SEPARATION OF SINGING AND PIANO SOUNDS

Yoram Meron *Keikichi Hirose*

Department of Information and Communication Engineering, Faculty of Engineering
The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

ABSTRACT

Database driven speech and singing synthesis systems, have been shown to produce good quality sound. Automatic training methods can make synthesis systems practical for a wide range of uses. Our goal is to develop a concatenation based singing synthesis system, in which the basic synthesis units (or "singing units") are automatically extracted from existing musical recordings (and then processed and modified) in various ways. This can be useful to the synthesis of songs with voices of existing singers, restoration of old recordings etc. In practical situations, most recordings feature a singer with instrument accompaniment, thus signal separation is required. This paper concentrates on the problem of separating singer from accompaniment, for the special case of piano accompaniment.

1. INTRODUCTION

Our aim is to construct a waveform concatenation singing synthesis system, with automated training using recordings of a singer with accompaniment.

In this paper, we concentrate on the signal separation block, for the special case of piano accompaniment (widely used in classical pieces).

Previous works on co-channel separation have shown improvements in SNR, VVR (voiced/voiced ratio) and intelligibility of target signal [1][2][3][4]. There are two main problems with the existing methods :

1. most methods require the frequencies of the spectral components of both signals to be known, in order to achieve good separation,
2. separation quality degrades when the co-channel signals have close spectral components (which, in music, happens frequently).

For the intended system, reliable frequency estimates for both signals are not available (at least for the singer), and the voice quality of the separated singing must not be degraded. Existing methods do not meet these requirements.

In this paper, we use the framework of the sinusoidal modeling approach. We use the fact that one signal, the instrument, is more structured, and try to use this structure to enhance the separation. We suggest the use of further sources of information, available to this specific task:

1. Advance knowledge of the music score sheet allows an automatic alignment of the score to the piano part in the recording. As a result, the system can know the begin and end time of each piano note. This is useful for obtaining reliable frequency estimations for the piano, as well as the singer.
2. The piano sound is represented by a model, which is then used to restrict the parameter estimation during the signal separation process:
   (a) Frequency components of the piano are assumed to have fixed frequency
   (b) A parametric model for the amplitude envelopes for the piano sound harmonics is used to overcome local spectrum corruption
3. Use of a large database of the same singer and instrument allows the estimation of the parameters of the instrument model.

In order to address the two above mentioned problems (lack of prior frequency knowledge, and voice quality degradation), experiments were carried with several separation methods. Section 3 describes an experiment to test separation of human and piano sounds, when no prior knowledge, except note information, is available. In order to improve the voice quality of the separated singer, a note modeling experiment, described in section 4, was performed. Separation quality of the proposed method is compared to separation methods mentioned in the references, followed by concluding remarks.

2. EXPERIMENT ENVIRONMENT

In the experiments described here, the sinusoidal modeling approach [5] was used for signal separation. A graphic tool was constructed (Linux X-Windows), allowing display, editing (automatic and manual), manipulation and synthesis of sinusoidal components of two separate channels simultaneously.

This tool allows to extract sinusoidal components from two separate files, then mix the two files, and load both sets of components simultaneously (displayed in different colors) for the mixed file.

The recording used were sampled at 44.1 kHz, analyzed using FFT windows of length 8192, and a maximum of 100 sinusoidal components for the two signals (together).

3. USING MUSICAL SCORE

As robust pitch detection for the multiple channel case is a difficult task, a separation system can not assume to be able to automatically extract reliable frequency estimations needed for the separation. In the first experiment, we compared the separation quality in two cases: 1) when prior knowledge of both signals' frequencies are known; 2) when only the musical score for the piano (which note was played at what time) is known.

Previous work [6] showed that automatic alignment of musical score to an audio recording is possible, with relatively low error in note onset timing detection. Therefore, we assume that it is possible to automatically obtain a close estimation of note begin (and, to a lesser degree, end) times.

3.1. Prior knowledge

For reference, the methods described in [3] were implemented in the graphic tool: 1) An LSE method to compensate for spectral interference due to the use of windowing, 2) A Multi-frame Interpolation method to overcome calculation instabilities where components' frequencies of the different signals are too close to each other.

Two separate files (with similar average energy) were used - one file of a singing voice, and another file of a piano, playing the same note several times (with different amplitudes and durations). The note played was chosen to be close to the average singer pitch (to make this a difficult separation task). The graphic tool was used to extract sinusoidal components for each file separately (using the peak picking algorithm [5]). The two files were mixed (by simple waveform addition). The mixed file is then loaded, with both sets of sinusoidal components (copies of the component frequencies estimated for the two original files). Next, The LSE or Multi-frame methods are used to estimate the parameters (amplitude and phase) for the two signals.

Once the amplitude and phase parameters are estimated, the separated singing voice is created by sinusoidal components summation of the singer's estimated parameters (standard sinusoidal modeling synthesis). Another method to create the separated singing is by synthesizing the piano signal (using the parameters estimated for it from the mixed file), and then performing a waveform subtraction of the synthesized piano signal from the mixed file. This waveform subtraction of the estimated piano signal is done in a frame by frame fashion. The resulting (subtracted) frames are then joined using a waveform OLA (overlap add). The subtraction method has the advantage of not subjecting the singer voice to any kind of modeling, thus retaining the voice quality of the singer. On the other hand, this method might lead to residual piano signal, if the piano signal parameters are not correctly estimated.

3.2. Musical score

In the second case, only the musical score information is available. The same files were used in this test. Estimated piano components were forced to have fixed frequencies, and be "almost" harmonic: in the graphic tool, note begin and end times were manually marked (this could be done automatically) and an estimate of the fundamental frequency is given (this could be done manually, or directly from the music score). The system automatically adds fixed frequency tracks at multiples of the given frequency (a small pitch deviation for the *whole* track can be allowed, to best match spectrum peaks along the whole note duration). Ultimately, the frequencies of the piano's harmonics could be extracted from the whole database, making the estimate more robust.

The assumption of fixed frequency is obviously problematic, especially in the attack part, but it is used as a first approximation.

Note: in this experiment, monophonic piano sounds were used. We assume that for ultimately, there will be some parts in the training database where the required note is played monophonically, so that frequency components could be estimated from it, and used for separation in polyphonic cases as well.

In the graphic tool, the mixed file is loaded, with the piano's fixed frequency sinusoidal tracks. The amplitudes and phases of the piano signal are estimated by sampling the FFT spectrum at the components' frequencies (or by an AbS procedure similar to the one described in [7]) . The parameters are used to perform OLA subtraction. Sinusoidal components are estimated (by peak picking) for the residual file. These components are then used as an estimate for the singer components, to be used (together with the fixed frequency piano components) for separation, in the same way prior frequency knowledge is used (using LSE and possibly multi-frame interpolation, or OLA subtraction).

3.3. Results

In order to compare the separation quality, a signal distance measure was used to compare the separated signal to it's original (unmixed) recording. A spectral distance measure was used as a measure of the distance between two (energy normalized) signals. The distance was calculated as an average Euclid distance between two spectral power vectors (calculated over a window of 25.6 msec, with a 10 msec shift). Table 1 shows the results for one typical pair of signals. High distance score indicates either the remaining of piano components, or a degradation in the voice quality of the singer. For comparison, the table also shows the result for comb filtering - the piano components' amplitudes and phases were sampled from the mixed file, and then OLA subtracted from it.

Informal listening confirmed the calculated scores. Waveform level subtraction and OLA exhibited better voice quality than the synthesis of the singer from esti-

Experiment	SM-synthesis	OLA-Subtraction
Prior - LSE	0.971	0.928
Prior - Multi-frame	0.968	0.823
Score - comb		0.896
Score - LSE	1.063	0.940
Score - Multi-frame	1.071	0.827

Table 1: Distance measure, comparing separately recorded signal, and the result of several separation methods, using either prior knowledge of frequencies or only music score knowledge

mated sinusoidal components (as the singer's voice was not subjected to any kind of modeling). Multi-frame interpolation improved the distance measure for most cases.

Separation quality did not degrade significantly when prior frequency knowledge was not used. In some instances, there was even an improvement in separation. This is probably due to fact that fixed components prevent the selection of wrong components in the automatic peak picking process for the piano. The attack part of the piano sound is still problematic for both cases. On one hand, the noisy character of the attack sound causes the peak picking process to pick many components, which then mask the singer. On the other hand, the fixed component approach does not try to include these components, thus leaving them in the residual file.

4. NOTE MODELING

In order to improve separation when singer and piano components are close, modeling of the piano notes was attempted. The modeling is performed automatically from the audio recording, using note and timing information from the musical score. This model is used to constrain the piano components, so resolution of close components can be improved.

4.1. Model structure and estimation

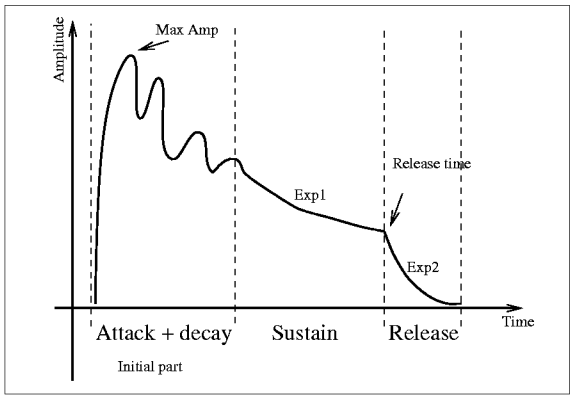

Figure 1: Note model structure

The note model, models the note amplitude envelope, and is based on the ADSR (Attack, Decay, Sustain, Release) method, which is often used in sampling-based musical synthesizers (see 1). In this model, notes are temporally divided to three parts : 1) An initial part (unmodified except amplitude scaling), 2) a sustained part, where amplitude envelope drops exponentially (with rate *exp1*) (this part can be skipped for short notes) 3) a final part, where amplitude decays exponentially with a faster rate (*exp2*).

The estimation of the model parameters was performed on a separate recording of repetitions of the same note, in several amplitudes and durations. For each fixed frequency component, the amplitude envelope is extracted, and an AbS process is performed to find the best model parameters. The varied parameters (for each harmonic) are: duration of the initial part, and the two exponential decay parameters. The set of parameters synthesizing the envelopes closest (in a least mean squares sense) to the measured envelopes (of all training notes), is chosen. Model parameters for each sinusoidal components (for each note) are saved to a file.

4.2. Note Instance Parameter Estimation

To use this model, only two parameters need to be estimated for any given note instance - *a)* amplitude *b)* release time (assuming note start and end time are approximately known). These two parameters are assumed common to all the note's sinusoidal components, so even if some harmonics are hidden by the singer's components, the parameters can still be estimated. Further, even if in some part of the note all components of piano and singer overlap, the piano note parameters can be estimated. In the singing example we used (classic singer), the singer's voice held a fixed frequency (which would completely overlap the piano's components) only for a short duration, relative to the note duration.

Estimation of the parameters (amplitude and release) for a given note instance, is done by an AbS process, trying to find the parameter set which will give a maximum match of measured and modeled amplitude envelopes over all frequency tracks. In this process, a lower weight is given to the match score where the singer's components interfere:

$$score_{A,R} = \sum_{t=1}^{T} \sum_{i=1}^{N} W(i,t) * (MD(i,t)_{A,R} - MS(i,t))^2$$

where T is the note duration, N is the number of piano components, A and R stand for the tested amplitude and release time, W gives a weight for each time and component (see below), MD is the amplitude calculated by the model (using A and R) for component number i at time t, and MS is the measured amplitude.

For calculating the weight function, first, the singer's components' frequencies and amplitudes are estimated (as described in subsection 3.2 - by peak picking from the residual signal after the piano's components are removed[1]).

[1]The residual signal might not have components in overlapping frequencies, thus distorting the weights. In principle, it is possible to repeat the process - use the amplitude calculated

For each piano component at each time frame $P(i,t)$, the closest singer component (within the main lobe of the windowing function) is found, $S_P(i,t)$, and the weight $W(i,t)$ is set proportional to the frequency distance between them, and inversely proportional to the amplitude of $S_P(i,t)$. In this way, the weight is higher where interference from singer components is small.

After the weights are set, the AbS process finds the best parameter amplitude and release time. If note begin time can not be accurately detected, the AbS process can also try to find the best time shift to the beginning point (this was confirmed by an experiment where note begin time mark was shifted).

The parameters found by the AbS are used to synthesize the amplitude envelope for all piano components, which are then OLA subtracted.

Experiment	OLA-Subtraction
Using original phase	0.555
Using phase sampled from mix	0.769

Table 2: Separation results, using piano note model. Phase information extracted either from separate piano file or from mixed file

Table 2 shows the distance measure results for separation using the note model approach. The spectral distance measure shows improvement relative to the previous experiments. Subjective listening also confirms an improvement in the separated voice quality. The model does not include any phase information, which causes it to be affected by phase estimation errors, which degrade separation quality. Phase continuity constraints may help improve this.

5. CONCLUSION

A method for the separation of singer and piano sounds, using musical score knowledge was presented. The method relies on the existence of a large database, for the construction of note models. Using note models increases the robustness of separation against an interfering signal, partly overlapping with the desired signal in the spectral domain.

Experiments showed that the note modeling method can improve separation, while preventing voice quality degradation.

Further work is needed to remove the in-harmonic components of the attack part of piano notes, which can not be modeled by a harmonic model.

6. REFERENCES

1. P.W. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection", *JASA 60*, 1976, pp. 911-918
2. B.A. Hanson, D.Y. Wong, "The Harmonic Magnitude Suppression Technique for Intelligibility Enhancement in the Presence of Interfering Speech", *ICASSP 84*, pp. 18A.5.1-4
3. T.F. Quatieri, R.G. Danisewicz, "An Approach to Co-Channel Talker Interference Suppression Using a Sinusoidal Model for Speech", *IEEE Trans. ASSP*, V.38, N.1, Jan 1990, pp 56-69
4. F.M. Silva, L.B. Almeida, "Speech Separation by Means of Stationary Least Squares Harmonic Estimation", *ICASSP* 90, pp 809-812
5. T.F. Quatieri, R.J. McAulay, "Shape Invariant Time Scale and Pitch Modification of Speech", *IEEE trans. Signal Processing*, Vol. 40, No. 3, March 1992, pp 497-510
6. E.D. Scheirer, "Extracting Expressive Performance Information from Recorded Music", *MSc thesis*, MIT 1995
7. E.B. George, M.J.T. Smith, "Speech Analysis / Synthesis and Modification Using an Analysis by Synthesis / Overlap Add Sinusoidal Model", *IEEE trans. speech and audio processing*, Vol. 5, No. 5, Sept. 1997, pp 389-406

by the initial model for the subtraction, and find the singer's component's from the new residual signal. In practice, as exact overlaps were short, this had only a minimal effect on the AbS results