A SIMPLE ARCHITECTURE FOR USING MULTIPLE CUES IN SOUND SEPARATION

William S. Woods, Martin Hansen, Thomas Wittkop, and Birger Kollmeier

AG Medizinische Physik, Carl von Ossietzky-Universität Oldenburg
D-26111 Oldenburg
Germany

ABSTRACT

The present work concerns a system aimed at enhancing a target talker under varying signal conditions based on the use of several different types of information or "cues". Toward this end, an architecture designed to combine separately operating estimators is described and evaluated. The architecture is currently implemented using spatial- and periodicity-based enhancement algorithms, and evaluated using a male target talker and female jammer talker under several spatial and target-to-jammer ratio (TJR) conditions. Using a TJR estimation algorithm, the implementation is shown to yield improved TJR under all tested input TJRs (-4, 0, 4, and 8 dB) and spatial conditions (target and jammer straight ahead; target ahead and jammer at 60 degrees). Improvement ranges from 1.4 to 4.5 dB.

1. INTRODUCTION

Existing techniques to improve the signal-to-noise ratio (SNR) of a target speech signal under common environmental conditions are limited by two main factors. These factors are the variability of these common conditions, and the level of complexity of current systems designed to deal with such variable conditions. Common conditions might include a moving target talker, or the presence of non-target sound sources varying in type or number. Such variability precludes wide-ranging success for systems designed to meet any one of such situations, since the systems will fail when their specific design conditions are not met. On the other hand, existing systems designed to face wide-ranging conditions [2], [8] are, in a sense, over-designed. These "blackboard-based" artificial intelligence systems are designed to handle the combinatorial explosion represented by the many possible received acoustic signals given varying source conditions. Such systems "... require a great deal of knowledge and flexibility." ([2], p. 206). It is questionable whether systems of such complexity are actually required in order to obtain target talker SNR improvement under a wide range of common conditions. The current work addresses this question.

The system being investigated in the current work falls between the two extremes outlined above. It is similar to the blackboard-based systems in that it uses more than one algorithm for SNR improvement simultaneously and reacts to changes in signal conditions, but it does so with a greatly reduced system complexity. The reduction is achieved through the use of a fixed number of algorithms operating in parallel, each designed for enhancement under particular conditions and continually providing an estimate for the spectrum of a single target sound, and through the tracking of parameters related to the target only. This greatly simplifies the system steering, essentially reducing it to the making of "fuzzy" decisions (explained below) concerning the quality of the target parameter estimates. Although these simplifications may limit the range of success of the system, it is believed that the limited range overlaps significantly with common operating conditions. The present paper describes an initial implementation of such a simplified system and processing results obtained in certain of these conditions.

2. SYSTEM DESCRIPTION

The current implementation, schematized in Figure 1, uses estimators based on spatial and periodicity information (derived from [6] and [10], respectively) as preliminary estimators. (Much of the present system remains unchanged if these two algorithms are replaced by others using the same cues.) Spectral estimates (S, P_L from the preliminary estimators and P_R in Fig. 1) are linearly combined based on measures of the "confidence" C_s and C_p each estimator gives to its estimate. In the current implementation,

$$E(n) = \frac{A(n) + S(n) \cdot \alpha_s \cdot C_s(n) + P(n) \cdot \alpha_p \cdot C_p(n)}{1. + \alpha_s \cdot C_s(n) + \alpha_p \cdot C_p(n)}, \quad (1)$$

where E is the frequency-dependent final estimate, n is the frame index, A is the STFT-amplitude of the received signal (left or right channel), C_s and C_p are the confidences in the spatial and periodicity frequency-dependent estimates S and P, respectively, and α_s and α_p are constants (used to suppress the input STFT-amplitude contribution when confidence is high in either of the preliminary estimates). The confidences are scalars ranging over [0, 1] (representing the result of a fuzzy decision, from "no confidence" to "complete confidence", concerning estimate quality – cf. [4]), and are measures derived from the inputs to the estimators. These

measures (detailed below) take into account how well the received signal conditions meet the assumptions, and values for the operating parameters, of the estimators, and how confident the system is in these values. These values are updated using the final, combined, estimate (as "attributes" of the "perceived" sound source) and fed back to the preliminary estimators for use in the next processing frame. Each parameter estimate also receives a confidence value.

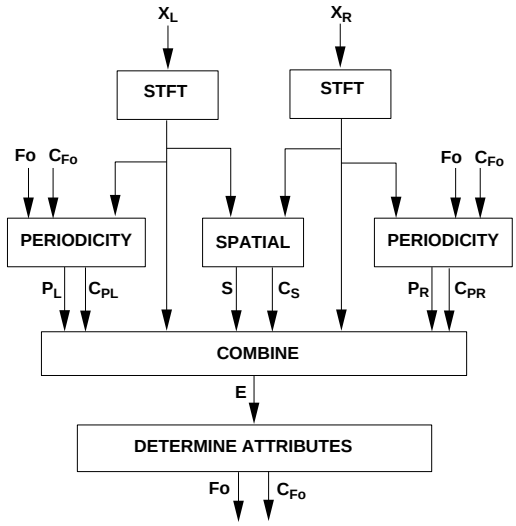

Figure 1: Schematic of the current separation system implementation.

As with any enhancement system, the present system must have some a priori knowledge concerning the target. This knowledge is embodied in the updating of attribute confidences. That is, assumptions are made concerning probable initial states of a target, and attribute confidences are updated based on this information. For example, in the evaluations presented here, the target talker is assumed to be positioned such that it elicits equal signals from two spatially displaced receivers (i.e., the target is "straight ahead"), and assumed to have a "male" pitch range. Confidence in the "pitch" target attribute Fo can only increase when the final estimate has significant periodicity in the expected range and from the expected direction. Note that the system admits operation under conditions in which these attributes vary over time. This is accomplished through the frame-by-frame update of the parameter values from the final estimate. This use of a priori expectations does not represent a limitation of the separation system per se. As mentioned, every separation system has to have some specification of what its "target" is.

2.1. Details of Current Implementation

As already mentioned, the system is evaluated assuming the target is always straight ahead. Thus, the spatial estimator always gives a unity confidence to its estimate. This estimator ([6]) operates in the frequency domain, applying increasing attenuation to frequency bins with higher inter-receiver phase or level differences and passing bins through unattenuated if these differences are small.

The fundamental-frequency-based estimator produces its estimate using the cepstrum of the received signal (cf. [10]). This estimator determines the cepstral position with highest cepstral value within a window around the expected fundamental frequency Fo, windows that position, and then inverts the windowed cepstrum. The confidence C_p in this estimate is

$$C_p(n) = g_C[y(n)] \cdot C_{Fo}(n), \tag{2}$$

where $g_C[\cdot]$ represents a two-parameter sigmoidal function ranging over [0, 1] (the subscript on g indexes the two-valued set of parameters required), y is the peak height relative to its neighbors, and C_{Fo} is the confidence in the input Fo (see Fig. 1). Qualitatively expressed, C_p is high (close to unity) if the system is confident in expected Fo and there is significant energy in the received signal with fundamental at or near Fo.

Parameter values ("attributes") for use in the next frame are computed after determination of the final estimate (see Eq. 1). The current system requires only two attribute values – the position of the "perceived" source, and its fundamental frequency. The former is determined from low-frequency inter-receiver phase differences, and the latter determined using the cepstrum of the final estimate. An initial estimate p of the fundamental frequency Fo is found by searching the cepstrum for a peak within the expected ("male") range. The perceived position $posn$ is then calculated as a weighted average of the inter-receiver phase differences across frequency for bins below about 1200 Hz, with weighting proportional to the amplitude in the final estimate in the given bin and a factor varying with the estimated fundamental frequency and its peak height in the cepstrum. This is done to ensure that the tracked source is coming from the expected direction (cf. [11]). Confidence $posnc$ in the $posn$ attribute is determined from the weighted variance of these phase differences. The final pitch estimate $Fo(n)$ and confidence $C_{Fo}(n)$ are updated according to

$$pc(n) = \{1. - C_{Fo}(n-1) \cdot g_C[z(n)] \cdot g_{\Delta p}[\Delta p(n)]\} \cdot h_R[R(n)] \cdot h_{posn}[posn(n) \cdot posnc(n)], \tag{3}$$

$$C_{Fo}(n) = min\{1., [pc(n) \cdot 0.2 - C_{Fo}(n-1) \cdot (1. - pc(n)) \cdot 0.01] + C_{Fo}(n-1)\}, \tag{4}$$

$$Fo(n) = p(n) \cdot pc(n) + Fo(n-1) \cdot (1. - pc(n)), \tag{5}$$

where $z(n)$ is the relative height of the peak in the current cepstrum, $h[\cdot]$ equals $(1. - g[\cdot])$, $\Delta p(n)$ is the difference between $p(n)$ and $Fo(n-1)$, divided by the frame rate, R is the distance of p beyond an upper pitch value delimiting the "male" pitch range, and $min\{a, b\}$ returns the minimum of a and b. In this way, Fo follows the fundamental frequency found in the "bottom-up" data, and the Fo confidence is increasing or equal to unity, if the data Fo is changing smoothly, is in the expected range, and is coming from

the expected direction. If these conditions are not met (i.e., $pc = 0.$) Fo remains constant and its confidence decays.

Time-waveform output is produced through an overlap-add process ([1]) using the final magnitude estimate E in combination with the input phase.

3. SYSTEM TEST

3.1. Methods

The ability of the implementation to deliver useful estimates was tested using a target (male) talker, in the presence of a jammer (female) talker. Different spatial conditions (target and jammer both straight ahead, and target ahead and jammer 60 degrees to the side) were simulated by convolving the original talker signals with impulse responses from an artificial head. Different target-to-jammer ratios (TJRs) were produced by applying gain to the jammer signal. Original jammer signals, and the impulse responses, were recorded in an anechoic chamber. The target signals were recorded in a slightly reverberant sound-attenuating chamber. A total of 5 target/jammer pairs (12.8 seconds of target talker sentences) was used in evaluation.

Performance is quantified using the change in TJR produced by processing. Due to the nonlinear nature of the processing, output TJR is not readily derived from input TJR. Thus, TJR was determined in 16 logarithmically-spaced frequency bands spanning 350-5000 Hz using a correlation procedure (cf. [7] and [5]). This was done by stimulating a gammatone filterbank ([9]) with a reference signal (the undisturbed target signal), passing the filter outputs through a simulation of auditory peripheral processing [3] and correlating the resulting "internal representation" with that found in response to a test signal (processed or unprocessed target-plus-jammer signal). The resulting correlations r_i are converted to TJR using $10log_{10}(r_i^2/(1. - r_i^2)$, and the TJRs are then averaged across frequency bands to yield a final TJR. This correlation method was "calibrated" by using it to determine the TJR of a target in the presence of a jammer over a range of known TJRs. Specifically, target and jammer were summed at a known TJR, and the correlation method then used to determine the TJR. When plotted against input TJR, the estimated TJRs, computed over an input range of -20 to +20 dB using 50 sentence pairs (total of 114 seconds per input TJR), were found to yield a line with slope 0.69 dB/dB and intercept -1.07 dB, with correlation 1.0. Thus, the procedure is found to be adequate for evaluating system performance.

3.2. Results

The reproduction of the target in the system output is qualitatively good. That is, little distortion of the target voice itself is heard, although the jammer is quite distorted due to removal of its pitched excitation by the periodicity algorithm, and varying attenuation produced by the spatial algorithm.

TJR IN	JPOS	BEST	SP	PER	SEP
8	0	7.9	-0.1	2.9	1.4
8	60	10.	2.3	3.6	2.9
4	0	4.2	0.0	3.3	1.9
4	60	6.2	2.8	3.2	3.6
0	0	0.5	-0.1	3.5	2.3
0	60	2.6	2.6	3.3	3.9
-4	0	-3.1	-0.1	3.5	2.0
-4	60	-1.1	2.6	3.5	4.5

Table 1: Results from the system test, shown as TJR gain in dB over TJR in the best ear (BEST) as a result of the separately performed spatial (SP) and periodicity (PER) processing, and the complete separation system (SEP), for different input TJRs (TJR IN)and jammer positions (JPOS).

Quantitative results for the system test are shown in Table 1 as the difference (or "gain") in TJR between the unprocessed "best ear" signal and the processed signals. The "best ear" is that with the higher received TJR, and is either right or left for target and jammer straight ahead (symmetric responses preclude a "best ear" for this condition), or the ear further from the jammer in the 60-degree condition. The TJR for the best ear is shown for reference in the column labelled BEST. Results are shown for different input TJRs and jammer positions (JPOS). Also shown for reference are results for processing using spatial information alone (SP) and periodicity alone (PER). The PER processing uses pitch information determined for target and jammer separately (as in [10]) to attenuate the jammer, and indicates the best performance possible for this type of processing (as do the SP results for the spatial processing). The results for the complete separation system are shown in the column labelled SEP.

Reference results in Table 1 indicate that the evaluation is consistent with expectations, despite the use of only 5 sentence pairs. For instance, TJR in the best ear changes on average 3.7 dB for a change of 4 dB in the input TJR, and the average difference between estimated and actual TJR in the straight-ahead condition is -0.4 dB. In addition, the average gain of 3.4 dB with the PER processing is near the approximately 4 dB inprovement found in [10] under similar conditions. Although these comparisons are expected to improve with an increased number of test sentences, they are consistently near expected values, and lend credence to the present evaluation.

Most significant of the results in Table 1 is the fact that the separation system always yields an improvement, that this improvement is always better than that from the SP processing alone, and is in some cases better than either processing alone. The latter occurs in the jammer off-axis conditions (JPOS = 60), where, most likely, the combination of the separation algorithms improves the system's ability to track the target pitch, thus improving the performance of the periodicity-based estimator. In addition, the improvement found for the separation system is greatest at the lowest TJR

tested, where improvement is most required. The extent to which this improvement is realized is being determined in ongoing investigations at lower input TJRs. The decreased average performance relative to the PER processing is most likely due to the fact that the separation system must locate the target pitch peak in the cepstrum of the *sum* of the target and jammer, while the PER processing uses the cepstrum of the target *alone* to find its pitch peak.

4. DISCUSSION AND CONCLUSION

The system evaluation indicates that the present system is able to combine the individual enhancement algorithm outputs in a favorable fashion. It was shown that the improvement obtained by the combination of algorithms was always better than that of the least effective algorithm, and in some cases better than the most effective. This indicates that the architecture can be successfully used for steering and combining different estimators.

The present system, which is much less complex than similarly motivated systems, has yielded TJR improvement in all of the conditions tested. Although these improvements were based on differences in pitch or spatial position of the target and jammer, such differences are expected to exist in common conditions. Of course, the improvement found in the present investigation will most likely decrease as the number of jammer sources increases, or when target and jammer attributes overlap more frequently. But this is true for any enhancement system. Additionally, it has been shown that the present architecture is able to track a target Fo after crossing with a jammer Fo, when the sources are in different spatial positions ([11]). The significant result here is that the present architecture was able to successfully combine separate algorithms, and do so in a very straightforward fashion.

Although the present system makes use of mainly low-level information and assumptions, integration of high-level knowledge is expected to be quite straightforward. For instance, a speech recognition system could be easily integrated by passing it the fundamental frequency and spectral estimates and confidences obtained by the present system, and allowing it to influence current processing by, for example, affecting the range for which periodicity is expected, or by influencing the confidence in fundamental frequency. Speech recognition could also take advantage of the confidence signals in decision making, giving higher weight to frames with higher-confidence attributes. Again, this interaction is simplified by the fact that decisions are only made concerning estimates of target attributes. This paper demonstrates that a system based on this idea can yield target talker improvement under a range of conditions. Further tests will show how wide this range is.

5. ACKNOWLEDGMENT

The first author is supported by the North Atlantic Treaty Organization under a postdoctoral fellowship awarded in 1996. Other support provided under DFG grant number Ko942/10.

6. REFERENCES

1. Jont B. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing*, 25 (3):235–238, 1977.
2. N. Carver and V. Lesser. Blackboard systems for knowledge-based signal understanding. In *Symbolic and Knowledge-based Signal Processing*, pages 205–250, A. V. Oppenheim and S. H. Nawab (eds.), Prentice-Hall, NJ, 1992.
3. T. Dau, D. Püschel, and A. Kohlrausch. A quantitative model of the "effective" signal processing in the auditory system: I. model structure. *J. Acoust. Soc. Am.*, 1996. in press.
4. Webster P. Dove. Knowledge-based pitch detection. Technical report, Number 518, Research Laboratory of Electronics, MIT, Cambridge, Massachusetts, 02139, USA, 1986.
5. R. Koch. *Gehörgerechte Schallanalyse zur Vorhersage und Verbesserung der Sprachverständlichkeit*. PhD thesis, Universität Göttingen, 1992.
6. B. Kollmeier, J. Peissig, and V. Hohmann. Real-time multiband dynamic compression and noise reduction for binaural hearing aids. *J. Rehab. Res. and Dev.*, 30 (1):82–94, 1993.
7. C. Ludvigsen, C. Elberling, G. Keidser, and T. Poulsen. Prediction of intelligibility of non-linearly processed speech. *Acta Otolaryngol*, Suppl. 469:190–195, 1990.
8. S. H. Nawab and V. Lesser. Integrated processing and understanding of signals. In *Symbolic and Knowledge-based Signal Processing*, pages 251–285, A. V. Oppeneheim and S. H. Nawab (eds.), Prentice-Hall, NJ, 1992.
9. R. Patterson, Nimmo-Smith I. J. Holdsworth, and P. Rice. An efficient auditory filterbank based on the gammatone function. In *Meeting of the IOC Speech Group on Auditory Modelling at RSRE*, pages 14–15, 1987.
10. R. J. Stubbs and Q. Summerfield. Effects of signal-to-noise ratio, signal periodicity, and degree of hearing impairment on the performance of voice-separation algorithms. *J. Acoust. Soc. Am.*, 66:497–500, 1991.
11. W. S. Woods, M. Hansen, T. Wittkop, and B. Kollmeier. Using multiple cues for sound source separation. In *Psychoacoustics, Speech and Hearing Aids*, pages 253–258, Kollmeier, B. (ed.), World Scientific, Singapore, 1996.