ノート

次元削減等の機械学習を用いた粒度分布測定結果解析手法の開発 －次元削減を用いたウルトラファインバブル粒度分布測定結果の解析－

水沼 達郎*

Analysis of Ultrafine Bubble Particle Size Distribution Measurement Using Dimensionality Reduction

MIZUNUMA Tatsuro*

1. 緒　言

ファインバブル(FB)とは，水中の気泡のことであり，その中でもバブル径が 1 μm 以下の微細な気泡をウルトラファインバブル(UFB)[1]と呼ぶ。UFB の特性は近年非常に注目されており，水産業・農業などで利用実績があるほか，UFB 水による洗浄効果が注目されており，家電製品・美容製品などでも UFB 発生機能を付加し，その効能を謳った製品が多数登場している[2]。

このような広範な用途で用いられる水中の UFB の数密度や粒子径分布の測定にレーザー回折・散乱法（LD 法）が用いられている。しかしながら，LD 法で UFB の測定を行うには課題点がある。

LD 法で得られるデータは，光強度の散乱光プロファイルである。どのような粒子であれ，媒体との屈折率の差に応じて特有の散乱特性を示し，測定すれば散乱光プロファイルを得ることができる。逆に言うと測定しているものが何なのかを区別するためには，散乱光プロファイルを識別することが必要であるが，粒子群の固有の粒度分布および屈折率に応じ散乱光プロファイルは変化するので識別するのは難しい。このことから LD 法による測定は，対象物が既知であることが前提である。しかし，測定対象物が確かに UFB であることを確認するのは容易ではない。水中に UFB が存在することを確認する手法としては，緩慢凍結を行い消泡を確認する方法[3]や，ソノルミネッセンスを利用した測定[4]などが提案されており，このような手法を用いて事前に UFB の存在を確認したうえで LD 法の測定結果を評価すると良い。しかしながら，これらの手法を併用することが困難な場合も多い。

そこで本報告では，限定された状況であるという制約がつくものの，LD 法で測定された粒子群が UFB であるのか，あるいはコンタミネーションがどの程度含まれているのかを簡易的に判別するための手法を機械学習の次元削減を用いて検討したので紹介する。

2. 次元削減を用いた LD 法測定結果の解析

2.1 次元削減について

次元削減は，高次元のデータを低次元データに圧縮する方法である。本所に設置されている粒度分布測定装置 SALD7500x10（(株)島津製作所製）から得られる散乱光プロファイルは 73 次元の数値ベクトルとなる。この 73 次元ベクトルを 2 次元平面上に射影することにより散乱光プロファイルの違いを可視化しようというのが基本的アイディアである。今回は線形的な次元削減である主成分分析(Principle Componet Analysis, PCA)，非線形的な次元削減として t 分布型確率的近傍埋め込み(t-distributed Stochastic Neigbor Embedding, t-SNE)[5]，Generative Topographic Mapping(GTM)[6]のそれぞれを検討した。

* 下越技術支援センター

2.2 検討内容

今回，想定した状況は以下のとおりである。

1. LD 法により得られた散乱光プロファイルが目的の測定対象物(UFB)のものであるか，そうでないのかを判断したい。
2. 目的の測定対象物(UFB)にコンタミネーションが含まれた可能性が考えられ，コンタミの度合いがおおよそどれくらいであるかを判断したい。

この状況を想定して以下の 2 つのタスクを検証する。

2.2.1 UFB と無機粒子の判別タスク

1 つ目は UFB 水と検定粒子 MBP1-10（MBP 粒子）を分散させた超純水（MBP 水）の判別タスクである。今回コンタミ粒子として実験に用いた MBP 粒子は，屈折率 1.92-1.94 で 1-10 μm に粒子径が分布するチタン・バリウム系ガラスである。前提条件として，この MBP 粒子が存在しない純粋な UFB 水の散乱光プロファイルを取得できたとした上で，コンタミ粒子として見立てた MBP 粒子が混入した場合に，2 次元平面上でどのように次元削減によりマッピングされるのか検証した。

2.2.2 UFB 水と水道水判別タスク

水道水は通常，さまざまな微粒子が混入している。その成分も配管由来のものや有機物など多様にわたるものと推測され，粒子径や個々の物性値も様々である。今回のタスクでは，UFB 水にこの水道水コンタミが含まれてしまった場合を考え，次元削減により 2 次元平面上にどのようにマッピングされるのかを検証した。

3. 実験手順

3.1 サンプルの準備および測定について

UFB の発生にはウルトラファインバブル発生装置 ultrafineGaLF（IDEC(株)社製）を用いた。発生装置は所定の圧力(270 kPa)に設定し，水量 8L の超純水を 1 時間循環運転をさせて UFB 水を得た。使用した超純水は 18.2 MΩ の抵抗値であることを確認した。

表 1 サンプル作製条件 UFB 水と MBP 水

	UFB100%	UFB91%	UFB83%	UFB67%	UFB50%	UFB25%	UFB0%
UFB水	10	10	10	10	10	10	0
MBP水	0	1	2	5	10	30	10

投入量(mL)

表 2 サンプル作製条件 UFB 水と水道水

	UFB100%	UFB75%	UFB50%	UFB0%
UFB水	10	15	10	0
水道水	0	5	10	10

投入量(mL)

また超純水中に MBP 粒子を適量混入させた水を作製し，これを MBP 水とした。MBP 粒子の濃度は，UFB の散乱光強度と同程度になるように希釈し，調整した。混ぜた条件は表 1 のとおりである。さらに実験室の水道水を採取し水道コンタミ水とした。水道水の採取にあたっては，最初の 3 分間水道水を流しっぱなしとして安定したコンタミ量となるように調整したのち，採取を行った。UFB 水と水道コンタミ水をそれぞれ混ぜたものを作製した。混ぜた条件は表 2 のとおりである。これらのサンプルを用いて粒度分布測定を行った。粒度分布測定には，回分セルを用いた。また UFB の屈折率 1.25 を設定し，定時点測定機能を用いて 5 秒間隔で 10 点測定を行った。

3.2 測定データの前処理と次元削減手法ライブラリ

測定で得られた各散乱光プロファイルに次元削減を適用して，73 次元から 2 次元に削減した。前処理として各手法を用いる前にデータの標準化を行った。PCA，t-SNE については，プログラミング言語 Python の Scikit-learn ライブラリ [7] を用いた。また GTM については，DCE-kit ライブラリ [8]を使用した。

4. 実験結果

4.1 UFB 水と MBP 水判別タスク

図 1 に PCA のマッピング結果を示す。図 1 左上にマッピングされた点●がコンタミの含まれていない純粋な UFB 水(UFB100%)である。それに対して体積換算で 9%の MBP 水を混入させた(UFB91% 以下同様表記)が図上部中央あたりにマッピングされ，点▲で示している。そのすぐ右となりに UFB83%の集団（点■）が存在する。UFB100%からの距離としては UFB91%のほうが近くなるはずでそのあとに UFB83%の集団が配置されるはずなので，コンタミ量に応じた距離の順序は合っているが，線形的な関係は見られなかった。残りの集団を見てみると UFB67%のデータが UFB100%群からかなり遠いところにあり，以降 UFB50%，UFB25%の距離も同程度となった。これらの一連のグループはおおまかにみると図左上から右下にむかって配置されているが，純粋なコンタミ粒子の散乱プロファイル UFB0%は図左下（点＋）にマッピングされた。PCA の結果は UFB25%の条件までの距離関係はおおよそ条件と合致したようにマッピングされているがコンタミ粒子の濃度が高くなりすぎるとマッピング面上に歪んで配置されているように見える。この原因についてはいくつか考えられるが，次元削減アルゴリズムが UFB と MBP 粒子が混ざっている水の散乱光プロファイルと MBP 粒子だけで構成された水の散乱光プロファイルが別物であると認識している可能性がある。これは粒子濃度が高くなることによる多重散乱が散乱光プロファイルに反映された可能性が推測される。

図 2 に t-SNE の結果を示す。こちらも傾向は PCA と同じ傾向であった。また，GTM（図 3）は，良いマッピング結果を示していて UFB25%程度のかなりコンタミ粒子の濃度が高くなる領域まで距離の順序をしっかり保ちつつマッピングすることができた。ただ，純粋なコンタミ粒子はやはり混合したサンプル群とは別の場所にマッピングされた。

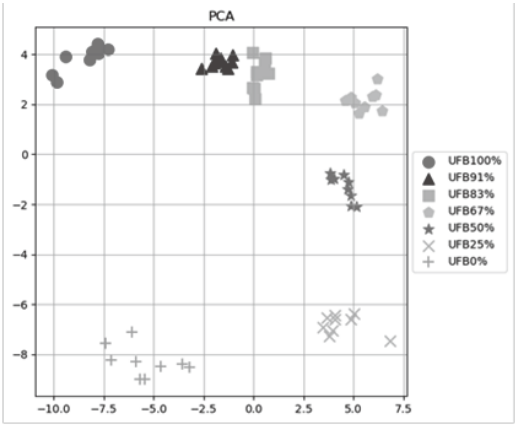

図 1　PCA マッピング結果　UFB 水＋MBP 水

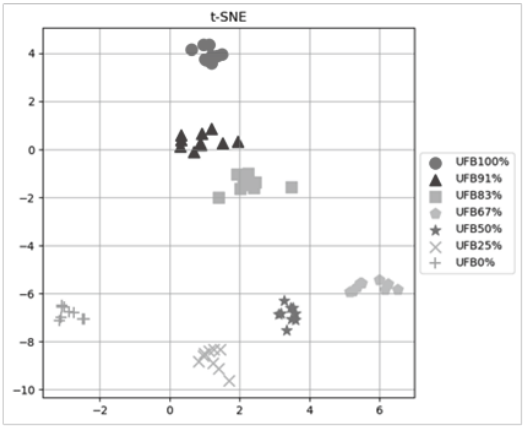

図 2　t-SNE マッピング結果　UFB 水＋MBP 水

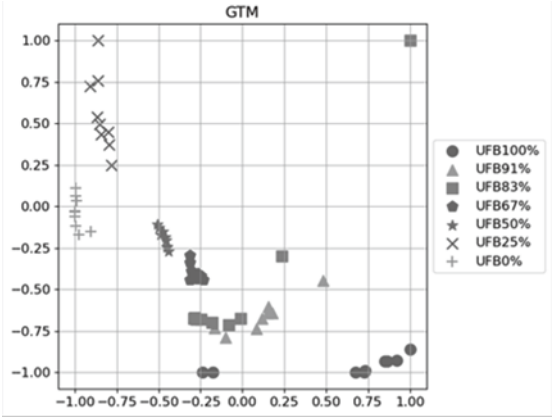

図 3　GTM マッピング結果　UFB 水＋MBP 水

4.2 UFB 水と水道水判別タスク

図 4 に PCA を適用した結果を示す。測定は 10 回の連続測定を再現性確認のため，数回おこなっており，水道水は 5 回×10 測定で合計 50 点あ

る。以降他条件の測定も再現性確認のため数回行ったものをすべて表示している。

UFB 水のみの結果 UFB100%は点●で表示されており，図中右下に分布する。また水道水のみ UFB0%の結果，点＋は図中左に分布している。この結果から水道コンタミ水と UFB 水に関しては PCA で完全に分離可能で，それらの判別は容易に行うことができる。つぎに水道水と UFB 水を体積で等量づつ混ぜたもの UFB50%，点★は，水道水と UFB 水の集団（クラスター）それぞれからの距離がちょうど同じくらいのところに分布していることが分かる。UFB 水と水道水の判別タスクにおいても基準となるデータとして純粋な UFB 水に加えて純粋なコンタミのものも用意できれば，両者の距離から両方の成分がどれくらいの比率で混ざっているかを判断することも可能である。なお，今回体積で等量混ぜておおよそ中間に位置が来ているのは，偶然にどちらの散乱光プロファイルも同じレベルの散乱光強度を示しているためである。当然片方が相対的に非常に濃い濃度であるならば体積で等量混ぜたとしても濃い濃度の粒子の散乱が支配的になってしまうため，このデータの見方には注意する必要がある。また，UFB100%もしくは UFB0%のどちらか片方のデータのみが分かる場合は，中間値の条件の距離は算出できないため違いがあるという判断のみが可能である点に注意が必要である。

つぎに UFB100%と UFB75%を比較すると UFB100%の周りにすこし離れてマッピングされており，水道水コンタミ粒子が存在していることは判別できている。しかしながら，UFB0%と UFB75%の距離を調べると一部のデータが UFB100%より遠いところにマッピングされてしまっていることが分かる。データを見てみるとこれは 2 回測定したうちの片方のデータが離れた集団でマッピングされてしまっていることが原因であった。測定時のデータの取り方に起因した可能性もあるものの，現時点でこのようにマッピングされた理由は追及していない。図 5 に t-SNE の結果を示す。t-SNE も PCA と同様の傾向であった。つぎに GTM の結果を図 6 に示す。GTM では，UFB75%が UFB100%と UFB 50%の中間にマッピングされうまく傾向をとらえることができたものの，UFB0%については，それら

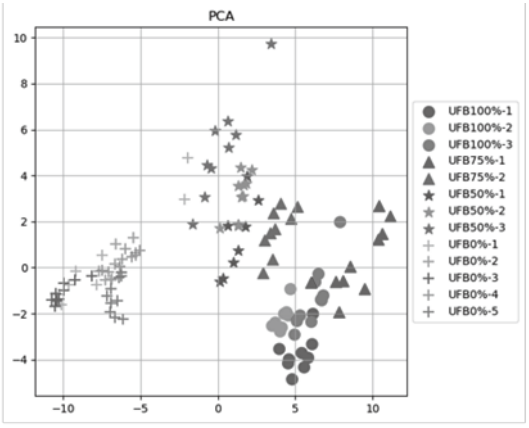

図 4　PCA マッピング結果　UFB 水＋水道水

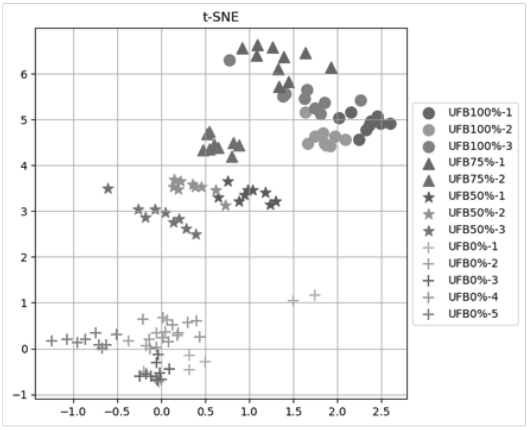

図 5　t-SNE マッピング結果　UFB 水＋水道水

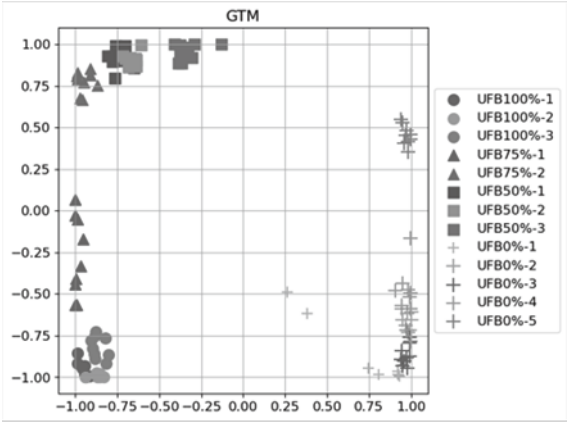

図 6　GTM マッピング結果　UFB 水＋水道水

とは離れた位置にマッピングされた。なお外周に沿ってマッピングされるのはGTM特有の傾向である。t-SNEやGTMは非線形の手法であり，それぞれチューニングする必要のあるハイパーパラメータが存在する。これらのハイパーパラメータの設定次第でマッピングの見た目はかなり変わり，極端に合わないハイパーパラメータを選択してしまうとそれぞれ条件と合致しないマッピングをされてしまう場合があった。

4.3 次元削減手法の選択と適用時の注意点

今回は，UFB水中の無機粒子をコンタミに見立てそれを判別する実験とUFB水と水道水成分を判別する実験をおこなった。今回検討した3手法とも純粋なUFB水とコンタミ粒子水の判別は可能であった。これらを混ぜ合わせたものについては，コンタミ粒子の濃度50%以下の領域は比較的良好な分離性能を発揮したと考えられる。ただ，コンタミ濃度が50%より大きくなる領域では，PCAではうまく距離関係がとらえられない場合があった。一方，GTMは良い結果を示し，コンタミ濃度の距離関係では良好なマッピングをした。しかし，GTMは，ハイパーパラメータの数が4つと多く，調整に苦労する。とくに今回はラベルありでの分離を行ったので，その結果をみてハイパーパラメータの調整を行うことができたが，実際の現場では，どの程度の比率で混入しているのか不明なサンプルを判別することになる。

このときにハイパーパラメータをどうするべきかという問題に対して，最適化するための指標がなくなってしまう。GTMは，EMアルゴリズム[9]を用いた最適化を行うため，指標の一つとしては尤度が高くなるものを選択するということが考えられる。計算機資源が潤沢であるならば，ある程度の範囲にパラメータをしぼってクロスバリデーションによる総当たりを行うのが良い。またはベイズ最適化[10]などのハイパーパラメータ選択手法を用いるのも良いだろう。また k3n-error[11]という高次元での距離関係と低次元での距離関係がどれくらい一致するかを計算する指標があり，この数値が一番よい結果を採用するのも選択肢の一つである。また t-SNE に関しては，ハイパーパラメータが一つだけだが，これも k3n-error を計算して選択するのが選択肢となる。

これらのハイパーパラメータを調整する必要のある非線形手法に対して，PCAはこのような調整を必要としないのが利点となる。計算資源も低コストで済み，一意性が担保できるという点で有力である。ただし，適用時の注意点がある。今回のような目的においては低次元で距離が近くにマッピングされるもの同士は，もとの高次元でも距離が近い必要があるが，残念ながらこの条件を満たさない場合がある。半面 t-SNE は高次元と低次元の距離関係がなるべく同じになるようにマッピングし，GTMは低次元で近くにきていれば高次元でも近いことが理論的に保証される[12]。したがって，本実験のようなケースの場合，目的に叶う手法はGTMとt-SNEである。しかし，今回のケースでは，どの散乱プロファイルも似たようなある範囲の数値しかとらないため，PCAの前述の条件が満たされないということはあまり気にする必要はないと考えられる。逆に今回のケースではなかったが，検出器の数値に極性があるような場合を判断するケース，たとえば電流値として＋－があるというような場合は，PCAを使う場合はこの点に気をつける必要がある。どの手法を用いるかについては計算機資源と目的に応じて柔軟に判断する必要がある。

また今回の実験では，コンタミ成分がいつも同等の散乱光プロファイルであることを仮定して分離している。水道水などは，その日によってコンタミ成分に違いが見られるため，注意が必要である。想定されるコンタミ成分が不明な場合は，コンタミを含まないサンプルに対してどれくらい近いかを判別することが可能である。しかし，どの程度コンタミが含まれるか定量的な判別を行う場合は，コンタミ成分のみの散乱

光プロファイルが必要となる点に注意が必要である。

また，より定量性のある手法の検討のため，同様のタスクをニューラルネットワークを用いて行ったので，この結果は別報にて紹介する。

5. 結　言

(1) UFB とコンタミ成分を判別するため次元削減で可視化し，含まれるコンタミ量を定量化することを目標とした。UFB 水と MBP 水の判別タスク，UFB 水と水道水の判別タスクを行った。おおよそのコンタミ量の違いに応じたマッピングが可能であり，判別することは可能であった。

(2) より高精度な定量性が欲しいケースでは次元削減では難しいと考えられる。もともと次元削減は，情報を捨て，判別できる程度に分けられれば良いという思想のものも多いと考えられ，数値的な定量を行う手段には向かない。また，基準となる純粋な散乱光プロファイルが必要となるという点も課題である。

参考文献

1) 例として https://shibata-corp.co.jp/babble/ufb/, 2024 年 3 月 29 日.

2) 経済産業省九州経済産業局, “ファインバブル活用事例集　追補版”, 平成 30 年 1 月.

3) 寺坂ら, “超純水中の空気ウルトラファインバブルの保存，輸送，濃縮，希釈および除去”, 混相流, 36 巻 1 号, 2022, pp.4-11.

4) 刈谷ら, “超音波を用いたウルトラファインバブルの評価手法に関する研究”, ソノケミストリー討論会講演論文集, 27(0), 2018, pp.47-48.

5) Laurens et.al, “Visualizing Data using t-SNE”, *Journal of Machine Learning Re-search*, 9, 2008, pp.2579-2605.

6) C.M.Bishop et.al, “GTM: The Generative Topographic Mapping”, *Neural Computation*, 10, 1, 1998, pp.215-234.

7) https://scikit-learn.org/stable/index.html, 2024 年 3 月 29 日.

8) https://datachemeng.com/dcekit/, 2024 年 3 月 29 日.

9) 例えば C.M.Bishop ら, “パターン認識と機械学習”, 下巻 第 9 章.

10) B Shahriari et. al, “A Review of Bayesi-an Optimization”, *Proceedings of the IEEE*, 104, 2016, pp.148-175.

11) H Kaneko, “K-nearest neighbor normal-ized error for visualization and recon-struction-A new measure for data visuali-zation performance”, *Chemometrics and Intelligent Laboratory Systems*, 176, 2018, pp.22-33.

12) 金子宏昌, “実践データサイエンス”, 第 5 章, 朝倉書店, p.202.