SPATIAL SEPARATION OF SPEECH SIGNALS USING CONTINUOUSLY-VARIABLE MASKS ESTIMATED FROM COMPARISONS OF ZERO CROSSINGS

Hyung-Min Park and Richard M. Stern

Language Technologies Institute and Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, U.S.A.
{*hmpark, rms*}*@cs.cmu.edu*

ABSTRACT

This paper describes an algorithm that achieves noise robustness in speech recognition by reconstructing the desired signal from a mixture of two signals using continuously-variable masks. In contrast to current methods which use binary masks, this approach estimates the relative contribution of the desired source in a mixture of sources and reconstructs the desired signal in proportion to its estimated contribution to each time-frequency segment. Estimation of the continuously-variable masks is based on the relationship between the relative intensity of each source and the interaural time difference (ITD). Estimation of the ITD is accomplished using zero-crossing-based methods. It is shown that the use of zero-crossing approaches to estimate ITDs and continuously-variable masks provide better speech recognition accuracy than cross-correlation-based approaches to ITD estimation and binary masks.

1. INTRODUCTION

Noise robustness remains a very important issue in the field of automatic speech recognition (ASR). While high recognition accuracy can be obtained in a controlled and noise-free acoustic environment, recognition accuracy is seriously degraded in a more realistic noisy environment. On the other hand, humans can understand speech even if it is subject to interference by competing speech or other noises. This observation has motivated the development of many types of signal processing approaches based on aspects of human auditory perception (*e.g.* [1]).

In his treatise on auditory scene analysis (ASA), Bregman [2] identified many cues that are used by humans to segregate a target sound from interfering sources (*e.g.* [2]), including the spatial location of sound sources. The primary acoustical cues for human sound localization are interaural time differences (ITDs) and interaural intensity differences (IIDs). ITDs serve as the localization cue primarily at frequencies below 1.5 *k*Hz [3], although the ITDs of the low-frequency envelopes of higher-frequency components of a sound can also be useful for sound localization. Many computational models have been developed to describe and predict binaural processing [4, 5, 6]. Most of these include a model of peripheral auditory processing which includes frequency analysis and subsequent nonlinear operations (*e.g.* [7]), a mechanism for estimating the interaural cross-correlation function on a frequency-by-frequency basis, and a mechanism to disambiguate the temporal analysis, typically exploiting the IID of the signal and consistency over frequency. These models have been incorporated into several systems that perform automatic speech recognition (*e.g.* [8, 9, 10]) . In recent work Kim *et al.* estimated ITDs by comparison of the zero crossings of the outputs of a bank of Gammatone filters rather than from the peaks of the cross-correlation function of these outputs [11]. They demonstrated that individual sounds from mixtures could be successfully reconstructed by selecting specific time-frequency segments on the basis of their estimated ITDs and IIDs.

So far, most algorithms for sound segregation and noise-robust speech recognition within the framework of ASA have used binary "masks" that specify which time-frequency components belong to a particular sound source on an "all-or-none" basis (*e.g.* [9, 10, 11]). This is clearly an oversimplification of how sound sources are combined. In this paper, we describe a method to estimate masks that represent continuously-variable estimates of the extent to which a desired sound source contributes to a particular time-frequency segment, based on estimates of ITD derived from zero crossings. We subsequently develop an estimate of that source in isolation by combining the time-frequency segments in proportion according to the weighting function. The use of a continuously-variable weighting function rather than a binary mask provides smoother transitions between segments that correspond to changes in the extent to which the desired signal is dominant, which provides improved speech recognition accuracy. In addition to comparing the effect of binary and continuously-variable masks, we also compare the use of zero-crossing-based and cross-correlation-based methods of ITD estimation in terms of the sample standard deviation of the ITD estimates and the resulting speech recognition accuracy.

While our work is based on the estimation of ITD information from a pair of microphones, we do not consider our approach to be closely based on models of auditory perception, despite the use of terms like ITD from the binaural hearing literature. Specifically, we make use of sensors (*i.e.* microphones) that are closely spaced to reduce the effects of spatial aliasing, rather than sensors that are spaced at approximately the distance between the two ears, and we assume zero IID between the signals to the sensors.

2. SPEECH ENHANCEMENT USING ZERO CROSSINGS

The zero-crossing approach is popular because it provides a convenient way to represent the synchronous response of low-frequency auditory-nerve fibers to the fine structure of a sound source. The ITD estimation algorithm used is very simple: zero-crossing points are detected for the signal arriving at one sensor, and a search is

This work was supported by the Information and Telecommunication National Scholarship Program sponsored by the Institute of Information Technology Assessment, Korea, and by the National Science Foundation (Grant IIS-0420866).

performed for the closest zero crossing of the signal from the other sensor. The difference in time between the observed zero crossings is regarded as a valid estimate of the ITD if it is smaller in magnitude than the time it takes for sound to travel from one sensor to the other, and it is discarded otherwise. The actual zero crossings are estimated by linear interpolation between the two points that straddle the zero crossings. As mentioned above, the two sensors are sufficiently close that there is no spatial aliasing for frequencies up to half the sampling frequency. Hence, the largest possible delay between the sensors is always smaller than half a period over all frequencies of interest, so the closest zero crossings provide the desired ITD value.

The mixture of speech signals from each sensor is first subjected to frequency analysis, typically accomplished by passing the mixture through a bank of Gammatone filters. Once the ITDs are estimated in each frequency band according to the procedure above, they must be converted into continuously-variable masks. Mask estimation is accomplished by deriving a relationship between the ITDs and the relative contribution of each source to the mixture. Let us approximate the outputs of the narrow bandpass filters as pure tones, one from each source as follows:

$$\begin{aligned} x_1(t) &= A_1\cos(\omega_1 t+\phi_1)+A_2\cos(\omega_2 t+\phi_2), \\ x_2(t) &= A_1\cos(\omega_1(t-d_1)+\phi_1) \\ &\quad +A_2\cos(\omega_2(t-d_2)+\phi_2), \end{aligned} \tag{1}$$

where A_i, w_i, d_i, and ϕ_i denote the amplitude, the frequency, the delay, and the phase for the i^{th} source, respectively. As noted above, the component amplitudes at the two sensors are assumed to be equal because of the closeness of the sensors. Without loss of generality, we assume that a zero crossing for x_1 occurs at time t_1, and that the nearest zero crossing for x_2 occurs at $t_1+\tau$, producing an ITD of τ.

We assume that $\omega_i(\tau-d_i)$ is small, which is especially valid at low frequencies, so $x_2(t_1+\tau)$ can be approximated by

$$\begin{aligned} x_2(t_1+\tau) &\approx -A_1\sin(\omega_1 t_1+\phi_1)\cdot\omega_1(\tau-d_1) \\ &\quad -A_2\sin(\omega_2 t_1+\phi_2)\cdot\omega_2(\tau-d_2). \end{aligned} \tag{2}$$

Since $x_2(t_1+\tau)=0$,

$$\begin{aligned} &\tau(A_1\omega_1\sin(\omega_1 t_1+\phi_1)+A_2\omega_2\sin(\omega_2 t_1+\phi_2)) \\ &\quad\approx A_1\omega_1\sin(\omega_1 t_1+\phi_1)\cdot d_1+A_2\omega_2\sin(\omega_2 t_1+\phi_2)\cdot d_2. \end{aligned} \tag{3}$$

Since $x_1(t_1)=A_1\cos(\omega_1 t_1+\phi_1)+A_2\cos(\omega_2 t_1+\phi_2)=0$ and τ is obtained from the nearest zero-crossing point,

$$\tau\approx\begin{cases} \frac{\omega_1\sqrt{A_1^2-A_2^2\cos^2(\omega_2 t_1+\phi_2)}\cdot d_1+A_2\omega_2|\sin(\omega_2 t_1+\phi_2)|\cdot d_2}{\omega_1\sqrt{A_1^2-A_2^2\cos^2(\omega_2 t_1+\phi_2)}+A_2\omega_2|\sin(\omega_2 t_1+\phi_2)|} & \text{if } A_1\geq A_2, \\ \frac{A_1\omega_1|\sin(\omega_1 t_1+\phi_1)|\cdot d_1+\omega_2\sqrt{A_2^2-A_1^2\cos^2(\omega_1 t_1+\phi_1)}\cdot d_2}{A_1\omega_1|\sin(\omega_1 t_1+\phi_1)|+\omega_2\sqrt{A_2^2-A_1^2\cos^2(\omega_1 t_1+\phi_1)}} & \text{otherwise.} \end{cases} \tag{4}$$

Through simulations and further derivations we realized that when the frequencies ω_i are uniformly distributed over a narrow band with $\omega_1\approx\omega_2$ and the phases ϕ_i are uniformly distributed over the interval $(-\pi,\pi)$, $\bar{\tau}$, the mean of the estimated ITDs, can be approximated by

$$\bar{\tau}\approx g(A_1,A_2)\cdot d_1+(1-g(A_1,A_2))\cdot d_2, \tag{5}$$

where

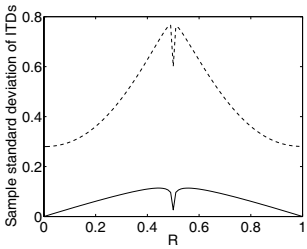
$$g(A_1,A_2)=\begin{cases} \frac{\pi(A_1^2-\frac{A_2^2}{2})-A_1^2\arctan\left(\frac{A_2}{\sqrt{A_1^2-A_2^2}}\right)-A_2\sqrt{A_1^2-A_2^2}}{\pi(A_1^2-A_2^2)} & \text{if } A_1>A_2, \\ \frac{1}{2} & \text{if } A_1=A_2, \\ \frac{-\frac{\pi}{2}A_1^2+A_2^2\arctan\left(\frac{A_1}{\sqrt{A_2^2-A_1^2}}\right)+A_1\sqrt{A_2^2-A_1^2}}{\pi(A_2^2-A_1^2)} & \text{otherwise} \end{cases} \tag{6}$$

Using Eqs. (5) and (6), one can easily relate the estimated ITDs to the relative amplitudes of the sources in a mixture. Note that Eq. (5) describes a monotonic function suitable for one-to-one mapping and does not depend on any frequency-specific parameters so the same function can be used in all bands. In each frequency band, this processing results in a series of ITDs estimated at each zero-crossing point, along with a set of estimated weighting factors for the desired signal corresponding to each zero crossing. The desired signal is reconstructed by multiplying the output of each bandpass analysis filter by a piecewise-linear amplitude modulation function that passes through the values of the weighting coefficients for the desired signal at each zero crossing point. A time-reversed version of the enhanced signal in each band is then convolved with the corresponding Gammatone filter to negate any across-frequency phase effects, and the outputs are time reversed again and summed across frequency.

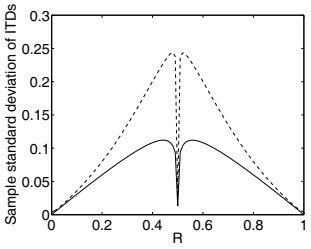
3. COMPARISON OF ITD-ESTIMATION METHODS

In the previous section, a method for developing continuously-variable mask estimates based on zero crossings was described. Since ITD estimation has traditionally been accomplished using cross-correlation, we compare the zero-crossing-based and cross-correlation-based methods in terms of the sample standard deviation of the resulting ITD estimates by simulation. For each value of the relative signal strength $R=A_2/(A_1+A_2)$, we obtained 100,000 samples of ITDs using Eq. (4). For each sample, the phases ϕ_i and frequencies ω_i were varied randomly over the intervals $(-\pi,\pi)$ and $(\omega_{cf}-BW/2,\omega_{cf}+BW/2)$, respectively, where ω_{cf} and BW are the center frequencies and bandwidths of the frequency channels. The simulations were obtained with parameter values $d_1=0$, $d_2=1/16,000$ sec, and a frame length $2T=25.6$ msec.

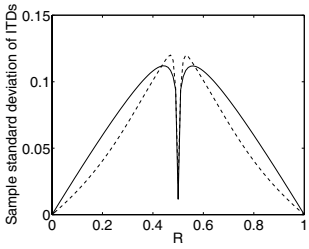
Fig. 1 shows the sample standard deviation of ITD estimates as a function of the relative strength R for three different frequency bands. To present corresponding results using a cross-correlation method, we also derive a function relating the amplitudes A_i to the ITDs based on cross-correlation in the Appendix. In the lowest frequency band, the zero-crossing-based ITD estimation method provided estimates of smaller variability than the cross-correlation-based estimate. The difference in sample standard deviations was reduced in the middle frequency band, and cross-correlation-based processing was slightly better in the highest frequency band for most values of R. It is worth noting that the zero-crossing-based method estimates an ITD at almost every zero crossing based on local signal information, while estimation of the cross-correlation function requires a greater length of input samples. In high frequency bands, the bandpass-filterbank outputs encounter zero crossings very frequently, and the duration needed for reasonable cross-correlation estimates may include hundreds of zero crossings. Because each estimate of the contribution of the desired signal based on zero crossings affects the reconstructed output over only a small neighborhood of samples, the impact of errors in estimating the weight of



(a) The lowest frequency band
(Center frequency CF: 170 Hz, Bandwidth BW: 43.87 Hz)



(b) The middle frequency band
(CF: 1,343 Hz, BW: 172.9 Hz)



(c) The highest frequency band
(CF: 6,430Hz, BW: 732.4Hz)

Fig. 1. Sample standard deviation of the estimated ITDs as a function of the relative strength R. They were computed when R varied with the step 0.01 from 0 to 1. The solid and dashed lines correspond to results obtained using the zero-crossing-based and cross-correlation-based methods, respectively. The vertical axis is normalized by dividing by d_2.

the desired contribution using the zero-crossing method is brief in duration, while errors in estimating the weights using the cross-correlation method affect sample points over a longer duration.

4. SPEECH RECOGNITION EXPERIMENTS

We evaluated the proposed method of signal enhancement by speech recognition experiments using the DARPA Resource Management (RM1) database [12] and the CMU SPHINX-III speech recognition system, which is based on fully-continuous hidden Markov models. Using 13^{th}-order mel-frequency cepstral coefficients as features, 2,880 RM1 sentences recorded in a quiet environment were used to train the recognition system, and 600 sentences were decoded to give a word error rate (WER). Each test utterance was corrupted by adding another interfering speech signal which had the same energy as the test utterance. To simulate the measurement of signals that would be obtained by microphones in close proximity, the target and interfering speech were combined with different simulated delays from sensor to sensor. The delays were created by upsampling the original speech signals (which were recorded at a sampling rate of 16 kHz), by a factor 4. The signals were combined with independent delays inserted on one side which were selected independently and randomly for the target and masker in the range of -3 to 3 samples at 64 kHz, except that delays for the target and masker were forced to be different from one another. As a result, the net difference in ITD of target and masker ranged from 1 to 6 samples or 15.6 to 93.8 μs. If the microphones are separated by about 21 mm to avoid spatial aliasing at 8 kHz, this would correspond to differences in azimuth of between about 14.9 and 100.7 degrees. The target signal was always assumed to be the component with the more positive delay. After combining the target and interfering speech, the resulting signals were downsampled back to 16 kHz.

Table 1. Comparison of the percentage WER obtained for the CMU SPHINX-III speech recognition system using continuously-variable masks versus binary masks and ITD estimation based on zero crossings versus cross-correlation. Results are also shown with no interfering masker, and with identical or statistically-independent white Gaussian noise added to the two mixtures. The frame size was 25.6 msec, and the frame rate was 10 msec.

added white Gaussian noise	target alone	target plus masker				
		no proc.	zero crossing		cross-corr.	
			binary	cont.	binary	cont.
none	7.3	90.2	36.7	11.8	88.3	23.0
independent	23.8	96.9	48.0	44.8	94.1	69.0
identical			45.6	24.0	89.5	39.3

We used a 40-channel bank of Gammatone filters with center frequencies spaced linearly in ERB (equivalent rectangular bandwidth) from 170 Hz to 6,430 Hz. Assuming that the actual delays for the desired target and interfering speech were known *a priori*, the ITDs of the combined signal at each frequency were converted into estimates of the relative strength of the target and masker according to Eqs. (5) and (6). For the cross-correlation case, we adopted Roman's approximation given by

$$\bar{\tau}_{cc} = \frac{d_1 + d_2}{2} + \frac{1}{w_{cf}} \left\{ \arctan \left[\frac{(A_2^2 - A_1^2)}{(A_1^2 + A_2^2)} \tan\beta \right] + k\pi \right\}, \quad k \in \{0, \pm 1\}, \quad (7)$$

where $\beta = w_{cf} \cdot (d_2 - d_1)/2 \in [0, \pi]$. If $\beta \leq \pi/2$, $k = 0$. Otherwise, $k = 1$ when $A_1 < A_2$ and $k = -1$ when $A_1 > A_2$ [10]. This corresponds to the solution when $\omega_1 = \omega_2$ and T goes to infinity. The time lag corresponding to the maximum of the cross-correlation function was estimated by differentiating a 20^{th}-order polynomial approximation to the cross-correlation function in a region around the observed discrete-time maximum within the time lags that are less than the delay time between the microphones and finding the root of the derivative closest to the discrete-time maximum using Newtonian iteration.

Table 1 presents the WERs obtained using mixtures of the target and interfering speech combined as described above. The WERs of clean target speech in the absence of an interfering signal are also included for comparison. It is seen that the use of the continuously-variable masks provides much better recognition accuracy than that of binary masks, and that estimation of ITDs using zero crossings is more effective than estimation obtained using the cross-correlation-based method. To obtain a crude characterization of the robustness of the methods considered we also obtained the WERs of the same signals in the presence of white Gaussian noise at a signal-to-noise ratio of 20 dB. This noise was presented both identically to the two sensors (which would correspond to an additional non-speech source of interference) and independently to the two sensors (which would correspond to sensor or measurement noise). The WER is affected adversely by the addition of noise in all cases, but the superiority of continuously-variable masks and estimation of ITD by zero crossings remains evident.

5. CONCLUSIONS AND FURTHER WORK

In this paper we have described a method that estimates continuously-variable masks for recognizing speech in the presence of interfering speech. Our approach estimates the relative contribution of the desired speech in mixtures by a derived mapping function between the masks and ITDs so that the recognizer can decode speech features successfully. In addition, estimation of ITD from zero crossings provides estimates with less variability than estimates obtained using cross-correlation, especially at low frequencies. The use of the continuously-variable masking weights and ITD estimation based on zero crossings provided good improvement in recognition accuracy for speech in the presence of masking speech when the target and masker arrive with different ITDs. While these results are promising, the approach remains to be extended to more difficult scenarios such as reverberant environments and multiple interfering speech sources.

6. REFERENCES

[1] H. Hermansky, "Should recognizers have ears?," *Speech Comm.*, vol. 25, no. 1-3, pp. 3–27, 1998.

[2] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, Cambridge, MA, 1990.

[3] J. W. Strutt (Third Baron of Rayleigh), "On our perception of sound direction," *Philosoph. Mag.*, vol. 13, pp. 214–232, 1907.

[4] J. Braasch, "Modelling of binaural hearing," in *Communication Acoustics*, J. Blauert, Ed., chapter 4, pp. 75–108. Springer-Verlag, Berlin, 2005.

[5] H. S. Colburn and A. Kulkarni, "Models of sound localization," in *Sound Source Localization*, R. Fay and T. Popper, Eds., Springer Handbook of Auditory Research. Springer-Verlag, 2005.

[6] R. M. Stern and C. Trahiotis, "Models of binaural perception," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. Gilkey and T. R. Anderson, Eds., chapter 24, pp. 499–531. Lawrence Erlbaum Associates, 1996.

[7] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification," *J. Acoust. Soc. Am.*, vol. 89, no. 6, pp. 2866–2882, 1991.

[8] M. Bodden, "Modelling human sound-source localization and the cocktail party effect," *Acta Acustica*, vol. 1, pp. 43–55, 1993.

[9] K. J. Palomäki, G. J. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Comm.*, vol. 43, pp. 361–378, 2004.

[10] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[11] Y.-I. Kim, S. J. An, R. M. Kil, and H.-M. Park, "Sound segregation based on binaural zero-crossings," in *Interspeech*, 2005, pp. 2325–2328.

[12] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallet, "The DARPA 1000-word resource management database for continuous speech recognition," in *ICASSP*, 1988, pp. 651–654.

A. APPENDIX

In order to estimate the variability of estimates of ITD based on cross-correlation, we follow almost the same procedure as in the case of the zero-crossing-based method. As before, we start from mixtures given by Eq. (1).

For finite-duration signals in a frame of duration $2T$, the cross-correlation function with time lag τ is expressed by

$$c(\tau) = \frac{1}{2T}\int_{t=-T}^{T} x_1(t)x_2(t+\tau)dt. \tag{A.1}$$

Since

$$\begin{aligned}
&x_1(t)x_2(t+\tau)\\
&= \tfrac{1}{2}[A_1^2(\cos(2\omega_1 t+\omega_1(\tau-d_1)+2\phi_1)+\cos(\omega_1(\tau-d_1)))\\
&\quad +A_1A_2(\cos(\omega_1 t+\omega_2(t+\tau-d_2)+\phi_1+\phi_2)\\
&\quad +\cos(\omega_1 t-\omega_2(t+\tau-d_2)+\phi_1-\phi_2)\\
&\quad +\cos(\omega_1(t+\tau-d_1)+\omega_2 t+\phi_1+\phi_2)\\
&\quad +\cos(\omega_1(t+\tau-d_1)-\omega_2 t+\phi_1-\phi_2))\\
&\quad +A_2^2(\cos(2\omega_2 t+\omega_2(\tau-d_2)+2\phi_2)+\cos(\omega_2(\tau-d_2)))],
\end{aligned} \tag{A.2}$$

the derivative of the cross-correlation function $c(\tau)$ is

$$\begin{aligned}
&\frac{dc(\tau)}{d\tau}\\
&= -\tfrac{1}{4T}[A_1^2(\sin(2\omega_1 T)\sin(\omega_1(\tau-d_1)+2\phi_1)\\
&\quad +2\omega_1 T\sin(\omega_1(\tau-d_1)))\\
&\quad +A_1A_2(\tfrac{2\omega_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T)\sin(\omega_2(\tau-d_2)+\phi_1+\phi_2)\\
&\quad +\tfrac{2\omega_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T)\sin(\omega_2(\tau-d_2)-\phi_1+\phi_2)\\
&\quad +\tfrac{2\omega_1}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T)\sin(\omega_1(\tau-d_1)+\phi_1+\phi_2)\\
&\quad +\tfrac{2\omega_1}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T)\sin(\omega_1(\tau-d_1)+\phi_1-\phi_2))\\
&\quad +A_2^2(\sin(2\omega_2 T)\sin(\omega_2(\tau-d_2)+2\phi_2)\\
&\quad +2\omega_2 T\sin(\omega_2(\tau-d_2)))],
\end{aligned} \tag{A.3}$$

where we assume $\omega_1 \neq \omega_2$ since they will be random variables uniformly distributed within the bandwidth of a band. Using the approximation of small $\omega_i(\tau-d_i)$, the ITD at the maximum of cross-correlation can be obtained from

$$\begin{aligned}
&\tau[A_1^2(\sin(2\omega_1 T)\omega_1\cos(2\phi_1)+2\omega_1^2 T)\\
&\quad +\tfrac{2A_1A_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T)(\omega_1^2+\omega_2^2)\cos(\phi_1+\phi_2)\\
&\quad +\tfrac{2A_1A_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T)(\omega_1^2+\omega_2^2)\cos(\phi_1-\phi_2)\\
&\quad +A_2^2(\sin(2\omega_2 T)\omega_2\cos(2\phi_2)+2\omega_2^2 T)]\\
&\approx d_1[A_1^2(\sin(2\omega_1 T)\omega_1\cos(2\phi_1)+2\omega_1^2 T)\\
&\quad +\tfrac{2A_1A_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T)\omega_1^2\cos(\phi_1+\phi_2)\\
&\quad +\tfrac{2A_1A_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T)\omega_1^2\cos(\phi_1-\phi_2)]\\
&\quad +d_2[\tfrac{2A_1A_2}{\omega_1+\omega_2}\sin((\omega_1+\omega_2)T)\omega_2^2\cos(\phi_1+\phi_2)\\
&\quad +\tfrac{2A_1A_2}{\omega_1-\omega_2}\sin((\omega_1-\omega_2)T)\omega_2^2\cos(\phi_1-\phi_2)\\
&\quad +A_2^2(\sin(2\omega_2 T)\omega_2\cos(2\phi_2)+2\omega_2^2 T)]\\
&\quad -[A_1^2\sin(2\omega_1 T)\sin(2\phi_1)+2A_1A_2\sin((\omega_1+\omega_2)T)\sin(\phi_1+\phi_2)\\
&\quad +2A_1A_2\sin((\omega_1-\omega_2)T)\sin(\phi_1-\phi_2)+A_2^2\sin(2\omega_2 T)\sin(2\phi_2)].
\end{aligned} \tag{A.4}$$