. Human has been found to use note-attacks, the breathiness and some spectral dynamics in source identification as well as the spectral envelope of the tone. In addition, it eliminates the problem of how a note-attack should be defined, since it is now automatically determined according to its mutual independence to the sus-

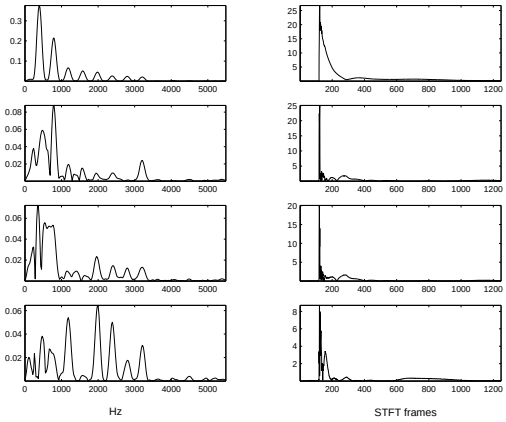

Fig. 1. ISA magnitude spectral bases (left) and magnitude temporal envelope coefficients (right) of a G4-piano tone.

tain portion. Similar decomposition has been found in other instruments used in this experiment.

When there are multiple sources in the mixture, the system hopefully will emit something closely enough to the original basis components of the sources, allowing physiologically intuitive use of such features described above in identification. Such a dramatically successful example is found in a mixture of Oboe and Bb-Clarinet playing note C4 concurrently with Bb-Clarinet lasting about 0.5 second longer. Despite having the same pitch and very similar sound, the spectral bases have been found to be rather well separated and are readily identified by a comparison to their isolated tone's first ISA spectral bases. Using a classifier in section 3, 7 out of 8 components are classified correctly with transient components matching with similar components in the trained prototypes. Admittedly, however, it is still required to be sufficiently non-overlapping, either temporally or spectrally, for such a healthy separation.

Though more than one bases may belong to the spanning set of one source subspace, we have to stop short of grouping them. Clustering of components belonging to the same source is problematic. This is not only because of the difficulty in estimating a reliable similarity measure as used successfully in [7] for a complex mixture, but also by the fact that they simply cannot be used to group transient and steady-state of the same sources together.

The advantage of using such a data-driven algorithm lies in its ability to do auditory grouping with no extra rules [7]. It does not rely on pitch estimation which is hard to do in a polyphonic signal. However, the drawbacks include its reliability on an exposure long enough for meaningful components to be learned. The current linear model is also limiting and only approximately true with respect to the use of magnitude. There is also no guarantee that the bases derived from a mixture will be the same as those learned from a single instrument, or even whether they will be separated like the example shown above. From experiments, this happens from time to time and brings down the identification performance. For example, a beating effect of nearby harmonics can cause the algorithm to yield a basis which is unidentifiable with any of the individual sources. In the next section, we will then examine how well the system can do in lights of these potential obstacles.

3. CLASSIFICATION SYSTEM

The Infomax ICA algorithm by Bell and Sejnowski [8] is used to learn $\mathbf{A}$ and $\mathbf{S}$ from $\mathbf{X}$ in the linear model, keeping only $N = 8$ components for maximum use in further classification. The window length is 10 ms to capture the transient with 50% hop size. The convergence is fast but annealing is also applied.

Various features are calculated from the magnitude spectral bases and temporal envelopes to be used as input to classifiers in the next stage. They include the Mel-Frequency Cepstral Coefficients (MFCC), the Perceptual-Linear Prediction Cepstra (PLPC). They are all calculated using Malcolm Slaney's Auditory Toolbox [9] with 40 frequency bands. The first coefficient is omitted to ignore scaling difference, leaving only twelve each (MFCC-12 and PLPC-12). They describe the shape of a spectral envelope in log-frequency scale similar to the human ears and have been enjoying a considerable success in the past recognition tasks, especially in speech. An additional spectral feature also tried in this experiment is the log-scale spectral centroid (SC) in kHz.

While it is possible to use temporal features, they are notably hard to extract from a polyphonic signal, requiring a good segmentation which is in general hard to do automatically. However, for a pre-segmented note in the case of the two-tonal mixtures, some easy-to-calculate temporal features are experimented. They are the temporal centroid (TC), as a ratio of total duration, the crest factor (CF) in peak/rms and amplitude modulation content, as a ratio of total energy, in the band 4-8 Hz (AM48) and 10-40 Hz (AM1040).

The k-nearest neighbor (k-NN) and Gaussian Mixture Models (GMM) are used as classifiers in this experiment. For k-NN, Mahalanobis distance is used to deal with different scaling and correlation among features. It almost always gives 2-3% better results in the experiments than using Euclidean distance. Each "independent" component and basis is individually classified before taking votes to decide which two sources make up the mixture in the experiment. The maximum number of eight components take part in the vote. If a draw occurs, the source assigned to more of the higher energy components prevails. If still undecided, the higher total number of k-NN's and the lower total distance, or the higher total log likelihood in the case of GMM, will

Instr. (%)	Flute	Bb-Clar	Cello	Oboe	Violin
Flute	100	0	0	0	0
Bb-Clar	0	83	0	17	0
Cello	7	0	86	0	7
Oboe	0	0	0	100	0
Violin	0	0	8	0	92

Table 1. Confusion matrix (%) of isolated tones of five instruments, using kNN-7 and PLPC-12 as features.

be considered until two sources are chosen.

Samples of instruments (about 60 each) were taken from the Iowa and McGill chromatic scale samples[1]. To limit the factor attributed to pitch, only the octave C4-C5 was used. 80% of the notes available were used in training, while the remaining will be combined exhaustively to make mixtures.

4. RESULTS

For comparison with two-tonal mixture, the identification result from isolated notes is shown in Table 1 using k-NN classifier with $k = 7$, $N = 8$ and PLPC-12 as features. The use of more than just the first component is found to be beneficial on average in some cases as shown in Fig. 3. Violin and Bb-Clarinet are mostly confused with Cello and Oboe respectively as should be expected.

The identification results on two-tonal mixtures are shown in Fig. 2. The best combinations on average for k-NN classifier, using only spectral features, is when $k = 7$ and PLPC-12 are used as features. For GMM classifier, the number of Gaussians $K = 40$ gives the best performance using MFCC-12 and SC. Since temporal features are not usually available in real recording, the performance when they are incorporated is also shown separately in Fig. 2. The best combinations for k-NN and GMM classifier, with additional temporal features, come from the set {PLPC-12,AM48}, $k = 7$, and {MFCC-12,TC,AM48}, $K = 40$, respectively. The best average rate of correct identification of both instruments is 45%, comparing with 10% for random guess, whereas that of identifying one instrument in the mixture correctly is 66%, compared with 40% for random guess (guessing two instruments in the mixture and get any one right). Generally, Flute, Bb-Clarinet and Oboe are not as well-identified as Cello and Violin probably due to the pitch-effect on spectral formants which are captured by PLPC and MFCC. GMM classifiers do not perform as well as the k-NN, probably due to a small set of training samples.

In Fig. 3, it is shown that using more than two ISA components (the first two usually correspond to the sustain components of each of the two sources) from the mixture in classification is not beneficial in general. This is in contrast to the identification of isolated tones as also shown on the same graph.

[1]Electronic Music Studio, University of Iowa and McGill Master samples, McGill University

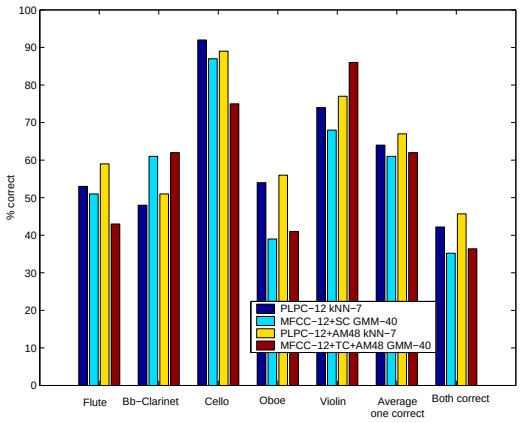

Fig. 2. Best average identification rate (%) of Flute, Bb-Clarinet, Cello, Oboe and Violin, with the average of getting one instrument correct and both correct on the two right most columns. See text for details of classifiers.

As a limiting case, the samples used to mix for testing are also trained and stored as prototypes. This can improve the result by at most 3% on average for all classifiers experimented with, indicating that most of the shortfall lies in the ISA algorithm which cannot achieve separation into original source components as in training.

As another comparison, Bb-Clarinet and Violin are removed from the experiment to remove confusion which might have brought down the performance. They are replaced by Piano and French horn. The result is relatively unchanged, having the best performance for spectral features only of 40% and 65% for both and each correctly identified, confirming that it is not the classification system but the ISA that is responsible for most incorrect identification.

In a real polyphonic song, it is not possible to determine how many components we should use in classification. Segmenting the phrase into individual notes is certainly difficult if not impossible which means temporal features can hardly be used. It is also even more unlikely that the attack of an instrument will be useful in identification then as similarly argued in [1]. Therefore, we here use only the features derived from spectral bases from each random segment of the song and then combine the score or the likelihood that the instruments are present in the given phrase. Here, we simply perform ISA on each chunk of the tree-structured non-overlapping phrases and calculate the score which are the percentage of number of k-NN in classification at each level (could use log-likelihood if GMM was employed). Level 1 corresponds to using the whole phrase,

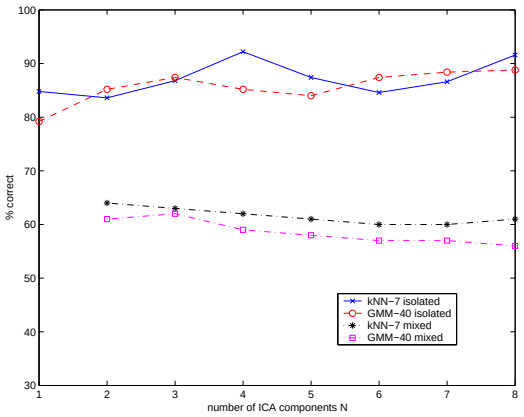

Fig. 3. Average identification rate (%) of isolated tones and two-tonal mixtures as a function of number of ISA components (N) used in classification for kNN-7 and GMM-30 classifiers

while Level 4 corresponds to segmenting the phrase into 2^4 non-overlapping chunks for independent basis learning and identification. Learning from different scales is essential to capture the note of instruments at different levels. The kNN-7 classifier using PLPC-12 and $N = 1$ on a 10-second phrase of a Piano-Flute-Violin trio playing simultaneously yields the score shown in Table 2. In Level 1, the piano which plays background arpeggio (and hence repetitive over time) is well-captured despite being softer, while at a higher level where phrases' duration are shorter, down to an individual note level, the violin can be captured. Combining the scores across levels will average out misidentification as seen in Cello and Alto-Saxophone.

5. CONCLUSION

The decomposition by ISA is shown to give physiologically intuitive features for instruments identification. The inclusion of the lesser energy components in classification can be beneficial for isolated tones but not so for the two-tonal mixtures. In real song, their roles are diminished. Only the sustain components will be learned and used in identification. A better learning algorithm is called for, for example, the non-negative matrix factorization which has been used for piano transcription in [10] might be investigated. Future works include more experiments on a larger database, pitch ranges and instruments, as well as its robustness to noise and percussive interference.

Score (%)	P	F	Bb	C	AS	FH	O	V
Level 1	43	29	0	14	0	0	0	14
Level 2	50	14	0	0	7	21	0	7
Level 3	29	32	0	7	4	11	4	14
Level 4	25	27	0	0	0	14	4	30

Table 2. k-NN score as percentage of 7-nearest neighbors at 4 levels for instruments Piano (P), Flute (F), Bb-Clarinet (Bb), Cello (C), Alto-Saxophone (AS), French-Horn (FH), Oboe (O) and Violin (V).

6. REFERENCES

[1] Martin, K., "Sound Source Recognition : A Theory and Computational Model", *Ph.D. Thesis*, MIT, 1999.

[2] Eronen A., and Klapuri A.,"Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features", *ICASSP'2000*, pp. 753-756, 2000.

[3] Eggink, J., and Brown, G.J., "A Missing Feature Approach to Instrument Identification in Polyphonic Signal", *ICASSP'03*, Vol.5, 2003, pp.553-556.

[4] Kashino, K., and Murase, H., "A Sound Source Identification System for Ensemble Music Based on Template Adaptation and Music Stream Extraction", *Speech Communication*, Vol.27, March 1999, pp.337-349.

[5] Kinoshita, T., Sakai, S., and Tanaka, H. "Musical Soundsource Identification Based on Frequency Component Adaptation". *IJCAI-99*, Stockholm, Sweden, 1999.

[6] Casey, M. and Westner, A., "Separation of Mixed Audio Sources by Independent Subspace Analysis", *ICMA'2000*, Berlin, August, 2000.

[7] Smaragdis, P., "Redundancy Reduction for Computational Audition, A Unifying Approach", *Ph.D. Thesis*, MIT, 2001.

[8] Bell, A.J. and Sejnowski, T.J., "An Information Maximization Approach to Blind Separation and Blind Deconvolution", *Neural Computation*, 7:6, 1995, pp.1129-1159.

[9] Slaney M., Auditory Toolbox version 2.0, *http://www.slaney.org/malcolm/pubs.html*.

[10] Smaragdis, P., and Brown, J.C., "Non-negative Matrix Factorization for Polyphonic Music Transcription", *WASPAA'03*, New Paltz, New York, 2003.